EGE Ecole de Guerre Economique

L'IA AU SERVICE DE LA CYBERSECURITE : APPROCHES, USAGES ET LIMITES DANS LES ORGANISATIONS

Fatima RAKHILA

Yves MOUVY

Amira HIDOUR
Jérôme CLAVILIER
Mohammed OUALI

MACYB03 de l'Ecole de Guerre économique

TABLE DES MATIERES

1 PREAMBULE

L'Intelligence Artificielle (IA) est devenue un élément incontournable et prépondérant de notre société, toutefois son développement rapide soulève de nombreuses questions en matière de cybersécurité. Comment respecter le juste milieu pour ne pas perdre l'Homme au centre de cette marche technologique, comment maitriser cet outil tout en restant étanche aux attaques qui pourraient en découler ? Comment protéger les données et mettre en place une réelle culture de la cybersécurité dans les organisations ?
Ce rapport explore les enjeux de cette convergence entre IA et cybersécurité, ainsi que les défis et les opportunités qui en découlent.
La technologie s'est complexifiée avec l'utilisation de gadgets de plus en plus sophistiqués, qui se surpassent continuellement en rendant obsolètes les avancées technologiques des décennies précédentes. Cette situation requiert de l'Homme une adaptabilité non seulement à ces nouveaux moyens technologiques, mais également à la maîtrise de la productivité. Une productivité basée sur les données, les automatisations et les apprentissages. L'intelligence artificielle cœur de notre démarche, trouve une place de choix dans ce panorama.
Dans « A quoi rêvent les algorithmes », (CARDON, 2015), Dominique Cardon rappelle que : *« si les logiques de personnalisation s'installent aujourd'hui dans nos vies, c'est parce qu'elles calculent une forme nouvelle du social, la société des comportements, où se recomposent la relation entre le centre de la société et des individus de plus en plus autonomes. »*
C'est une société nouvelle que nous sommes appelés à construire ou déconstruire. Les algorithmes et les processus d'apprentissage nous ouvrent la voie. Au-delà des enjeux cyber que cela peut induire, se pose une réelle problématique de la place de l'Homme dans cet univers, à savoir un nouveau type de société qui favorise le déploiement des algorithmes et leur permet d'avoir des rêves.
L'IA est multiforme et trouve ses applications dans plusieurs techniques. *L'apprentissage machine, les réseaux de neurones profonds, l'analyse prédictive, le traitement du langage naturel, les agents conversationnels*[5], sont autant de domaine de développement de l'IA dont l'objectif serait de faciliter la vie de l'Homme.
Comment se fait-il que l'Homme, avec toute « sa puissance » et sa maîtrise de la nature, soit aujourd'hui contraint de confier son quotidien à des interfaces technologiques ? Autrefois, la question était de maîtriser la technologie, mais aujourd'hui, c'est de lui faire confiance. La sécurité et la fiabilité des nouvelles technologies soulèvent des préoccupations quant à la protection de l'Homme face à leur potentiel d'erreur et de danger.
En 1995, Pierre Lévy prônait (LEVY, 1995) déjà l'intelligence collective et prévoyait une anthropologie du cyberespace. Le terme "cyberespace", d'origine américaine, a été utilisé pour la première fois par l'écrivain de science-fiction William Gibson dans son roman "Neuromancien" (GIBSON, 1984). Gibson y définissait le cyberespace comme l'univers des réseaux numériques, un lieu de rencontres et d'aventures, mais également un enjeu de conflits mondiaux, une nouvelle frontière économique et culturelle. Comment l'Homme peut-il trouver sa place dans cet univers inconnu, aux confins incertains et aux réalités non maîtrisées ? Ces questionnements trouvent des essais de réponses avec une approche humaniste du cyberespace.

1.1 Définition du cyberespace

Si pour Gibson, le cyberespace est un univers de réseaux numériques intangibles, une approche anthropologique le transforme en lieu d'exercice de pouvoirs et de contrepouvoirs, où se mêlent des enjeux géopolitiques, commerciaux, ludiques, sportifs et criminels. Selon Bernard Stiegler, "Le cyberspace est un espace de production de l'individu augmenté, c'est-à-dire d'un individu qui se constitue lui-même en produisant des connaissances, des relations sociales, des formes de vie et des dispositifs techniques qui le renforcent." (STIEGLER, 2016). De cette approche, il découle que l'Homme "s'augmente" dans le cyberespace grâce à la maîtrise de la technologie et des processus.

[5] Tous ces types d'intelligence artificielle sont développés plus loin dans le document

Le sociologue Serge Proulx va plus loin en précisant que *"Le cyberspace est un lieu de production, de circulation et de transformation de l'information, qui est gouverné par des règles et des normes spécifiques et qui est marqué par une forte dynamique de l'innovation technologique."* (Serge Proulx, "Internet et la transformation de la communication", 2006). Cette notion de production, transformation et innovation est essentielle pour comprendre le rôle central du cyberespace dans les mutations sociales, politiques, économiques et intellectuelles. Le bien-être de l'Homme doit être placé au centre de son action et favoriser l'innovation.
Cette créativité sans cesse renouvelée est accompagnée de la menace constante de cybercriminalité et de nouveaux dangers. La cybersécurité est donc essentielle pour protéger les utilisateurs de ces menaces. Une approche anthropocentriste de l'innovation et de la cybersécurité est nécessaire pour garantir que les innovations technologiques sont au service de l'Homme et non l'inverse. Cela implique de développer des technologies conviviales, respectueuses de la vie privée et des droits de l'Homme, et de mettre en place des mesures de sécurité efficaces pour protéger les utilisateurs.

1.1 Mutations structurantes dans le développement humain vis-à-vis des innovations

La maîtrise de la technologie a connu une évolution en plusieurs étapes majeures. La préhistoire a vu l'apparition des premiers outils en pierre taillée, il y a environ 2,6 millions d'années, qui ont été utilisés par les premiers hominidés pour chasser et se nourrir. Par la suite, de nouvelles compétences ont été développées et améliorées tout au long de l'histoire humaine, de l'Antiquité jusqu'à la Renaissance.
L'Antiquité : Les civilisations antiques ont apporté des avancées significatives dans l'agriculture, l'irrigation, la construction de bâtiments et la médecine, améliorant ainsi la productivité de la production agricole et des soins de santé, et permettant une expansion des villes et des centres de commerce.
Au Moyen Âge, l'émergence de guildes et de corporations a permis une meilleure organisation de la production et du travail, conduisant à une augmentation de la productivité de l'artisanat et de la production de biens manufacturés.
La Renaissance a vu l'émergence de nouvelles technologies telles que l'imprimerie, la navigation et l'horlogerie, qui ont permis une augmentation de la productivité dans la production de livres, l'exploration de nouveaux territoires, et la mesure du temps, permettant une meilleure planification et une meilleure utilisation du temps.
La Révolution industrielle, au XVIIIe et XIXe siècle, a marqué une étape importante dans l'histoire de la technologie, avec l'émergence de nouvelles technologies telles que la machine à vapeur, la production de masse, les chemins de fer et la télégraphie. Cette évolution s'est poursuivie au XXe siècle, avec la maîtrise technologique orientée vers l'informatique, la médecine, la conquête de l'espace et des territoires inconnus.
Le XXIe siècle voit l'avènement des nouvelles technologies passionnantes telles que l'intelligence artificielle, la réalité virtuelle, la robotique, la biotechnologie, les énergies renouvelables, bientôt peut être la fusion nucléaire, et bien d'autres encore. Toutes ces avancées technologiques ont pour objectif ultime le bien-être de l'homme et la standardisation de son quotidien.
Cependant, ces nouvelles technologies posent également de nouveaux défis, notamment en termes de cybersécurité. Selon le rapport du groupe Cisco, (CISCO, 2022), il y aurait environ 3,6 objets connectés par personne, créant ainsi de nouvelles vulnérabilités pour les cyberattaques. À mesure que les technologies évoluent, la question de leur fiabilité et de leur sécurité devient de plus en plus préoccupante, avec des enjeux d'intégrité, d'éthique et de protection de l'utilisateur. Les organisations qui sont en mesure d'adopter rapidement les nouvelles technologies tout en garantissant leur sécurité ont désormais un avantage concurrentiel sur les autres.

L'interaction avec Alexa se fait sans avoir besoin de taper ou de cliquer, ce qui peut avoir des implications sur la façon dont les gens interagissent avec les technologies.

Christina Dunbar-Hester a examiné l'impact de l'assistant vocal d'Amazon sur les perceptions culturelles de la voix et de la communication. Elle a noté que l'interaction avec Alexa peut changer la façon dont les gens parlent et perçoivent leur propre voix. Par exemple, les utilisateurs peuvent adopter une façon de parler plus directe ou plus robotique lorsqu'ils interagissent avec Alexa, ce qui peut influencer leur façon de communiquer avec les autres.

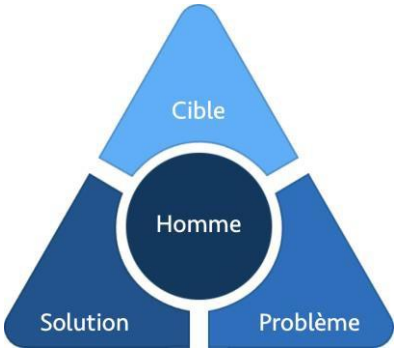
Cette observation soulève des questions sur la manière dont les technologies de l'IA peuvent influencer les pratiques sociales et culturelles. L'IA est souvent conçue pour répondre aux besoins et aux désirs des utilisateurs, mais elle peut également affecter leur comportement et leur perception de la réalité.

1.2 L'homme, l'intelligence artificielle et la cybersécurité

L'importance de la sécurité et du contrôle est primordiale dans la maîtrise de la technologie dans les organisations, surtout dans le contexte de l'intelligence artificielle (IA). La cybersécurité est un élément crucial dans le développement et l'utilisation de l'IA, au regard de la quantité de données qu'elle traite et stocke. Les cyberattaques peuvent entraîner des conséquences désastreuses telles que la compromission de la confidentialité des données, des perturbations de service, voire des dommages physiques dans certains cas.

Des études menées par des chercheurs de l'Université de Californie à Berkeley et de l'Université de Washington ont montré que les utilisateurs d'assistants vocaux intelligents et de systèmes de sécurité domestique ont des préoccupations quant à la sécurité de ces systèmes. Cependant, ces préoccupations sont souvent atténuées par la commodité et l'utilité des assistants vocaux et la nécessité de se sentir plus en sécurité chez soi.

Il est essentiel de revenir aux fondamentaux et d'impliquer l'homme dans la démarche de sécurité. Les entreprises doivent mettre en place des systèmes de contrôle pour garantir la sécurité de leurs données et actifs contre les cyberattaques. Selon Lianne Potter, une cyber anthropologue, l'homme est à la fois le problème, la cible et la solution de la cybersécurité.



Il est donc important de prendre en compte les perspectives des utilisateurs lors de la conception et le déploiement des systèmes d'IA sécurisés, car leur relation avec ces systèmes est souvent complexe et nuancée. Les utilisateurs peuvent être prêts à accepter des risques en matière de sécurité en échange d'autres avantages, comme la commodité ou la sécurité perçue. En conclusion, la sécurité et le contrôle doivent être intégrés dans le développement et l'utilisation de l'IA, avec une attention particulière aux besoins et aux préoccupations des utilisateurs.

1.3 Perception et enjeux de la Cybersécurité et l'IA

La perception de la cybersécurité et de l'IA varie considérablement d'une personne à l'autre. Certaines considèrent la cybersécurité comme une mesure de protection contre les attaques malveillantes ou les pertes de données, tandis que d'autres la voient comme une atteinte à leur vie privée et à leurs libertés individuelles. De même, l'IA peut être perçue comme une menace pour l'emploi et la vie privée, ou comme une opportunité de résoudre des problèmes complexes et de créer de nouveaux produits et services.

Il est essentiel de noter que la perception de la cybersécurité et de l'IA peut évoluer avec le temps, en fonction des expériences et des événements. Par exemple, un utilisateur peut être sceptique à l'égard de l'IA jusqu'à ce qu'il en voie les avantages concrets, ou ne pas accorder d'importance à la cybersécurité jusqu'à ce qu'il subisse une attaque ou une violation de données.

La perception de chacun sur la cybersécurité et l'IA est influencée par son environnement et son expérience personnelle. C'est pourquoi il est crucial de prendre en compte les différents points de vue et les préoccupations des utilisateurs lors de la conception et de l'utilisation de technologies telles que l'IA. En fin de compte, les entreprises doivent s'assurer que les avantages de la technologie l'emportent sur les risques, tout en protégeant la vie privée et les données des utilisateurs.

Le déploiement de l'intelligence artificielle (IA) a un impact important sur la cybersécurité, car il augmente le nombre de points d'entrée potentiels pour les cyberattaques. Les algorithmes d'apprentissage machine et les réseaux de neurones profonds sont des exemples d'IA qui peuvent être utilisés pour détecter et prévenir les cyberattaques. Ces technologies peuvent également être utilisées par les cybercriminels pour développer des attaques plus sophistiquées et plus difficiles à détecter.

Par conséquent, la cybersécurité doit évoluer pour suivre le déploiement de l'IA. Cela inclut l'adoption de nouvelles technologies de défense, telles que l'IA elle-même, pour protéger les systèmes contre les attaques (MATANIA & RAPAPORT, 2022). Dans les « guerres du futur » in (MATANIA & RAPAPORT, 2022), les auteurs rappellent la nécessaire question de l'équilibre offensive – défensive. Autant l'on déploie des algorithmes d'aménagement du quotidien, de préparations offensives, autant les outils de défenses doivent être d'une acuité certaine.

Les experts en cybersécurité doivent également être formés à la compréhension de l'IA et de ses applications en matière de sécurité. Les gouvernements et les organisations doivent également mettre en place des politiques et des réglementations pour s'assurer que l'IA est utilisée de manière responsable et éthique. Le développement de l'intelligence artificielle entraîne des conséquences importantes sur le marché du travail.

1.4 Limites de l'IA

Sans une réelle coordination à tous les niveaux, on pourrait se retrouver dans des travers jusque-là craints. Dans « Les défis de la cybersécurité », (AVOINE & KILLIJIAN, 2020), les auteurs reviennent sur les risques qui guettent tous les acteurs qui manipulent les données de productions de l'IA : fuite de données personnelles, piratages massif, espionnage économique, usurpation d'identité, infection de systèmes informatiques sensibles etc.

L'IA a des implications sociales et sociétales importantes, notamment en ce qui concerne la distribution des richesses et du pouvoir. Les personnes qui contrôlent l'IA et ses données peuvent devenir très puissantes et entrainer des inégalités économiques et politiques. C'est le cas notamment des GAFAMI qui captent aujourd'hui une grande partie des richesses produites par l'IA.

L'IA peut également affecter les relations sociales entre les personnes. Par exemple, les assistants virtuels peuvent ajuster la façon dont les personnes interagissent avec les autres, en les rendant plus isolées ou en changeant les normes sociales. La discussion avec les dispositifs dits « intelligents » relève plus des commandes déshumanisées. Le risque serait de le voir, si ce n'est déjà le cas, dans les relations sociales. Le revers peut être de croire en une grande proximité avec les produits de l'IA. Geneviève Bell a mené une étude sur la façon dont les gens interagissent avec les assistants virtuels tels que Siri et Alexa. Elle a montré comment les gens développent des relations affectives avec ces assistants virtuels et les considèrent comme des membres de la famille ou des amis, même s'ils savent qu'ils ne sont que des machines.

Une autre préoccupation concerne la question de la discrimination et des biais dans les algorithmes d'IA. Les algorithmes sont « influencés » par les données qu'ils traitent, et si ces données sont biaisées ou discriminatoires, cela peut conduire à des résultats discriminatoires. Par exemple, un algorithme utilisé pour le recrutement peut être biaisé contre les femmes ou les personnes de couleur, en raison de données historiques qui ont elles-mêmes été influencées par des préjugés et des discriminations.

Sur le plan technologique, bien que l'IA ait connu des avancées spectaculaires ces dernières années, elle présente encore des limites importantes. Par exemple, les algorithmes de reconnaissance d'images peuvent être biaisés et donner des résultats erronés, en particulier pour les minorités ethniques. De même, les chatbots et autres programmes de traitement du langage naturel peuvent mal interpréter la langue, ce qui peut causer des problèmes dans les interactions avec les clients. En ce qui concerne la cybersécurité, ces limites peuvent être exploitées par les attaquants pour contourner les défenses de l'IA, telles que les systèmes de détection des intrusions.

En 2020, la ville de Nice a utilisé des caméras de surveillance équipées de logiciels de reconnaissance faciale pour surveiller les comportements des citoyens dans l'espace public. Cependant, des experts ont mis en évidence des erreurs et des biais dans les résultats, tels que la confusion de personnes d'origines ethniques différentes. De plus, les défenseurs des droits de l'Homme ont soulevé des préoccupations concernant la vie privée et la surveillance de masse, et ont appelé à l'arrêt de l'utilisation de la reconnaissance faciale. Ces systèmes peuvent également être vulnérables aux attaques d'usurpation, dans lesquelles des attaquants utilisent des images ou des vidéos falsifiées pour tromper le logiciel de reconnaissance.

Un autre problème important de l'IA est le manque de transparence dans les décisions prises par les algorithmes. Les processus de prise de décision peuvent être opaques et difficiles à comprendre, ce qui peut causer des problèmes d'éthique et de responsabilité. Ce manque de transparence peut rendre difficile la détection des failles de sécurité ou des tentatives de sabotage de l'IA.

Toujours en France en 2019, la Commission nationale de l'informatique et des libertés (CNIL) a infligé une amende de 50 millions d'euros à Google pour violation du règlement général sur la protection des données (RGPD). La CNIL a constaté que Google n'avait pas informé clairement les utilisateurs de la manière dont leurs données étaient collectées et utilisées, en particulier en ce qui concerne les publicités personnalisées basées sur l'IA. Ce manque de transparence dans les pratiques de collecte et d'utilisation des données peut également poser des problèmes de cybersécurité, car les attaquants peuvent exploiter ces lacunes pour accéder à des données sensibles.

Comment garantir que les décisions prises par les algorithmes d'IA soient transparentes et responsables ? Ces questions sont encore largement débattues, et de nombreuses organisations et gouvernements travaillent à la mise en place de réglementations et de normes éthiques pour encadrer le développement et l'utilisation de l'IA.

En France, la question de la responsabilité liée à l'utilisation de l'IA est un enjeu important. En effet, l'IA peut prendre des décisions qui entraînent des conséquences graves, par exemple en matière de santé ou de justice. Il est donc crucial de déterminer qui est responsable en cas de préjudice causé par une décision prise par une IA.

Dans ce contexte, la loi française a introduit en 2016 le concept de "responsabilité de plein droit" pour les activités liées à l'IA. Cette disposition impose aux concepteurs et fabricants d'IA de prendre toutes les mesures nécessaires pour éviter les dommages causés par leurs produits, sous peine de devoir indemniser les victimes sans qu'il soit nécessaire de prouver leur faute.

En outre, la Commission Nationale de l'Informatique et des Libertés (CNIL) a publié en 2019 des recommandations sur l'utilisation de l'IA en matière de santé. Ces recommandations soulignent la nécessité d'assurer la transparence et la traçabilité des décisions prises par les systèmes d'IA, ainsi que de garantir l'information et le consentement des personnes concernées.

L'ANSSI, l'Agence Nationale de la Sécurité des Systèmes d'Information, a publié en 2020 un livre blanc sur la cybersécurité des systèmes d'IA. Ce livre blanc met en garde contre les risques de cyberattaques visant les systèmes d'IA, qui pourraient entraîner des conséquences graves en termes de sécurité et de vie privée des personnes concernées.

Le sujet de l'IA et de la cybersécurité ainsi posé, nous ramène à une remarque essentielle : Quelle société souhaitons-nous construire grâce à l'IA ? De quels types d'IA disposons nous ? Quelles sont leurs domaines d'application ? Avec quelles protections et quels process sont-elles déployées ? Dans quel cyber espace ?

2 INTRODUCTION

2.1 Notre sujet de mémoire

L'Intelligence Artificielle (IA) révolutionne et façonne notre monde, elle redéfinit les rapports de puissance entre les nations. Elle prend dans ce contexte une place de plus en plus importante au niveau de toutes les strates de la société et devient au-delà de ses applications un véritable enjeu d'influence et de pouvoir.

Qu'en est-il de la place de l'IA dans la cybersécurité et plus largement la sécurité ? Comment l'IA est-elle déjà appliquée à la cyberdéfense ou la cyberattaque ? Quels sont les risques et les opportunités associés à cette technologie ? Quels sont les impacts de l'IA sur les organisations ?

Telles sont les quelques questions que nous aborderons dans le mémoire. Pour ce faire, nous étudierons les principaux concepts liés à l'intelligence artificielle ainsi que ses champs d'application civiles ou militaires. Nous passerons ensuite en revue les principales doctrines utilisées en cybersécurité ainsi que les concepts associés. Il nous est apparu nécessaire d'étudier les aspects réglementaires encadrant cette technologie car ils traduisent les questionnements de la société sur cette technologie et illustrent les rapports de force entre d'une part les différents acteurs d'une même société mais également les tentatives et stratégies mises en œuvre par les nations pour se protéger. Nous abordons par la suite des aspects plus techniques en essayant dans la mesure du possible d'illustrer les applications de l'IA au domaine de la cybersécurité. Nous étudierons également les failles de sécurité spécifiques à ce type de système. Au-delà des illustrations, il nous apparaît essentiel pour une meilleure compréhension des enjeux d'expliquer le mode de fonctionnement de certains algorithmes. Nous pensons qu'il est important de fournir au lecteur un minimum d'explications sur les modalités de traitement des informations, ainsi que les principes utilisés pour le développement des IA dans toutes leurs diversités, afin de donner les outils critiques pour une meilleure compréhension du sujet et des enjeux associés. La description des applications existantes ou en cours de développement permet également de démystifier le sujet. Nous pensons que l'IA est une technologie relativement ouverte et que tout citoyen avec un minimum de bagage scientifique est capable de comprendre les principes sous-jacents à cette technologie et que compte tenu de la disponibilité des algorithmes en libre accès, il est possible d'expérimenter avec cette technologie. C'est ce que nous essayons de montrer en développant par exemple un anti-malware statique (antivirus) en annexe du présent document. Nous tenterons, par la suite, de donner une certaine visibilité du marché de l'IA dans la cybersécurité afin d'identifier les principaux pôles d'innovation et les tendances à court ou moyen termes. Enfin, il n'existe pas vraiment de technologie « neutre », dans ce contexte, nous tenterons d'identifier les impacts sociaux-économiques dans un proche avenir.

Notre objectif est d'étudier modestement les possibilités offertes par les technologies issues de l'IA, d'en explorer les opportunités et les limites afin d'apporter un éclairage sur le sujet. Pour ce faire et le lecteur l'aura compris, nous traitons le sujet sous le prisme historique, social, économique, scientifique et technique. Il nous apparaît intéressant d'essayer d'adopter cette approche multidisciplinaire pour tenter d'enrichir notre réflexion.

2.2 Une première définition de l'IA

Nous y reviendrons plus tard, mais dans le cadre de cette introduction et afin de fixer les idées et définir au moins génériquement dans un premier temps le concept d'IA, nous reprendrons les propos du Professeur Stuart Russell (Berkley University), qui la définit comme « *l'étude des méthodes permettant aux ordinateurs de se comporter intelligemment* », ce qui englobe les tâches telles que l'apprentissage, le raisonnement, la planification, la perception, la compréhension du langage et la robotique (Miailhe, 2018), (Russell, 2016).

2.3 Une révolution en marche ?

Il apparait essentiel d'évoquer rapidement les problématiques liées aux enjeux de pouvoir, de souveraineté ainsi que les différents champs d'application. L'IA n'étant pas une technologie neutre, il est important de dresser un panel de ses impacts sur nos sociétés avant de regarder plus spécifiquement un usage qui se voudrait plus « neutre », à savoir la cybersécurité.

La révolution des usages de l'IA et son essor récent, bien que les premiers travaux théoriques sur le sujet remontent à plus de 70 ans (Loiseau, 2019), sont rendus possibles grâce à la convergence du réseau internet, du « big data » (massification des données) et des moyens de calcul. En effet, le développement de l'IA se fait dans un contexte technologique marqué par la « *mise en données* » du monde, qui touche des domaines aussi variés que la robotique, la blockchain, le calcul intensif ou le stockage massif et des secteurs comme la banque, l'industrie, la défense, l'administration et la politique (Villani, 2018).

Pour illustrer notre propos, actuellement, des systèmes sont capables de reconnaître la parole articulée et de la retranscrire. D'autres systèmes d'IA sont capables de faire de la reconnaissance faciale avec précision ou d'attribuer des empreintes digitales. Il est possible d'entrainer une IA à comprendre des textes écrits en langage naturel. Des centres d'appels sont augmentés par des agents conversationnels synthétiques reposant sur l'IA. Des voitures sont pilotées par des systèmes d'IA, des automates diagnostiquent parfois mieux que des dermatologues des mélanomes à partir de photographies de grains de beauté prises par des téléphones portables. Des chaînes de fabrication sont complètement automatisées dans les usines.

Dans le domaine des mathématiques et des sciences physiques, des IA démontrent ou aident à démontrer des théorèmes mathématiques et construisent automatiquement des connaissances à partir de masses immenses de données.

Dans le domaine de la biologie, les techniques d'apprentissage sont utilisées pour déterminer la fonction de macromolécules biologiques, en particulier celle de protéines et de gènes, à partir de la séquence de leurs constituants, acides aminés pour les protéines, nucléotides pour les gènes (Ganascia, 2022).

Plus généralement, toutes les sciences subissent une rupture épistémologique majeure avec les expérimentations dites *in silico* sur des données massives (Ibid.).

2.4 Un vecteur de « soft-power »

L'IA est devenue un véritable enjeu de puissance entre les différents États et multinationales. En témoigne les investissements dans la recherche et dans l'industrie de l'IA qui atteignent actuellement des sommes très importantes, notamment en Chine et aux États-Unis. Ces financements traduisent l'importance stratégique prise par ces technologies. Ainsi, les dépenses pour le développement des technologies relatives à l'IA aux États-Unis atteindront 120 milliards de dollars d'ici 2025 (Coret, 2022); la Chine avait déjà investi en 2017 plus de 70 milliards de dollars (Schaeffer, 2020). L'Union européenne est à la traine et a investi pour la période 2018-2020 1,5 milliards d'euros dans le secteur, complété de 2,5 milliards d'euros supplémentaires au titre de la politique européenne du numérique (Fekhi S.,2021). Récemment la Commission européenne semble avoir mesuré son retard. Elle prévoit d'investir 1 milliard d'euros par an dans le secteur. Elle devrait également mobiliser *« des investissements supplémentaires du secteur privé et des États membres afin d'atteindre un volume d'investissement annuel de 20 milliards d'euros au cours de la décennie* » (strategy, 2023). Ce chiffre de 20 milliards est néanmoins trompeur, puisqu'il ne s'agit absolument pas d'investissements directs de la part de l'Union européenne, selon une stricte interprétation des propos de la Commission européenne.

Dans la mesure où les chaînes de valeur, surtout dans le secteur numérique, sont désormais mondiales, les pays qui seront les leaders dans le domaine de l'IA seront amenés à capter une grande partie de la valeur des systèmes qu'ils transforment, mais également à contrôler ces mêmes systèmes, mettant en cause l'indépendance des autres pays (Villani, 2018).

Nous assistons depuis longtemps déjà à l'essor de véritables « empires numériques » plus ou moins soutenus et/ou contrôlés par des États qui financent le développement des bases scientifiques et techniques sur lesquelles ces entreprises innovent et prospèrent (Miailhe, 2018). Ces organisations provoquent ou renforcent un mouvement global de centralisation du pouvoir dans les mains d'une poignée d'acteurs ayant un accès privilégié aux données. Cela étant dit, la spécificité des géants du secteur qu'ils soient américains GAFAMI (Google, Apple, Facebook, Amazon, Microsoft, IBM) ou chinois BHATX (Baidu, Huawei, Alibaba, Tencent, Xiaomi) tient plus à leur modèle économique novateur dans lequel le consommateur tient une place centrale, qu'aux technologies déployées.

2.5 Un risque de perte de souveraineté

Les masses critiques de données (big-data), les capacités de traitement pour l'analyse et la montée en puissance de l'IA rendent possible cette concentration/centralisation du pouvoir numérique entre les mains de ces « empires numériques ». Ils sont les seuls à disposer des données de leurs clients/citoyens, de la puissance de calcul nécessaire et d'un vaste panel de compétences pour extraire la valeur ajoutée de ces données en les transformant dans le meilleur des cas en service et dans le pire des cas en un vecteur d'influence politique.

Dans son rapport sur l'IA de 2018, Cédric Villani pointe le risque de captation de la valeur et de la compétence par des « organisations étrangères ». Ceci reflète bien évidement le point de vue français. Mais force est de constater, qu'à bien des égards, la France et l'Europe peuvent d'ores et déjà faire figure de « colonies numériques » (Villani, 2018). En effet, les grands acteurs du domaine (tous étrangers) captent toute la valeur ajoutée : celle des cerveaux quelles recrutent, celle des applications et services qu'elles créent et par les données qu'elles absorbent. Le mot est certes fort, mais techniquement, il s'agit bien d'une démarche de type « soft-colonisation » en exploitant les ressources locales et en mettant en place un système qui attire la valeur ajoutée.

L'Europe a accumulé un profond retard dans le domaine. Dans les faits, son approche consiste essentiellement à profiter de son marché de 450 millions de consommateurs pour jeter les bases d'un modèle industriel « éthique » de l'IA, tout en essayant de négocier un partenariat stratégique avec les États-Unis et en acceptant *de facto* une forme de « vassalisation douce ». La stratégie consiste plus à réguler la révolution de l'IA qu'à l'accompagner et la favoriser (Miailhe, 2018). Son ambition s'articule autour d'une volonté de reconquête d'une certaine forme de souveraineté, de recherche de puissance et de respect de la personne humaine. La bataille du contrôle des données et de l'IA se joue pour les européens essentiellement sur le front de la régulation et du droit, compte tenu des niveaux d'investissement consenti par l'Union européenne (Tabouy, 2022). Cette stratégie vise à éviter ou au moins atténuer le risque de « cyber-colonisation » (Miailhe, 2018).

Du point de vue du « soft-power », le développement (et l'utilisation) de l'IA est constitutif d'un type de puissance permettant d'influencer, par des moyens non coercitifs le comportement d'acteurs, ou la définition que ces acteurs ont de leurs intérêts. En ce sens, il s'agit d'une certaine façon d'un projet politique de la part de ces empires numériques, en parallèle de la recherche d'un profit légitime.

2.6 Les impacts sociétaux de l'IA

Les données (et notamment nos données personnelles) sont l'une des ressources les plus précieuses à l'heure de l'économie de la connaissance. Elles sont au cœur du fonctionnement des intelligences

artificielles, elles déterminent notre capacité à organiser les connaissances, à leur donner un sens, à augmenter nos facultés de prise de décision et de contrôle des systèmes. Selon Cédric Villani, l'IA est l'une des clés du pouvoir de demain (Villani, 2018).

L'accumulation d'importante volumétrie de données sur les consommateurs/citoyens a bien évidement pour buts de gagner des parts de marché et de disposer d'un pouvoir d'influence sur les modes de pensées. Dans ce contexte, l'analyse de nos données à travers les technologies de l'IA est un enjeu de sécurité et d'influence pour les compagnies privées mais également pour les États.

L'un des exemples les plus frappant qui doit nous inciter à réfléchir, sur la nécessaire régulation de l'usage des données et de l'IA, est le scandale de la société britannique de conseil politique Cambridge Analytica. Cette entreprise a pu accéder aux données personnelles issues des comptes Facebook de 87 millions citoyens américains, et qui par leur exploitation a influencé le résultat des élections présidentielles américaines de 2016 (Vang, 2020). D'ailleurs, Cambridge Analytica ne semblait pas particulièrement cacher ses activités puisque ces représentants déclaraient utiliser « *des techniques scientifiques avancées de recherche et d'analyse sociale, adaptées à un usage civil à partir d'application militaires (opérations psychologiques), pour mieux comprendre les comportements des électeurs* » (Ibid.).

Ce n'est pas le seul fait d'armes de cette compagnie puisqu'elle a pesé sur l'élection du président Uhuru Kenyatta au Kenya en 2013 et en 2017 (Ibid.). Et, bien que Cambridge Analytica ne soit plus en activité depuis 2018, il reste de multiples entreprises avec ce type d'activités telles que Clarity Campaigns, BlueLabs ou Civis Analytics pour n'en citer que certaines (Ibid.).

Sans parler de cas aussi « extrêmes », l'usage courant et normal d'internet et des réseaux sociaux peut poser parfois question. Ainsi, les résultats de recherche ou les contenus qui s'affichent dans le fil d'actualité d'un média social sont tout sauf neutres et résultent le plus souvent d'une personnalisation. Développé depuis 2011 par Eli Pariser dans son livre « *The Filter Bubble : What the Internet Is Hiding from You* » (Pariser, 2012), le concept de la bulle de filtres repose sur l'idée que les algorithmes « prédictifs », au cœur des moteurs de recherche et des réseaux sociaux trient et « personnalisent » les informations en fonction des centres d'intérêt de l'utilisateur allant dans le sens de ses choix habituels (Asan, 2021). C'est grâce à ces méthodes par exemple que Netflix propose notamment des recommandations ou qu'Amazon connait par avance le prochain achat. Ainsi, en ayant accès uniquement aux informations conformes à ses propres opinions, le citoyen risque de se confronter à un isolement idéologique lié à la réduction de son champ informationnel (Ibid.).

2.7 Un vecteur de « hard-power »

L'IA est un enjeu de sécurité nationale prioritaire pour les puissances militaires du XXIe siècle. Sans surprise, les États-Unis et la Chine sont aujourd'hui en tête de cette nouvelle « course aux armements » profitant de partenariats public-privé (Miailhe, 2018). Elle répond à trois contraintes opérationnelles : simplifier, automatiser, prédire (Noël, 2018).

Les États-Unis sont des leaders dans ce domaine, l'armée américaine est convaincue que l'IA révolutionnera l'ensemble des opérations sur le champ de bataille. Le Pentagone souhaite généraliser l'usage des technologies de l'IA au sein de l'armée (Kriegler, 2020). Pour les USA fidèles à leur doctrine, l'intérêt stratégique de l'IA est sans surprise de maintenir un écart entre les capacités des armées américaines et celles de leurs adversaires potentiels, pour empêcher tout risque de lutte à armes égales (Noël, 2018). La supériorité dans le domaine conventionnel, obtenue notamment grâce à l'IA, doit être telle qu'aucune autre puissance ne pourra envisager de défier les forces américaines sur le champ de bataille sans s'exposer à coup sûr à la débâcle.

L'US Army investigue néanmoins la possibilité d'un conflit sans être en situation de domination absolue consciente de la montée en puissance de la Chine. Dans ce contexte, elle étudie les options pour mettre en place ce qu'elle appelle des fenêtres de domination transitoires sur le champ de bataille (Kriegler, 2020). Le renseignement en conjonction de l'utilisation de l'IA pourrait aider à l'identification de telles opportunités sur les batailles à venir (Ibid.).

La victoire d'AlphaGo en 2016 contre les meilleurs joueurs humains a été une révélation pour les militaires chinois. Elle a prouvé tout le potentiel de l'IA, capable de mener seule des analyses complexes et d'établir des stratégies gagnantes en environnement complexe (Kania, 2019). Ainsi pour la Chine, l'IA est une technologie stratégique essentielle dans toutes les dimensions de la compétitivité nationale, avec le potentiel de transformer les paradigmes actuels de la puissance militaire. La décision de Pékin est de donner la priorité à l'IA pour améliorer son développement économique, et ses capacités militaires (Ibid).

Les stratèges militaires chinois anticipent une transformation de la forme et du caractère des conflits à venir. Ils anticipent une évolution de la guerre « informatisée » actuelle vers une future guerre « intelligente » avec un usage intensive des technologies issues de l'IA. De leur point de vue, il s'agit de relever les défis que l'IA représente ainsi que les opportunités historiques qu'elle permet (Ibid.). La Chine voit dans l'adoption des technologies de l'IA une opportunité unique de moderniser son armée, sa doctrine militaire et de combler son retard technologique vis-à-vis de l'armée américaine.

Ainsi, l'armée chinoise explore et expérimente activement de nouveaux concepts et capacités pour tirer parti de l'IA afin d'améliorer sa puissance de combat, ses stratégies et sa dissuasion.

De leur côté, les observateurs russes explorent de nombreuses voies pour mettre en échec les stratégies américaines et chinoises. Les experts russes conscients des limites de leurs moyens, savent qu'ils ne peuvent fournir des efforts sur une échelle comparable aux américains ou aux chinois. Aussi, le recours potentiel aux armes nucléaires stratégiques et tactiques demeure l'option préférentielle face aux agressions directes contre la Russie. Par l'ampleur des dommages potentiels causés, l'atome continue de balayer tous les avantages conventionnels que peuvent offrir l'IA. Les contre-stratégies efficaces ne passent pas forcément par davantage de technologie (Noël, 2018).

La France essaye de rester dans la course, mais les moyens consentis ne sont clairement pas là. Les niveaux d'investissement de la France pour l'IA militaire restaient en 2018 largement en deçà de ceux des États-Unis et de la Chine : 100 millions d'euros annuels de financement public annoncés, contre 1 à 3 milliards de dollars pour les États-Unis et 22 milliards de dollars prévus pour la Chine (59 milliards d'ici 2025). En outre, la réglementation sur la protection des données personnelles n'autorise pas la France à mener les collectes massives que se permettent la Chine et les États-Unis sur leur population – des données essentielles à l'apprentissage automatique des machines – (Thibout, 2018).

2.8 Les applications militaires

En reproduisant les processus cognitifs au moyen d'algorithmes et de traitements automatisés du big data, l'IA est capable d'effectuer un nombre grandissant de tâches spécifiques dans lesquelles elle surpasse les performances humaines. L'énorme potentiel de cette technologie n'a pas échappé aux organisations militaires de l'ensemble des pays développés.

L'IA permet en effet de gérer et simuler l'environnement opérationnel, de détecter des menaces, de traiter et simplifier les masses d'informations collectées et d'en livrer, dans le meilleur des cas, une analyse élémentaire (Kriegler, 2020).

L'IA peut être utilisée dans un vaste éventail d'applications militaires tant au niveau des missions de combat que du soutien aux opérations. La Chine, les USA (et bien sûr certains États développés ou qui aspirent simplement à garder leur souveraineté) poursuivent des recherches et des développements sur tous les aspects militaires imaginables (Kania, 2019). Sans être exhaustif, parmi les domaines qui font l'objet d'une recherche et de développements, nous pouvons citer :

- L'analyse de données massives pour :
    - La détection et le traitement automatisé des données (par exemple : le sons émis par des navires de guerre, des sous-marins, des avions ou des engins militaires peuvent être recueillis, analysés et attribués avec une précision supérieure à celle des meilleurs analystes humains).
    - L'aide à la décision aux commandants ou aux opérateurs de plates-formes spécifiques (avions de chasse et sous-marins).
    - L'appui aux opérations interarmées et l'amélioration de l'intégration et du traitement des informations pour la plateforme de commandement intégrée.
    - L'automatisation de l'analyse d'image notamment pour la reconnaissance faciale à large échelle (dans une foule par exemple).

- La maintenance des équipements via la prédiction des occurrences de panne sur les équipements pour ainsi simplifier la logistique des armées.

- L'étude des interactions homme-machine, en impliquant de nouveaux modèles pour améliorer la fiabilité et l'ergonomie des équipements.

- L'utilisation de réseaux de neurones pour le guidage des missiles de croisière afin de permettre leur plus grande autonomie pour le contrôle et le ciblage (génération de trajectoires alternatives). Dans la même veine, l'application des réseaux de neurones au guidage des véhicules planeurs hypersoniques pour permettre un contrôle plus précis ainsi que leur autonomie.

- L'amélioration des algorithmes de reconnaissance automatique de cibles (ATR) pour gagner en précision, notamment pour l'identification de cibles multiples en temps réel via l'utilisation de réseaux de neurones ainsi que le développement d'armes semi-autonomes ou autonomes.

- L'utilisation de l'apprentissage en profondeur (deep learning) et d'autres algorithmes pour modéliser la dynamique de l'attaque et de la défense en combat aérien libre. À titre d'illustration, l'armée de l'air américaine a expérimenté sur simulateur de vol des combats aériens avec des pilotes expérimentés contre une IA surnommé ALPHA (Noël, 2018). L'IA a systématiquement remporté la victoire dans ces simulations.

- De nouveaux algorithmes et architectures pour l'intelligence en essaim visant à permettre le "combat en essaim", dont une application évidente peut être la saturation des systèmes de défense antiaérienne à l'aide de vecteurs à bas coûts.

- Le développement de méthodes de modélisation et d'évaluation des équipements sans pilote pour en tester la fiabilité et la fonctionnalité. Et plus généralement, le développement de systèmes "sans pilote" comme des véhicules aériens, des véhicules terrestres, des navires de surface, ainsi que des sous-marins autonomes.

- L'exploration d'options pour les opérations psychologiques. Il est ainsi possible d'utiliser des réseaux antagonistes génératifs exploités pour la manipulation d'images et la mise en place de stratégies de désinformations à travers les « deep fakes ».

- L'application de la réalité virtuelle et augmentée pour la simulation et l'entraînement au combat réel. L'IA peut utilement être utilisée dans le « war-gaming » comme outil d'entraînement et d'évaluation des dynamiques d'affrontement intelligent.

- L'utilisation des technologies de l'IA pour la cybersécurité et la cryptographie, y compris en stéganographie avancée. Les applications de ces technologies sont évidentes pour l'amélioration des communications et sécuriser les réseaux contre le brouillage. En outre, les techniques de cyberdéfense ou cyberattaque peuvent être utilisées pour contrer ou subvertir les systèmes d'IA d'un adversaire via la manipulation de données et/ou l'exploitation de vulnérabilités matérielles.

- L'utilisation du traitement automatique du langage naturel pour l'analyse dans le renseignement militaire, de même que le développement de systèmes portables destinés à améliorer la connaissance de la situation et la prise de décision sur le champ de bataille.

On le voit le champ d'application militaire de cette technologie n'a de limites que l'imagination humaine, qui en la matière est toujours très fertile. Mais sans surprise, certains développements ont – bien évidemment - une application à la fois militaire ou civile. Cela est rendu d'autant plus simple que l'ensemble des algorithmes d'IA sont dans le domaine public et disponibles en « open source ».

2.9 Les applications dans la lutte contre le terrorisme

L'IA s'applique évidemment dans le cadre de la gestion des conflits armés asymétriques et par extension dans le cadre de la lutte contre le terrorisme. L'analyse d'informations à partir des données textuelles, sonores ou encore vidéos, provenant par exemple des réseaux sociaux, permet de faciliter la surveillance des comportements suspects. Le programme DEXTER met au point un prototype utilisant l'IA pour la détection des explosifs et des armes à feu, sur les lieux très fréquentés tels que les métros ou les aéroports (OTAN, 2022).

En 2021, l'ONU a publié un rapport intitulé *"Countering Terrorism Online with Artificial Intelligence"* (A Joint Report by UNICRI and UNCCT, 2021) qui décrit les cas d'usage pour la lutte contre le terrorisme. Parmi les applications, ce rapport cite :

- L'analyse prédictive des activités terroristes en se basant sur les données collectées sur les réseaux sociaux afin d'analyser les comportements et prévenir les actes terroristes. L'IA sert ici à analyser des masses de données ouvertes en ligne.

- L'identification des signaux d'alarme mettant en avant une radicalisation. La détection s'appuie sur le traitement automatique du langage comme le cas précédent et l'analyse comportemental qui implique un suivi dans le temps.

- La détection de la désinformation et des « fake news » diffusées par les terroristes à des fins stratégiques, l'IA permettrait de détecter les faux comptes ou l'utilisation des bots sur les réseaux sociaux

- La lutte contre les récits terroristes et extrémistes par le biais de l'IA dans le traitement automatique du langage.

2.10 Les applications en police prédictive

La police prédictive est également un champ d'application de l'IA, ce qui ne va pas sans poser de questions d'ordre philosophique ou sociétale. Sommes-nous ainsi prêts à sacrifier un peu de notre droit à la vie privée, de notre liberté pour plus de sécurité et de « justice » ?

L'idée est de prédire un crime et sa localisation avant même qu'il ne se produise. À titre d'exemple, nous pouvons citer une étude publiée par des chercheurs aux États-Unis décrivant l'utilisation de l'IA pour la prédiction de la localisation des crimes et de leurs typologies dans la ville de Chicago (Rotaru, et al., 2022), (Communication Tri Duc Tran, 2022). Les chercheurs ont pu mettre en place un algorithme permettant de prédire une semaine à l'avance la survenance d'un crime et sa localisation approximative, avec une exactitude de 90 %. Les chercheurs ont utilisé pour cela les données d'environ 350 000 crimes commis entre janvier 2014 et décembre 2016. Les capacités de prédiction de l'IA sont également utilisées dans le domaine de la gestion des récidives relative aux comportements violents (Berly, et al., 2022)

Dans le document de travail du Centre d'Innovation d'Interpol (INTERPOL, 2022), les outils d'IA font partis des réflexions pour l'avenir de l'action policière concernant notamment :

- La prévention de la criminalité afin de déterminer à quels endroits déployer des patrouilles ou identifier les personnes jugées les plus susceptibles de commettre des infractions, de récidiver ou d'être victimes d'infractions.

- Les enquêtes et la criminalité à l'ère numérique avec notamment l'utilisation de la reconnaissance faciale, l'usage des scanners biométriques portatifs et de l'imagerie 3D.

- De nouvelles relations publiques s'appuyant par exemple sur des systèmes de conversation (chatbots ou voicebot) afin de fournir des services plus rapides et plus adaptés à la population, tout en réduisant les pressions sur des ressources limitées.

2.11 Les applications civiles

L'IA permet aussi de détecter des signaux faibles pour développer des systèmes d'alertes précoces. Cette propriété devrait permettre une gestion des risques plus efficace dans divers domaines de l'activité humaine (Communication Tri Duc Tran, 2022).

Ainsi, d'après l'agence des Nations unies pour le développement spécialisée dans les technologies de l'information et de la communication (ITU), les Nations unis soutiennent directement 228 projets relatifs à la sureté (ITU, 2021), dans divers secteurs stratégiques, comme :

- Le nucléaire civil où l'application de l'IA servirait à prévenir les accidents et à contrôler le bon fonctionnement des centrales nucléaires en surveillant les différents paramètres liés au combustible radioactif (Ibid., p.33).

- L'agroalimentaire avec la lutte contre les insectes ravageurs (Ibid., p.24) via l'analyse d'images, en provenance par exemple de téléphones mobiles. Citons également, l'analyse de l'impact de la crise sanitaire de la Covid-19 sur les chaînes d'approvisionnement afin d'éviter de potentiels crises alimentaires. Il y a également la lutte contre les famines (Ibid., p.184) à travers des systèmes de surveillance et de visualisation de données.

- L'anticipation des catastrophes naturelles et des crises climatiques (Ibid., p.94) en utilisant des données satellites et démographiques et les prévisions météorologiques.

- La lutte contre les fraudes électorales et la désinformation (Ibid., p.89), l'outil IVerify est déployé au Zimbawe. (https://iverify.org.zm/)

- La lutte contre la désinformation en analysant les données des réseaux sociaux comme pour la crise de la Covid-19 (ibid., p.93) ou les mythes et les fausses croyances sur la contraception (Ibid., p.114).

- La gestion des pandémies : analyse de la dynamique de la propagation de la Covid-19 (ibid., p.119), du choléra (Ibid., p.191) à partir des données démographiques et satellites.

- La gestion de la sécurité des flux migratoires (Ibid., p.132) en analysant automatiquement les tendances concernant les contenus haineux ou discriminatoires.

- La surveillance des cultures illicites (Ibid., p.163) à partir des images satellites ou encore la surveillance et la visualisation des trafics de drogues.

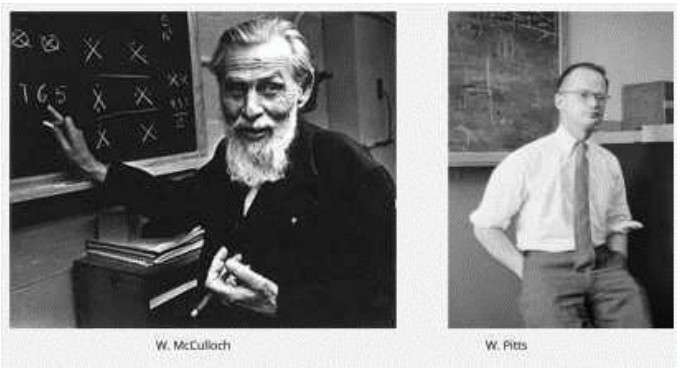
Cette liste est loin d'être exhaustive, mais elle traduit la richesse et la puissance de cette technologie et les impacts qu'elle a et aura sur notre vie. La seule limite reste l'imagination humaine.

3 LES PRINCIPAUX CONCEPTS DE L'IA

L'objectif de ce chapitre, n'est pas de produire un cours sur l'IA ou de reprendre de manière exhaustive tous les algorithmes existants dans le domaine. Nous tenterons simplement de donner une idée la plus exacte possible des principaux concepts utilisés.

3.1 Quelques considérations autour de l'IA

Au carrefour des sciences mathématiques, informatique, neurobiologique, logique, statistiques et cybernétique, les concepts de l'intelligence artificielle (IA) sont apparus dès les années 40 (McCulloch & Pitts, 1943). En étudiant les analogies entre les machines numériques et le cerveau humain, Mc Culloch et Pitts furent les premiers à proposer un modèle mathématique simple, capable de reproduire le comportement essentiel du neurone biologique à savoir sa capacité de s'activer à partir d'un seuil de potentiel électrique.



Mais, la véritable rupture vient des travaux de Frank Rosenblatt (1928 – 1971) du laboratoire d'aéronautique de l'université de Cornell, et sa publication de 1958 *« The Perceptron : a probabilistic model for information storage and organization in the brain »*, dans Psychological Review. Frank Rosenblatt est le concepteur du perceptron, la première machine capable d'apprendre en simulant de manière simplifiée un neurone.

Depuis, les chercheurs du monde entier mènent des recherches en IA afin d'effectuer des tâches nécessitant une « capacité de réflexion » de « raisonnement », de manière autonome et adaptable. De nos jours, l'IA est utilisée dans tous les secteurs d'activités, même dans notre vie quotidienne. À titre d'exemple, une recherche sur Google implique des mécanismes de deep learning (Yonnet, 2022).

C'est à partir de 2010 que l'IA commence véritablement à percer dans tous les domaines, des mots comme apprentissage automatique, machine learning, deep learning ou encore réseaux de neurones réservés jusque-là au monde de la recherche passent dans le langage du grand public. Le machine learning et plus précisément le deep learning arrivent à surpasser l'être humain dans certaines tâches spécifiques.

Utilisée en entreprise, l'IA a pour objectif d'améliorer les performances et la productivité en contribuant à l'automatisation des processus ou des tâches. En théorie, dès qu'une tâche nécessite l'intervention d'un être humain, l'IA pourrait le remplacer. De façon surprenante, en septembre 2022, en Chine, une IA est devenue CEO de la filiale spécialisée dans le jeu vidéo de l'entreprise NetDragons (Challand, 2022).

Pour ce qui concerne plus spécifiquement la cybersécurité, les cyberattaquants intègrent déjà cette technologie, les attaques utilisant l'IA sont en augmentation (Guembe, et al., 2022). En réponse à cette tendance, l'intégration de l'IA peut être utilisée dans la gestion des risques cyber, notamment en contribuant à détecter et arrêter les attaques. L'IA fournit une aide précieuse pour automatiser l'analyse des flux de données en temps quasi-réel, elle est utilisée en cybersécurité pour identifier les

comportements à risques et prévoir les attaques potentielles d'un système d'information. Dans une stratégie de défense en profondeur, l'IA peut apporter une aide sur chaque couche de défense.

3.2 L'IA forte

Précisons tout de suite et avant même toute définition, qu'à l'heure actuelle, l'IA forte n'existe pas. Et, il n'est absolument pas démontré que la création d'une telle IA soit même possible. L'avènement de cette IA transformerait notre rapport homme-machine et bouleverserait le devenir de notre espèce. Les impacts philosophiques et anthropologiques seraient considérables.

L'IA forte serait capable d'accomplir l'ensemble des tâches que réalisent les humains et voir au-delà. Elle pourrait proposer une aide à la décision sur des affaires stratégiques, en prenant en compte de manière exhaustive un nombre de données considérable appartenant à des domaines de plus en plus vastes (Noël, 2018). Cette IA ne serait théoriquement pas soumise aux mêmes contraintes et biais que les êtres humains.

Une intelligence forte regrouperait l'intelligence analytique, déjà présente dans l'IA faible, et les intelligences situationnelle et émotionnelle (Jean, 2020). Cette forme d'IA pourrait s'adapter à plusieurs domaines, à plusieurs situations, ses caractéristiques seraient :

- La capacité de raisonner, de résoudre des problèmes, de porter des jugements, de planifier, d'apprendre et de communiquer. Il s'agit en quelque sorte des capacités cognitives permettant la manipulation de concepts appris du monde extérieur.

- Avoir des « pensées objectives », une « conscience de soi », une « sensibilité ». Pour résumer, ce sont les capacités à avoir une réflexion intérieure propre sur soi et une méta analyse de son existence. Nous laissons ici volontairement les implications philosophiques et les difficultés associées. Disons simplement, que le développement d'une IA forte démontrerait, au moins de manière indirecte, que la conscience n'implique aucun processus psychologique indépendant, distinct du cerveau. En d'autres termes, la conscience ne se situerait pas au-dessus et au-delà du fonctionnement physique du cerveau, elle émergerait de la structure physique même de ce dernier.

3.3 L'IA faible

L'IA faible est présente dans les outils que nous utilisons dans notre vie quotidienne, elle se concentre sur une tâche, un domaine, ou une problématique spécifique. Ses performances peuvent dépasser largement les capacités d'un humain effectuant les mêmes tâches.

Nous pouvons cumuler les différents types d'IA faible au sein d'une même "machine" comme la reconnaissance vocale, la reconnaissance des objets ou encore l'analyse textuelle. Mais chaque capacité est indépendante et programmée d'une manière spécifique. Nous ne pouvons cumuler les IA faibles pour obtenir une IA forte.

Tous les systèmes d'IA ne sont que des IA faibles. Tout au long de ce mémoire quand nous faisons référence à l'IA, nous nous référons à la notion de l'IA faible, qui en dernier ressort n'est qu'un outil informatique adapté à un problème particulier dans un contexte donné.

3.4 L'IA comme outil de modélisation des processus cognitifs

L'IA modélise les cinq fonctions cognitives suivantes (Ganascia, 2017) :

- **Les fonctions réceptives** : elles autorisent l'acquisition, le traitement, la classification et l'intégration de l'information. Les champs d'applications possibles sont :
  - L'interprétation automatique d'images : la reconnaissance de formes, la reconnaissance d'images, l'identification d'objets, les moteurs de recherche dans des bases d'images.
  - L'Interprétation automatique de vidéos : détection d'intention.
  - Le traitement de la parole et du texte : compréhension du langage, traduction, identification de l'interlocuteur.
  - L'analyse et fusion d'informations issues de capteurs : « désambiguïsation » de l'image par le son.

- **La mémoire et l'apprentissage** permettant le stockage et le rappel de l'information, les champs d'applications possibles sont :
  - La représentation des connaissances.
  - Les logiques de description (graphes de données).
  - L'extraction de connaissances.

- **Le raisonnement** : cette fonction concerne aussi l'organisation ou la réorganisation de l'information ainsi que son utilisation. Les champs d'applications possibles sont :
  - Les inférences (raisonnement logique, démonstration de théorèmes).
  - Les retours d'expériences (traçabilité, capitalisation de connaissances).

- **Les fonctions exécutives de prise de décision et d'action** comme :
  - Les systèmes autonomes et/ou semi-autonomes.
  - L'aide à la décision.
  - La planification des tâches.

- **Les fonctions expressives** qui rendent possibles la communication avec notamment :
  - Le langage naturel (écrit/oral) (traitement du langage, système question-réponse).
  - Les interfaces homme-machine.

Dans une approche orientée opérationnelle, l'utilisation de l'IA est pertinente et est exploitée (dans un contexte industriel ou grand public) pour la résolution de problème, la gestion des connaissances et le raisonnement (Haton, 2000) :

- **La résolution de problèmes** : il s'agit de concevoir des stratégies efficaces d'exploration d'espaces de solutions souvent très vastes comme pour les jeux de réflexion.

- **La reconnaissance et l'interprétation de données** : ces données peuvent être de nature très variée et, par suite, nécessiter des traitements très différents (informations symboliques, signaux temporels, images, etc.). Les applications pratiques ont trait à la reconnaissance de l'écriture (lecture optique de textes), le traitement d'images (télédétection, biomédical, industrie), le diagnostic (médical, industriel, financier), la surveillance et la conduite de procédés industriels, la compréhension de signaux industriels ou biomédicaux, etc.

- **L'aide à la décision** : le but est d'aider un décideur humain dans les choix qu'il a à faire en présence d'informations diverses, hétérogènes et incertaines. Comme nous l''avons vu, les applications se trouvent dans tous les domaines : bancaire, financier, industriel, médical,

militaire, etc. Dans ce cadre l'IA permet de modéliser des connaissances pour aider à la prise de décision efficace et rapide.

- **La planification d'actions et la robotique** : il s'agit de définir précisément la suite d'actions élémentaires permettant de mener une tâche complexe, telle que celle que doit mener un robot mobile dans un environnement plus ou moins bien connu.

- **Le traitement du langage naturel, écrit et parlé** : des progrès importants ont été faits dans ce domaine complexe. Même si le problème est loin d'être résolu, des applications réelles existent comme la traduction de textes techniques, l'analyse et l'indexation de documents écrits (moteurs de recherche sur internet), la reconnaissance de la parole (dictée vocale de textes, accès à des informations par télématique vocale). Avec les moteurs de recherche et les chatbots, l'IA permet de traiter des données textuelles orales et écrites, dans le seul but de répondre de manière quasi instantanée aux requêtes diverses.

- À cela nous pouvons ajouter **les capacités de prédiction** qui est un domaine adjacent à la reconnaissance et l'interprétation de données.

Dans une approche de méta analyse, les apports de l'IA pour la gestion de risques s'appuient sur :

- La reconnaissance et l'interprétation des données
- L'aide à la décision
- Le traitement du langage naturel
- La prédiction

3.5 Revue des domaines de l'IA

L'IA ne se réduit pas uniquement au machine learning et au deep learning. Cette association rapide est liée au fait que la majorité des systèmes de deep learning sont extrêmement performants et sont déployés dans les environnements de production.

L'IA s'appuie sur deux approches, une approche symbolique basée essentiellement sur une modélisation des règles et une approche statistique dans laquelle le machine learning est largement utilisé dans les systèmes en production actuellement.

3.6 L'IA symbolique

Du point de vue historique, le symbolisme remonte à la fin des année 1950 et regroupe des chercheurs travaillant sur la résolution de théorèmes comme Herbert Gelernter (créateur d'un système de démonstration de théorèmes géométriques en 1959), Allen Newell et Herbert Simon (inventeurs du General Problem Solver en 1957), John McCarthy (inventeur de l'appellation d'intelligence artificielle en 1955 ainsi que du langage LISP en 1958), James Slagles (concepteur du premier système expert, SAINT, de traitement de formules mathématiques, en 1961), Thomas Evans (et son programme ANALOGY de 1963 qui pouvait résoudre des problèmes de tests de QI) et aussi les Français Alain Colmerauer et Philippe Roussel (inventeurs du langage Prolog en 1972).

L'IA symbolique s'appuie sur un système de "règles" de type « Si X Alors Y », ce sont les systèmes experts. Ces règles sont construites par des experts métiers puis intégrées aux machines afin de guider une prise de décision autonome. Il s'agit de modéliser les concepts/les symboles manipulés par l'esprit humain, afin de générer un modèle explicable et donc prouvable.

Nous parlons alors de « boîte transparente » ou modèle explicable par opposition à la « boîte noire » créé par les algorithmes de machine learning. Les outils utilisés par l'IA symbolique se basent sur des moteurs d'inférence manipulant les mécanismes de la logique formelle. Pour illustrer le propos, dans le traitement du langage naturel, l'approche symbolique consiste à s'appuyer sur la modélisation de la grammaire formelle ainsi que des techniques d'analyse syntaxique. Les règles grammaticales et les données lexicales sont alors codées dans les systèmes "IA".

Dans l'IA symbolique, nous retrouvons, par exemple, les systèmes experts qui s'appuient sur une modélisation de la connaissance d'un domaine spécifique auprès de spécialistes du domaine métier.

Tout système expert est composé de trois (3) principaux éléments :

- Une base de règles spécifiques au domaine métier : les connaissances sont implémentées sous une forme déclarative : SI "La condition est vraie" Alors "l'action à déclencher est…".

- Des données d'entrée nommées « base de faits » qui servent au raisonnement et auxquelles s'appliquent les règles logiques.

- Un moteur d'inférence qui applique les règles sur les données en entrée. L'ensemble constitué du modèle de faits et de la base de règles est aussi appelé la base de connaissances.

Dans le monde industriel les systèmes experts se retrouvent sous la dénomination de BRMS (Business Rules Management System) ou DMS (Decision Management System).

L'enjeu majeur est la modélisation des règles avec un expert métier. Il s'agit d'un travail de coopération entre l'expert de l'outil d'inférence et l'expert métier pour déterminer avec exhaustivité les règles applicables.

Par exemple, l'usage des systèmes experts peut s'appliquer au calcul des risques de dossiers de crédits immobiliers, à la lutte contre la fraude et le blanchiment d'argent, au calcul de primes, de retraites, du prix de billets de train, aux plateformes de jeux, etc.

Les avantages d'un système expert reposent sur :

- L'explicabilité : le raisonnement peut être tracé et expliqué.
- L'exploitation simplifiée : la transmission entre l'équipe de projet et l'équipe de production est facilitée par le fait que les règles sont lisibles. Cependant, cela peut être un inconvénient sur de grandes bases de connaissances pour lesquelles plusieurs experts sont intervenus sur des temporalités différentes.

Les limites sont :

- La capacité d'apprentissage automatique : les règles ne sont pas créées automatiquement, l'intervention humaine est nécessaire pour enrichir la base de connaissances.
- Une inadaptation à la reconnaissance, la détection des formes (images, vidéos) ou au traitement automatique des langues.

Parmi les systèmes experts qui ont marqué l'histoire, nous citerons Deep Blue d'IBM qui en 1997 a battu le champion du monde d'échecs Garry Kasparov. L'algorithme fonctionnait sur un supercalculateur parallèle IBM capable de calculer quelques 200 millions de positions par seconde. Deep Blue s'appuyait sur un algorithme systématique de force brute, où tous les coups envisageables étaient évalués et pondérés (GEO, 2022).

3.7 L'IA statistique

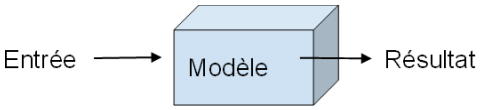
L'IA statistique s'appuie comme son nom l'indique sur des modèles probabilistes qui sont utilisés par les systèmes informatiques pour effectuer une tâche spécifique, sans utiliser d'instructions explicites, en se basant sur des modèles et des inférences. Les règles sont créées via une intervention humaine, puis codées « en dur » dans un programme statique.

Les algorithmes de machine learning construisent un modèle mathématique basé sur des échantillons de données, appelés « ensemble d'apprentissage ». L'algorithme apprend des règles en établissant des corrélations entre les entrées et les sorties.

Actuellement, la majorité des applications en production contenant de l'IA relèvent de l'IA statistique plus spécifiquement du machine learning avec des modèles de type réseaux de neurones. Le machine learning couvre un large spectre de méthodes mathématiques telles que les classificateurs Bayésien, les régressions/discrimination linéaires ou quadratiques, les arbres de décision, les forêts aléatoires, l'analyse en composantes principales (ACP), les plus proches voisins (K-NN), le « support vector machine » (SVM), les algorithmes génétiques ou encore les réseaux de neurones (qui se déclinent en différentes architectures comme les réseaux simples, les réseaux récurrents, les réseaux de convolution, etc.). Cette liste n'est – bien évidemment – pas exhaustive.

3.8 L'apprentissage machine

L'apprentissage machine ou machine learning s'appuie sur des données afin de calculer un modèle. Pour simplifier, un modèle contient un ensemble de règles permettant de trouver la réponse désirée à partir d'une entrée. C'est une approche d'analyse statistique, à partir d'un grand ensemble de données, des règles implicites sont extraites automatiquement et sont encapsulées dans un modèle numérique.



L'objectif est de construire un modèle mathématique qui prédit des valeurs (résultats) à partir des données d'entrée, et de minimiser l'écart entre ces valeurs prédites et les valeurs réelles. Pour ce faire, un algorithme d'optimisation d'une fonction erreur est utilisée (comme la méthode du gradient).

Il existe plusieurs méthodes d'apprentissage en fonction du type de données disponibles ou de l'objectif à atteindre :

- **L'apprentissage supervisé** : la détermination du modèle s'appuie sur des données dites annotées de leurs sorties pour l'entraînement. Lorsque le modèle est entraîné, d'autres jeux de données annotés (classés) sont utilisés pour déterminer les performances du modèle. Le travail d'annotation ou d'étiquetage/classification des données est primordial.

- **L'apprentissage non supervisé** : les données en entrée ne sont pas annotées. L'algorithme d'entraînement s'applique dans ce cas à trouver seul les similarités et les distinctions au sein de ces données, et à regrouper ensemble celles qui partagent des caractéristiques communes.

- **L'apprentissage semi-supervisé** : ce type d'apprentissage est une combinaison de l'apprentissage supervisé et non supervisé, où seule une petite partie des données sont annotées. Le modèle utilise les données étiquetées pour apprendre les règles et les structures générales des données, puis utilise ces connaissances pour prédire les étiquettes des données non étiquetées.

- **L'apprentissage par renforcement** : ce type d'apprentissage est adapté pour résoudre des problématiques de prises de décisions séquentielles (comme pour les jeux, la planification de

tâches, le contrôle de voitures autonomes). La méthode consiste à récompenser les comportements souhaités et à sanctionner les comportements non désirés. L'objectif est de permettre à un agent (entité virtuelle : robot, programme, etc.), placé dans un environnement interactif (ses actions modifient l'état de l'environnement), de choisir des actions maximisant des récompenses quantitatives en essayant toutes les façons possibles et en apprenant de ses erreurs. Les données proviennent donc de l'environnement réel ou simulé. Notons que l'apprentissage par renforcement peut également être combiné avec d'autres méthodes d'apprentissage telles que l'apprentissage supervisé ou semi-supervisé pour améliorer les performances du modèle en utilisant des données étiquetées pour guider l'apprentissage.

Les algorithmes de machine learning sont utilisés pour opérer les fonctions suivantes :

- **La Régression** consiste à trouver des corrélations entre différents ensembles de données, ce qui permet de comprendre comment les données sont liées les unes aux autres. Nous pouvons utiliser la régression pour faire de la prédiction ou encore pour trouver les causes des anomalies.

- **Le Clustering** permet d'analyser des ensembles de données et de les regrouper en fonction de leurs caractéristiques générales. Le clustering fonctionne directement avec de nouvelles données sans tenir compte des exemples précédents. C'est de l'apprentissage non supervisé.

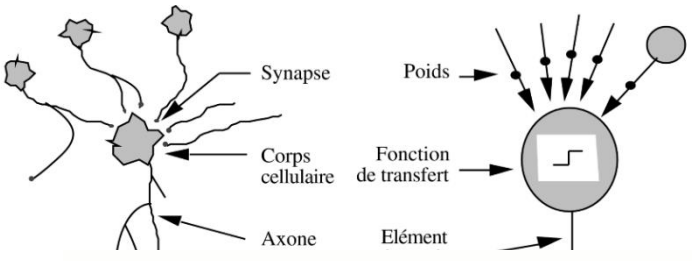
- **La Classification** consiste à classer/catégoriser les données. Par exemple, classer un fichier binaire dans des catégories telles que logiciels légitimes, logiciels publicitaires ou rançongiciels.

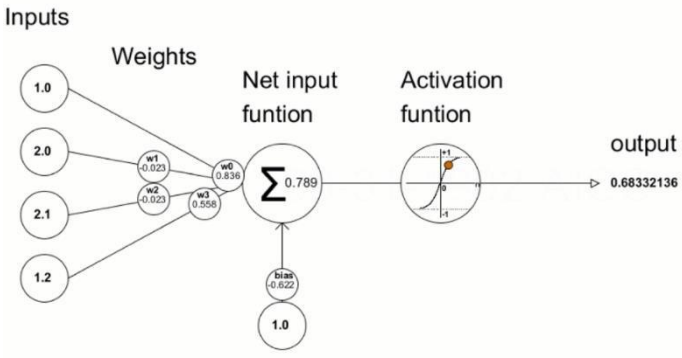
Les méthodes d'apprentissage supervisé sont adaptées lorsqu'on dispose d'une connaissance sur la problématique à résoudre. Par exemple, en reconnaissance faciale nous avons au préalable l'information de l'image associée à un individu ou un objet, l'objectif est d'apprendre à l'automate à identifier l'individu ou l'objet parmi une multitude d'images.

Les méthodes d'apprentissage non supervisé sont adaptées lorsque aucune connaissance sur les données ne sont disponibles. Par exemple, nous avons des images de voitures, d'avions et de bateaux, l'objectif est de regrouper les photos selon ces trois catégories.

3.9 Le deep learning

Le deep learning est une variante de l'IA statistique. Nous retrouvons par exemple son usage dans la reconnaissance faciale ou dans la reconnaissance de texte. Le deep learning est une des méthodes de machine learning qui repose sur un modèle de réseaux de neurones multicouches. Dans les faits, un réseau de neurones dispose d'une couche d'entrée, d'une ou plusieurs couches cachées et d'une couche de sortie.

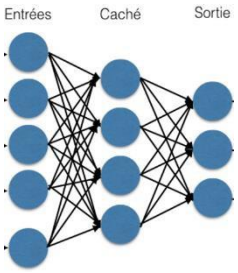




Un réseau de neurones est une interconnexion de neurones formels (Cf. illustration ci-contre). La puissance d'un tel système réside dans le fait qu'il peut approximer n'importe quelle fonction non-

linéaire, en jouant sur le nombre de neurones, les connexions et leur poids. N'importe quelle fonction mathématique peut ainsi être modélisée à partir d'un réseau de neurones. Ceci explique la capacité de ces structures mathématiques à modéliser le monde.

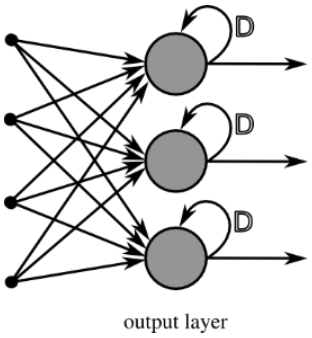
3.9.1 Les réseaux de neurones à propagation avant



Pour simplifier, l'entraînement d'un réseau de neurones consiste pour un jeu de données en entrée (dont nous connaissons a priori les valeurs de sortie à trouver) à comparer les résultats calculés par le réseau de neurones avec les valeurs à trouver. Les poids sont ajustés puis le calcul est réitéré jusqu'à ce qu'il n'y ait plus de progrès dans la réduction des erreurs observées. L'entraînement a pour objectif de minimiser l'écart entre la sortie attendue et la sortie effective (celle fournie par le réseau de neurones). Cet écart peut être calculé par la somme du carré des erreurs entre la sortie effective et la sortie attendue. Il existe néanmoins d'autres métriques de l'erreur, on parle de *loss function*. Il s'agit de minimiser cette fonction de perte.

Ces « Feedforward Neural Network » (FFNN) en anglais, sont les réseaux les plus simples. Chaque perceptron d'une couche est connecté uniquement à l'ensemble des perceptrons de la couche qui le suit immédiatement.

3.9.2 Les réseaux de neurones récurrents



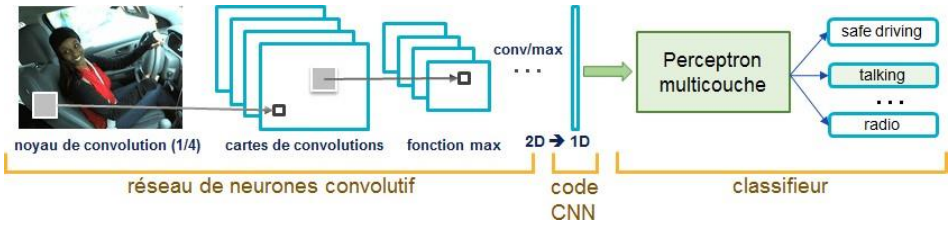
Les « Recurrent Neural Network » (RNN) (en anglais) sont une classe de réseaux de neurones qui travaillent à partir de données séquentielles brutes, et apprennent des motifs dans la succession des états présentés. Pour ce faire, Un RNN présente des connexions récurrentes entre les neurones de chaque couche. Ils se distinguent donc des réseaux de neurones traditionnels (FFNN) en ce sens qu'ils utilisent des boucles de rétroaction qui permettent aux informations de persister. La boucle de rétroaction signifie que les sorties calculées précédemment sont utilisées en même temps que l'entrée courante pour produire une prédiction. Les informations calculées par le RNN sont en quelque sorte stockées en mémoire pour être réutilisées. Ce mécanisme les rend donc particulièrement bien adapté pour traiter des données séquentielles en tenant compte des informations précédemment vues dans la séquence pour prédire la suite. Par exemple, dans le cas d'une application de type reconnaissance de la parole, le modèle évalue chaque mot d'une phrase, l'un après l'autre, mais en gardant en mémoire les mots qu'il a déjà traité. Ceci permet au moins dans une certaine mesure, au modèle de comprendre le contexte de chaque mot et sa place dans la phrase car les mots précédemment traités vont être analysés de manière récurrente. Les RNN sont ainsi utilisés dans la génération automatique de texte, en reconnaissance automatique de la parole ou de l'écriture manuscrite - plus généralement en reconnaissance de formes - ou encore en traduction automatique. Ils présentent le défaut majeur d'être lents et difficiles à entrainer.

3.9.3 Les réseaux de neurones à convolution

Les « Convolutional Neural Network » (CNN). Ce sont des réseaux de neurones qui présentent en amont d'un réseau FFNN plusieurs couches de convolution (au sens produit de convolution[6]). Ces réseaux sont particulièrement bien adaptés à la reconnaissance d'images. Ces filtres de convolution fonctionnent en quelque sorte comme un extracteur de caractéristiques des images. Une image est

[6] Le produit de convolution généralise l'idée de moyenne glissante et est la représentation mathématique de la notion de filtre linéaire.

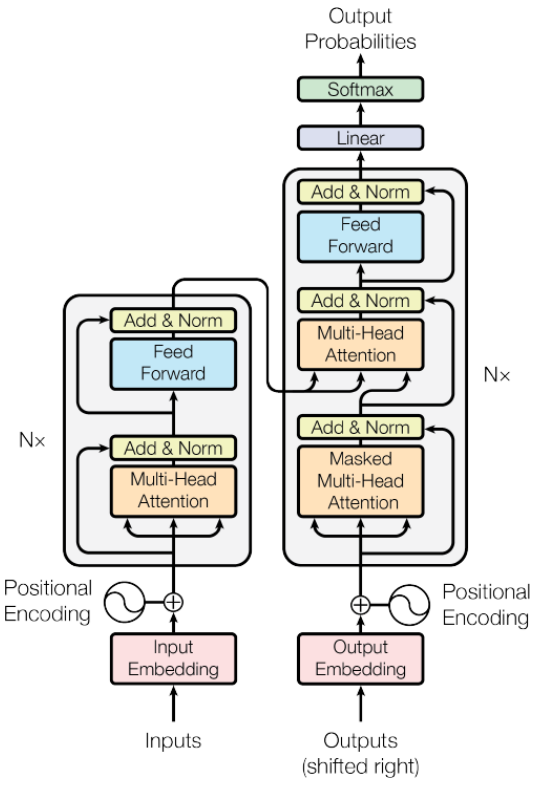
ainsi passée à travers une succession de filtres, ou noyaux de convolution, créant de nouvelles images. La partie convolutive est ensuite branchée en entrée d'une deuxième partie, constituée de couches entièrement connectées (FFNN). Le rôle de cette partie est de combiner les caractéristiques extraites pour classer les images, par exemple. De même, il existe de multiples variantes architecturales, que nous ne détaillerons pas ici.



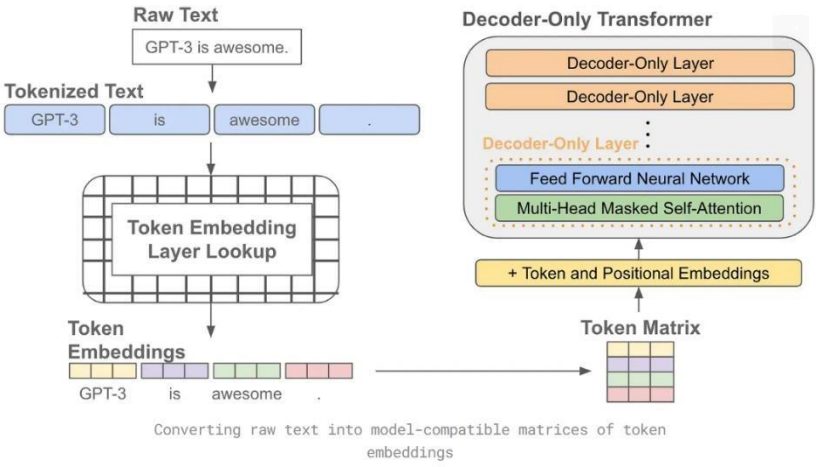
Ci-dessus un exemple d'architecture de CNN d'après : https://blog.octo.com/classification-dimages-les-reseaux-de-neurones-convolutifs-en-toute-simplicite/

3.9.4 Les « large langage model » ou LLM

Pour clore cette section, nous ne pouvions pas ne pas évoquer les LLM ou « large langage model », tellement l'actualité relative à l'IA est « saturée » par ces nouveaux algorithmes de NLP (natural language processing) qui visent à modéliser la structure du langage et finalement à calculer la sortie ou réponse la plus probable à une phrase ou question en tenant compte de la succession du contexte et des mots. Cette nouvelle classe de modèles constitue une véritable révolution conceptuelle et marque selon toute vraisemblance l'avènement de la généralisation des usages de l'IA dans tous les domaines de l'activité humaine, car ces modèles dépassent largement le cadre de la NLP à présent. Cette révolution remonte à Décembre 2017 avec la publication d'un article rédigé par une équipe de chercheurs de Google intitulé *« Attention is All you Need »* (Ashish Vaswani, 2017). Les idées contenues dans cet article qui se focalisait essentiellement sur des applications de traduction de textes, sont à l'origine des outils d'IA populaires comme ChatGPT, GitHub Copilot et quelques autres. L'article décrit essentiellement deux concepts, la notion de « transformer » et la notion « d'attention ». Pour simplifier, un transformer est une architecture de réseaux de neurones qui se compose d'un encodeur et d'un décodeur. L'attention est une fonction mathématique dont l'objectif est de modéliser le contexte d'un mot à partir des autres mots de la phrase et de leur position. Plus généralement, l'attention est une fonction mathématique qui permet au modèle de comprendre les relations entre les différents éléments d'une séquence. Le transformer est composé d'un encodeur qui prend en entrée une séquence de mots et la transforme en une représentation vectorielle. Cette représentation vectorielle contient des informations sur chaque mot dans la séquence et leur relation les uns aux autres. Le décodeur est la seconde partie du transformer qui utilise la représentation vectorielle générée par l'encodeur pour générer une sortie. Le décodeur prend en entrée un état caché et une séquence de mots générée précédemment pour produire un nouveau mot dans la séquence de sortie. L'encodeur et le



décodeur peuvent être parallélisés car les calculs pour chaque item dans la séquence d'entrée et de sortie sont indépendants. Par conséquent, les architectures de réseau de neurones qui utilisent des encodeurs et des décodeurs peuvent utiliser le calcul parallèle pour réduire les temps d'apprentissage. Ceci représente un très grand progrès par rapport au réseau de neurones récurrent dont la phase d'apprentissage est séquentielle. À titre d'illustration, nous présentons ci-dessus l'architecture du transformer telle que présentée dans l'article de l'équipe de google » (Ashish Vaswani, 2017) . le schéma montre que l'encodeur encode les données puis y applique une fonction auto-attention et une normalisation afin de mémoriser le contexte. L'encodage est réalisé par un réseau de neurone classique (FFNN). Les résultats sont ensuite transférés au décodeur qui est composé lui aussi d'une fonction d'attention et d'un réseau de neurones standards. L'empilement en couche de transformer permet d'améliorer les résultats d'apprentissage. Les LLM modernes utilisent des procédures de préapprentissage génériques pour résoudre une grande variété de tâches sans nécessiter d'adaptation en aval (c'est-à-dire pas de modifications architecturales, d'ajustements, etc.). Il existe de multiples variantes de l'architecture transformer. Par exemple dans le cas de GPT, ce dernier utilise une architecture de type « decoder-only ». Au-delà de ces couches de décodeurs, l'architecture GPT contient des couches d'intégration qui stockent des vecteurs correspondant à tous les jetons possibles dans un vocabulaire de taille fixe. Nous présentons ci-contre l'architecture de GPT avec ses multiples couches de décodeurs constitués d'une fonction d'attention et d'un réseau de neurones. Le schéma montre une unique étape d'encodage des entrées. Par la suite, l'architecture GPT 3 correspond à l'empilement de 96 couches de décodeurs pour environ 175 milliards de paramètres à optimiser (Wolfe, 2022). La force de ce type d'architecture réside dans sa capacité à modéliser et mémoriser un corpus de textes immenses ainsi que dans l'utilisation de modèles pré-entrainés pour réaliser des tâches plus spécifiques. Les LLM modernes peuvent être réutilisés comme un outil pour créer des modèles de base génériques qui résolvent divers problèmes sans avoir besoin d'adapter ou d'affiner le modèle. Le fait de réutiliser des modèles existants permet d'accélérer la phase d'apprentissage. Ces algorithmes ne se limitent pas à la modélisation du langage ou à la réalisation d'assistants vocaux, mais sont également appliqués avec succès sur des sujets aussi variés que le repliement des protéines qui est l'un des problèmes majeurs en biologie moléculaire ou la découverte de nouveaux médicaments en pharmacologie.



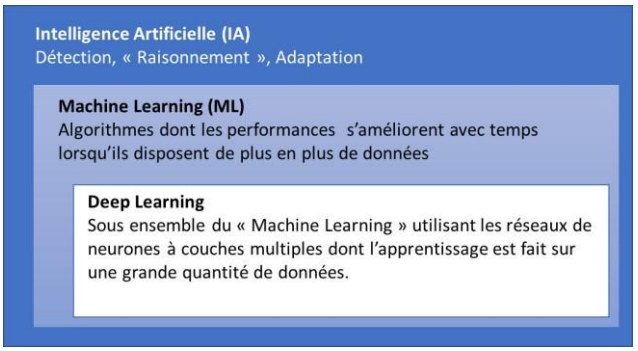
Enfin, pour ce qui concerne la méthodologie d'apprentissage employée par les concepteurs, elle repose essentiellement sur trois phases :

- Un préapprentissage non supervisé.
- Un apprentissage supervisé avec des données annotées.
- Une dernière phase utilisant l'apprentissage par renforcement à l'aide d'une supervision humaine.

Pour conclure cette section, le lecteur intéressé par un « bestiaire » des architectures de réseaux de neurones, pourra utilement visiter le site : https://www.asimovinstitute.org/neural-network-zoo/ (Van Veen, 2016).

3.10 Articulations des concepts étudiés

Les algorithmes utilisés en IA nécessitent l'intervention d'experts humains pour définir l'architecture du réseau, le tester, le valider et le maintenir. Les données fournies à l'algorithme doivent être choisies, sélectionnées, préparées et toutes les informations transformées en tableaux de valeurs avant d'être utilisées. Les paramètres doivent être analysés afin de trouver la bonne fonction d'activation à utiliser pour chaque type d'apprentissage. Le schéma ci-contre montre l'articulation des trois concepts utilisés dans ce chapitre (l'intelligence artificielle, l'apprentissage machine, l'apprentissage profond).



4 LES FONDEMENTS DE LA CYBERSECURITE

La cybercriminalité concerne ici les entreprises, les services publics et les associations. Le vol de données sensibles et les actes de sabotages sont à l'origine de pertes financières pour ces organisations qui se chiffrent en centaines voire en milliers de milliards d'euros à travers le monde, sans compter les pertes immatérielles telles que la crédibilité ou l'image de marque (Bouaziz, 2021). Cet état de fait, rend nécessaire d'investir dans des technologies de protection parmi lesquelles figurent l'IA. Étant entendu, que ces technologies peuvent servir, comme souvent, à la fois des buts défensifs mais également servir d'armes aux mains des « hackers » informatiques (Larue, 2022).

4.1 Définition de la cybersécurité

La définition de la cybersécurité donnée par l'ANSSI sur son site est : *« L'état recherché pour un système d'information lui permettant de résister à des événements issus du cyberespace susceptibles de compromettre la disponibilité, l'intégrité ou la confidentialité des données stockées, traitées ou transmises et des services connexes que ces systèmes offrent ou qu'ils rendent accessibles. La cybersécurité fait appel à des techniques de sécurité des systèmes d'information et s'appuie sur la lutte contre la cybercriminalité et sur la mise en place d'une cyberdéfense »* (ANSSI, s.d.).

La cyberdéfense est en retour définie comme : *« L'Ensemble des mesures techniques et non techniques permettant à un État (ou une organisation) de défendre dans le cyberespace les systèmes d'information jugés essentiels »* (cf. ibid.).

4.2 Champ d'application

Le champ d'application de la cybersécurité concerne tous les secteurs régaliens (défense, sécurité intérieure, administrations publiques), les infrastructures vitales (énergie, transport, santé, etc.), le secteur privé (industries, banques, commerces, services, etc.) et les populations.

La menace peut toucher n'importe quelle entité ou entreprise, quel que soit son secteur d'activité ou sa taille.

La menace est permanente, diffuse et diverse : vol, écoute, brouillage, abus et usurpation d'identité et de droits, altération de données, etc. Les menaces informatiques ont connu une forte croissance depuis le début des années 2000 avec l'essor d'internet dans nos sociétés. Même les systèmes industriels interconnectés (IoT) aux réseaux sont des cibles privilégiées fortement exposées.

4.3 Quelques statistiques d'intérêts

Le coût du cybercrime est estimé à 6 mille milliards de dollars et atteindra les 10,5 mille milliards en 2025 d'après l'entreprise Packetlabs (cf. https://www.packetlabs.net/posts/cybersecurity-statistics-2021/).

En 2021,

- 85 % des incidents de sécurité impliquaient un élément humain.
- 61 % étaient dus à des informations d'identification d'utilisateur volées ou compromises.
- L'ingénierie sociale a été observée dans plus de 35 % des incidents.

4.4 Système de management de la sécurité du système d'information (SMSI)

La cybersécurité et son pilotage reposent sur un système de management de la sécurité de l'information (SMSI) qui est un système d'organisation préconisé par notamment la norme ISO 27001 ou encore le standard SOC-2. Ces standards en définissent les caractéristiques. Au niveau le plus général, un « système de management » est un système qui doit permettre :

- D'établir une politique
- De fixer des objectifs
- D'atteindre les objectifs fixés

Le SMSI est donc avant tout un mode d'organisation que l'entreprise doit mettre en place pour préserver la disponibilité, l'intégrité, la confidentialité et la traçabilité (DICT) de l'information (Fernandez-Toro, 2018). Il permet de gérer les risques relatifs à l'information au moyen de processus, et définit les différentes responsabilités. Il se définit par :

- Des processus (exemple gestion des incidents, supervision, etc.)
- Une politique de sécurité
- Des structures de pilotage et de contrôle
- Des outils de sécurité
- Une démarche d'amélioration continue

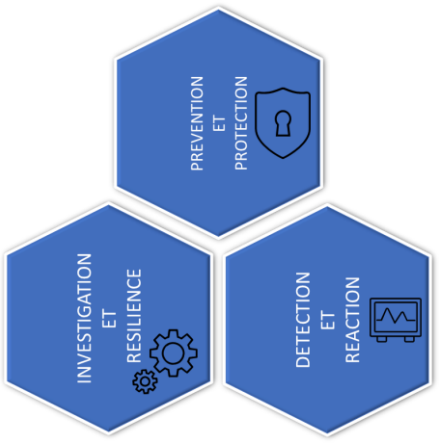
Nous avons évoqué dans cette section la norme ISO 27001 car elle s'est imposée comme référence en matière de management de la sécurité de l'information (Ibid.). Elle est parfois complétée par des audits de type SOC-2 (notamment dans les pays anglo-saxons). Nous ne détaillerons pas dans le présent travail le contenu de ce type de normes.

4.5 Les trois piliers de la cybersécurité

La protection et la sécurisation du système d'information passent par la mise en place d'outils, de processus, d'organisations capables de répondre aux objectifs précédemment évoqués que sont la disponibilité, la confidentialité, l'intégrité et la traçabilité. La cybersécurité repose sur les trois piliers majeurs que sont la « prévention et protection », la « détection et réaction », « l'investigation et la résilience » (Gicat, 2015).

Le pilier « **prévention et protection** » a pour objectifs :

- D'anticiper et de prévoir les menaces et vulnérabilités et d'en déduire les risques.
- De définir les architectures.
- D'installer, configurer et maintenir en condition les ressources.
- De se mettre en conformité avec les meilleures pratiques ou normes du domaine.
- D'établir les procédures associées.
- De sensibiliser et former les personnels.



Ce champ concerne l'ensemble des solutions techniques ou organisationnelles permettant d'éviter ou de réduire l'apparition des incidents et des sinistres sur une infrastructure et de s'y opposer.

Le pilier « **détection et réaction** » concerne essentiellement :

- La détection des incidents et sinistres.
- La collecte et l'analyse des flux et des comportements sur les systèmes afin de détecter un incident au plus tôt.
- La mise en place d'une réaction adéquate afin de circonscrire et maîtriser l'incident pour en réduire au maximum l'impact.

Ce champ concerne l'ensemble des solutions techniques ou organisationnelles permettant la détection et le blocage des incidents et des sinistres sur une infrastructure.

Le pilier « **investigation et résilience** » a pour objectifs :

- L'analyse des incidents afin d'empêcher leur reproduction
- La collecte des preuves en cas de malveillance
- La continuité du service

Ce champ réunit l'ensemble des solutions techniques ou organisationnelles permettant de minimiser les préjudices des incidents et des sinistres, d'analyser les faits et de revenir à l'état initial le cas échéant.

4.6 Segmentation fonctionnelle

Le découpage fonctionnel de la cybersécurité s'appuie sur la segmentation suivante (Gicat, 2015) :

- L'identification et l'authentification
  - Le contrôle de l'accès à un site est la première mesure de sécurité pour garantir la protection physique du système d'information et de ces différents composants.
  - Les mesures de contrôle s'appuient sur des moyens permettant de reconnaître physiquement une personne ou un matériel, c'est l'identification. Les mesures en place doivent également permettre de vérifier l'authenticité des informations communiquées.

- La gestion des identités et des accès
  - La traçabilité de l'information et l'imputabilité des actions sont des enjeux majeurs pour toutes les organisations. L'objectif d'un système de gestion des identités et des accès est de maîtriser la problématique « Qui a accès à quoi ? », de manière que seules les personnes ou les applications ou équipements autorisés aient accès aux ressources (données, informations, etc.) auxquelles elles ou ils ont droit.
  - Les processus de demande et de révocation des droits doivent être gérés de manière continue, audités et contrôlés, en application de la politique de sécurité.

- La sécurité des données
  - Il s'agit de garantir la disponibilité, la confidentialité, l'intégrité et la traçabilité des informations. Par exemple, la signature électronique permet de marquer son engagement sur des données (non-répudiation) et de garantir leur intégrité. Il est aussi important de garantir l'intégrité des données sur le long terme et de préserver la valeur légale des preuves numériques dans le temps.
  - Les outils de chiffrement permettent de gérer la confidentialité et l'intégrité des données y compris entre les différents personnels de l'entreprise mais aussi lorsque ces données sont partagées, échangées ou stockées.

- La sécurité des applications, des infrastructures et des équipements
    - Il s'agit d'assurer la sécurité de son architecture en garantissant la confidentialité, la disponibilité et l'intégrité de toutes les briques la composant. Cette assurance passe par la sécurisation du système dans son ensemble, de chaque équipement et technologie qui le constitue, des interconnexions et des paramètres de configuration. Ici, il est fait usage notamment des techniques de cloisonnement réseau, le filtrage de flux (firewalls) et de leur chiffrement, le durcissement des équipements, la mise en place de solutions antivirus, anti-spam, etc.

- La supervision, la commande, le contrôle et l'aide à la décision
    - Il s'agit de mettre en place une organisation et une surveillance adéquate afin de détecter les incidents de sécurité touchant le système d'information. Les solutions permettant cette surveillance doivent gérer la journalisation d'activité des composants du système et analyser les flux dans le but d'informer les équipes de sécurité. Pour améliorer la détection d'attaques complexes, elles doivent également permettre d'associer les événements entre eux et présenter les synthèses appropriées. Ce type de démarche repose sur la mise en place d'un système de gestion des événements et des informations de sécurité (SIEM).
    - Des équipes spécifiquement formées doivent avoir la capacité de détecter, qualifier et réagir aux attaques.

- Le renseignement et la collecte d'information
    - Le contrôle de l'information est devenu un enjeu majeur pour les entreprises et les services de l'État. Le développement des plateformes d'échanges de données et des réseaux sociaux sont des canaux dont la surveillance est essentielle.
    - La diversité et le volume des informations multilingues échangées dans le monde numérique d'aujourd'hui nécessitent des outils de plus en plus performants afin de recueillir, analyser et proposer les informations pertinentes et dont la valeur est reconnue.

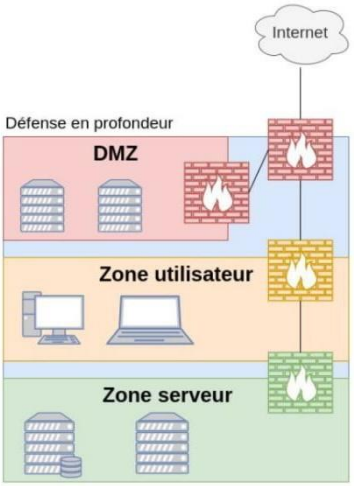
- L'investigation et la recherche de preuve
    - Après un incident de sécurité, il est nécessaire d'investiguer sur ce qui s'est produit afin de pouvoir judiciariser l'événement et/ou éviter sa reproduction.
    - Cette investigation numérique nécessite de s'assurer de l'intégrité des preuves numériques et de leur conservation, d'analyser ces preuves et enfin de présenter un rapport compréhensible par des personnels non experts du domaine.

- L'audit, le conseil, l'exploitation et la formation
    - Le traitement de la complexité et la diversité des systèmes et technologies interconnectés nécessitent des experts à la fois en sécurité des systèmes d'information et dans chaque technique utilisée pour manipuler l'information.
    - Ces spécialistes possèdent les qualifications nécessaires afin de contenir un incident sur le système et réagir avec promptitude et efficacité pour en minimiser l'impact. Ces consultants sont à même de rendre des prestations de niveau gouvernance, contrôle, conception/intégration.

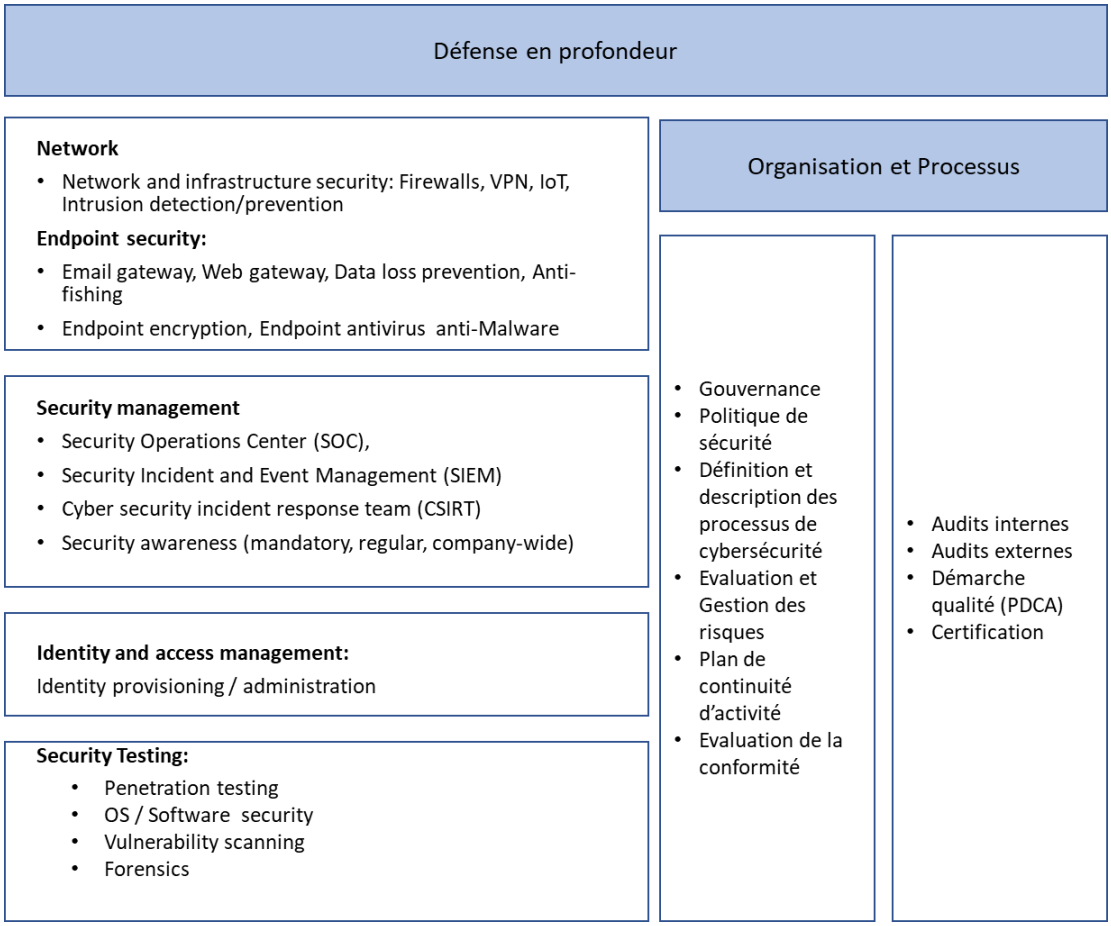
4.7 Doctrine de défense en cybersécurité

La défense en profondeur est la doctrine de cyberdéfense la plus communément acceptée et déployée au sein des organisations. Une série de mécanismes défensifs indépendants sont superposés afin de protéger des données et des informations précieuses (Dorigny, 2022). Si un mécanisme échoue, un autre intervient pour déjouer ou au moins freiner une attaque. Cette approche multicouche à redondances intentionnelles augmente la sécurité d'un système dans son ensemble et s'attaque à de nombreux vecteurs d'attaque différents. En France, cette doctrine a été formalisée en détail par la DCSSI (actuellement l'ANSSI) dans son document « *La défense en profondeur appliquée aux systèmes d'information* » paru en 2004 (DCSSI, 2004).

L'implémentation des principes de cette doctrine se retrouve dans la mise en place d'une défense multicouche pour un Système d'Information avec par exemple :

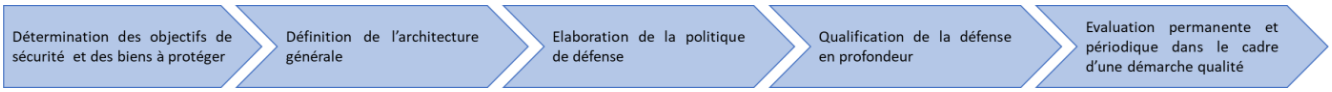


- Le découpage en sous-réseaux (DMZ) multiples par niveau de criticité et/ou de services exposés avec un filtrage des flux entrants/sortants via plusieurs niveaux de pares-feux.
- L'implémentation de zones réseaux par type d'activités. Le schéma ci-contre illustre le concept (schéma repris de Mickael Dorigny , 2022).
- Le déploiement de système de détection sur les principaux nœuds réseaux (IDS, IPS, sondes).
- La mise en place de système antivirus / EDR sur l'ensemble des systèmes.
- L'utilisation systématique de bastions d'administration avec enregistrements des commandes (effectués par les administrateurs).
- Le déploiement de l'authentification multi-facteurs.
- L'utilisation du chiffrement pour les communications des applications.
- L'implémentation de VPN pour le chiffrement des connexions.
- Le durcissement des systèmes pour la réduction de la surface d'attaque sur les équipements.
- La formation des utilisateurs.



Le tableau ci-dessus illustre une déclinaison du concept de « défense en profondeur » du point de vue technologique et organisationnelle.

Il convient de garder à l'esprit que la définition des systèmes de défense et leur positionnement doivent être axés autour de deux éléments : la menace et les biens à protéger. Ceux-ci doivent donc être clairement définis avant toute démarche continue de sécurisation : c'est l'objectif de l'analyse de risque. La mise en œuvre de cette doctrine suit classiquement les phases illustrées sur le schéma ci-dessous et s'inscrit dans une démarche d'amélioration continue.



4.8 Les usages de l'IA en cybersécurité

Les cas d'usages possibles de l'IA dans la cybersécurité sont multiples. Une simple recherche avec le moteur de recherche « Google » permet d'analyser les pages web abordant les sujets de l'usage de l'IA en cybersécurité. Le tableau ci-dessous est une synthèse des cas d'usages possibles remontés sur une vingtaine de résultats de Google, à partir des requêtes "AI Use case Cybersecurity", "Cas d'usage IA cybersécurité" et "Apports IA cybersécurité". Cette « recherche » n'est pas exhaustive ou ciblée pour un type de publication, il s'agit d'articles provenant de portails d'information ou d'entreprises intervenant dans le secteur de la cybersécurité. Les résultats résument simplement, les cas d'usages investigués par l'industrie et les centres de recherche.

Domaine d'apport de l'IA	Cas d'usage	Données
La reconnaissance et l'interprétation des données	Détection de malwares. Traitement de grand volume de données. Support pour l'analyse des attaques. Détection des menaces 0 days. Analyse des menaces réseaux. Détection des Deepfakes. Identification des fraudes. Sécurisation des authentifications. Identification des tentatives de phishing. Analyse des comportements malveillants. Détection de l'usurpation d'identité. Cartographie des systèmes d'information.	Données des logs. Tableau de bord des alertes. Données des outils d'analyse réseaux. Données en provenance des sources caméra, enregistrements audio, images, etc. Données des transactions bancaires, transactions cryptomonnaie. Données comportementales de l'utilisateur.
Aide à la décision	Surcouche de synthèse pour les plateformes SIEM, SOAR, EDR … Corrélation d'événements. Analyse et évaluation de risques.	
Le traitement du langage naturel	Détection de spam. Ingénierie sociale Veille pour les techniques d'attaques, et les attaques à venir. Identification des tentatives de phishing. Désinformation. Identification.	Données textuelles. Données des réseaux sociaux.
La prédiction	Prédiction des risques.	Données sur l'utilisation des ressources de la machine. Données textuelles.

Source (Tri Duc TRAN, communication privée)

Dans le domaine de la cybersécurité, l'usage de l'IA est justifié par la promesse d'une meilleure protection et des gains de productivité, en prenant en charge les tâches chronophages et répétitives via une automatisation de l'analyse de données massives et des actions associées.

4.9 L'IA appliquée aux moyens des cyberattaques

La plupart de tâches chronophages effectuées par un « hacker » dans le cadre de ses activités, comme scanner une machine et trouver les failles non patchées, sont relativement simples à automatiser sans avoir besoin de faire appel à l'IA.

Il est néanmoins possible de réutiliser un système d'IA commercial ou en libre-service, pour une utilisation malveillante. Par exemple, il est possible de réutiliser le « deep fake » ou un équivalent vocal pour tromper des cibles soigneusement sélectionnées. Il est ainsi envisageable de réutiliser des productions de l'IA « clef en main », sans avoir besoin de la créer et de l'entraîner.

Un autre exemple trivial qui ne nécessite pas de compétences particulières en IA, consiste à utiliser des logiciels de traduction comme Google Translate qui permettent à un attaquant de déployer du phishing dans quasiment toutes les langues sans aucune compétence linguistique.

Les attaques de type « Spear fishing » qui sont des attaques très ciblées, qui ne visent qu'une seule personne peuvent bénéficier des techniques d'IA. Il s'agit en effet d'une technique très chronophage qui consiste à récupérer des informations sur la personne ciblée et produire un message personnalisé afin de mieux la tromper. L'IA peut alors être utilisée pour sélectionner les cibles vulnérables grâce aux données des réseaux sociaux, mais également pour les prioriser et toucher les plus rentables selon le niveau de vie identifié par l'IA grâce aux photos ou encore au vocabulaire utilisé dans les commentaires écrits par la cible.

Des attaques de type le déni de service (DDoS) augmentées par IA peuvent être envisagées. L'IA en défense peut servir à différencier les comportements de navigation d'une machine de ceux d'un humain. Cependant, il est possible d'entraîner une IA attaquante contre une IA défensive pour reproduire le comportement d'un humain dans ses schémas de « clics » et dans sa vitesse de navigation, ce qui devrait permettre de tromper l'IA de défense (Larue, 2022). Il est cependant clair que ce type d'attaque nécessite non seulement de disposer de compétences poussées en IA et en réseaux mais également d'avoir accès à des données en grande quantité pour reproduire des trames TCP/IP non détectables.

4.10 L'IA et la désinformation

Grâce à la recherche menée par l'entreprise OpenIA, de nombreux outils spécialisés dans la création de contenu par l'IA sont à notre disposition. Nous y retrouvons GPT (Larue, 2022) qui permet d'écrire des articles avec quelques mots clés et DALL·E 2 de générer des images (Ibid.). Nous pouvons également citer les applications comme les *Deepfakes* et *Faceapps/*filtres pour la création de vidéo.

Ces technologies rendent possibles les attaques par génération d'informations à grande échelle dirigées par des automates (génération automatique de contenu) qui peuvent être utilisées pour inonder les canaux d'information de bruit, c'est-à-dire de fausses informations, ce qui rend plus difficile l'acquisition d'informations réelles. En bref, les techniques d'IA rendent accessibles les opérations de désinformation en réduisant les moyens nécessaires à de telles manœuvres.

4.11 Attaques sur les systèmes utilisant l'IA

Comme tout système d'information, l'IA peut faire l'objet d'une grande variété d'attaques (Larue J., 2018). Des attaquants peuvent ainsi modifier le comportement d'un système d'IA soit en phase d'apprentissage soit en phase de production. Les systèmes à apprentissage en continu sont particulièrement vulnérables aux piratages. Nous étudierons plus loin cette problématique en regardant la taxonomie des attaques sur les systèmes d'IA. La course entre les « hackers » et les équipes en charge de sécuriser le système d'information des organisations s'étend également à ce domaine et la recherche y est très active.

5 L'EXPLOITATION DES DONNEES ET LES DEFIS DE LA REGULATION DES SYSTEMES D'IA/ML

Ce chapitre examine les questions qui se rapportent aux données utilisées dans un contexte d'IA et analyse les enjeux que soulèvent l'exploitation des données en quantité gigantesque sur lesquelles les algorithmes s'entrainent, des enjeux qu'ils soient juridiques ou éthiques.

5.1 La donnée, un enjeu stratégique

5.1.1 Les données, la matière première

Il convient de rappeler que les systèmes d'IA/ML mobilisent de multiples disciplines scientifiques à savoir l'électronique, l'informatique nécessaire au traitement des données collectées, les mathématiques au travers de la modélisation et de l'apprentissage sur des données de qualité ainsi que les sciences sociales pour analyser les impacts sociétaux occasionnés par les systèmes d'IA/ML.

Sur le plan stratégique, le plus important, ce ne sont pas uniquement les disciplines scientifiques mobilisées mais également les données, elles sont la matière première, le carburant nécessaire à l'apprentissage d'une tâche. La valeur créée par l'IA provient plus de l'exploitation des données que des algorithmes.

L'exigence de la qualité et de la pertinence des données utilisées est fondamentale en matière de l'IA/ML. En effet, la capacité des algorithmes à apprendre « correctement » sur de grandes bases de données est étroitement liée à une sélection de données de qualité. Autrement dit, les données doivent être correctement sélectionnées pour éviter l'entrainement des algorithmes sur des données incomplètes, incorrectes, ou redondantes, dont les conséquences seront la production de résultats biaisés ou discriminatoires.

Pour illustrer le propos, nous prendrons l'exemple du robot conversationnel Tay de Microsoft, lancé en 2016 sur la plateforme Twitter. En analysant les conversations de plusieurs réseaux sociaux, Tay s'est très vite mis à tenir des propos racistes et misogynes. Cet exemple nous démontre qu'une pollution de données du système lors du processus d'apprentissage continu peut influencer la qualité des résultats produits par les algorithmes.

Si ces données réutilisées par l'IA/ML appartiennent à une personne précise, nous parlons alors de données privées ou personnelles et/ou confidentielles. En outre, ces données peuvent avoir un intérêt statistique pour l'étude de différents phénomènes. Nous parlons alors de données d'intérêt statistique.

Prenons pour exemple, l'analyse de contenus des livres que nous lisons sur les Sites Web dédiés. Les systèmes d'apprentissage analysent nos préférences, et comme les algorithmes ont besoin pour apprendre de très grands échantillons de données fiables et pertinentes, la machine profilera toutes personnes qui ont un attachement à la lecture et surtout le même genre en littérature que les nôtres dans un intérêt statistique.

En outre, les données utilisées par l'IA/ML peuvent également être des données dites non personnelles, à savoir des données qui ne sont pas rattachées directement ou indirectement à une personne physique comme par exemple les données économiques, etc. L'exploitation de ces données est examinée plus loin dans ce chapitre au travers de l'implication des textes existants qui régissent la protection des données personnelles dans un contexte d'IA.

5.1.2 Risques de détournement des données

L'utilisation croissante des données par les systèmes d'IA/ML pourrait être propice à un détournement de leur usage initial par des acteurs malveillants. Les attaquants ont recours à ces nouveaux usages afin d'augmenter l'efficacité de leurs attaques. Prenons pour exemple, une automatisation d'attaque par ingénierie sociale exécutée avec une imitation artificielle du style d'écriture des contacts de la cible visée par l'attaque, dont le but est de manipuler sa victime.

Nous pouvons nous questionner sur le caractère dual des systèmes d'IA/ML. Des techniques développées dans le domaine de la cybersécurité pourraient – elles être également utilisées à mauvais escient par les attaquants ? Si les applications d'IA/ML utilisées sont identiques, elles n'ont pas les mêmes finalités. Les données personnelles étant à la fois la source et la cible des systèmes d'IA, il est d'une importance essentielle au moment de leur développement de s'interroger sur les mesures de sécurité à intégrer dès la phase de conception *(Privacy by Design)* et d'instaurer également une éthique dès la conception des projets *(Ethics-by-design)* (Levy, 2021). Les bonnes pratiques en matière de sécurité des systèmes d'IA/ML sont étudiées plus loin dans ce chapitre au travers de la proposition législative de la Commission européenne sur l'IA, appelée « Artificial Intelligence Act ». Ce texte vise à rendre l'utilisation des systèmes IA/ML digne de confiance.

C. Villani dans son rapport (Council of europe , 2018) préconise d'étendre la protection des données à une régulation de l'IA orientée par les valeurs, qui seraient fondées principalement sur trois principes décrits en deçà, des exigences que nous retrouvons dans le projet de la Commission européenne[7] :

- Évaluation et gestion des risques ;
- Accent mis sur les acteurs de la chaine de valeur de l'IA qui engloberait le fournisseur, le distributeur et l'utilisateur ;
- Approche fondée sur les valeurs sociales et éthiques.

La sécurité des données est un des enjeux de la protection des données collectées et traitées par un système d'IA/ML. La technologie blockchain semblerait être potentiellement une réponse aux risques de piratage des données. Les données porteraient leur propre sécurité via cette technologie décentralisée, anonyme et cryptée. Toutefois, les réseaux de blockchain peuvent éventuellement être confrontés à des cyberattaques.

5.2 Les données dans un contexte big data

5.2.1 Notion big data

Le big data désigne à la fois des données en quantité gigantesque mises à disposition des fournisseurs du numérique et les techniques innovantes pour les traiter. Ce terme « big data » a été démocratisé par John Mashey, informaticien chez Silicon Graphics dans les années 1990 (Bourany, 2018). Il a noté que les données trop grandes, non structurées et complexes ne pouvant pas être traitées avec des techniques traditionnelles, impliquaient d'innover dans de nouveaux outils d'analyse.

Et, c'est dans le courant des années 2000, que les premiers procédés du big data sont apparus avec la croissance massive de données numériques via les moteurs de recherche. En effet, les données ne pouvant plus être gérées par des techniques classiques, les géants du numérique à l'affût de nouvelles méthodes d'analyse proposent une alternative, Google crée en 2001 Big Table, ainsi que l'algorithme MapReduce publié en 2004 (Brasseur, 2015) pour simplifier la parallélisation des calculs.

[7] Artificial Intelligence Act, publié le 21 avril 2021

5.2.2 La règle des 5V

En 2001, l'analyste Doug LANEY[8] du cabinet Meta Group (devenu Gartner), décrivait le big data d'après le principe des trois V, ensuite élargie à cinq V (Zolynski, 2015).

Le premier V renvoie à un **volume** important de données numériques à traiter. Et face à la croissance exponentielle de la taille des données, il est souvent nécessaire d'investir dans de nouvelles méthodes de stockage, des serveurs hébergés à distance *(cloud computing).*

À l'augmentation croissante du volume des données, s'ajoute la **variété,** les données prennent différentes formes. Elles peuvent être structurées ou non structurées. Si les techniques traditionnelles permettent d'analyser les données numériques dites simples, toutefois elles s'avèrent moins performantes pour traiter les données dites complexes comme les images pixélisées.

Le troisième V renvoie à la **vitesse** de collecte et de traitement des données. La vélocité croissante de création des données numériques suppose une analyse optimisant de manière efficace les temps de traitement afin d'en extraire de nouvelles informations qui étaient jusqu'alors inconnues.

Le quatrième V, renvoie à la **véracité** des données exploitées. La fiabilité des données joue un rôle central, car elle impacte l'ensemble du processus (Ridzuan et al., 2022) (Journodev.tech, 2022), surtout depuis les polémiques provoquées par la propagation de faux messages politiques via les réseaux sociaux, vecteurs de désinformation et de manipulation psychologique (ingérence de la Russie lors de la campagne présidentielle américaine de 2016, par exemple).

Et enfin, le cinquième V qui renvoie à la **valeur** potentielle de ces données, une incitation pour les entreprises à développer des procédés du big data. Si les projets liés au big data généraient près de 42 milliards d'euros de revenu mondial de marché en 2018, les revenus générés via la technologie big data à horizon 2030 sont évalués à plus de 100 milliards d'euros, selon les estimations émises notamment sur le portail Statista (Bourany, 2018).

Depuis l'émergence des premiers signes de la strate big data, nous notons que l'intelligence artificielle s'invite dans notre vie au quotidien, (assistants vocaux, géolocalisation, objets connectés …). Et si les nouvelles technologies comme l'intelligence artificielle et le big data sont possibles aujourd'hui, c'est notamment grâce au progrès des systèmes de stockage combinés au développement de techniques algorithmiques de plus en plus sophistiquées, permettant d'analyser rapidement ces données volumineuses, voire parfois en même temps qu'elles sont produites pour éviter leur stockage éventuel.

Pour illustrer la puissance de la convergence entre le big data et l'IA/ML, nous évoquerons le domaine de la pharmacologie, où l'IA/ML a permis la découverte d'un nouvel antibiotique en analysant une base de 107 millions de molécules (Stockes, Cell 2020) (Laurain, 2022). Le big data est le point de rencontre entre une base d'apprentissage de qualité, les besoins d'analyser ces données abondantes et une technologie développée.

5.3 Cadres règlementaires juridiques spécifiques à l'IA/ML

En dépit de l'utilisation croissante de l'IA/ML, les systèmes sont développés actuellement en l'absence de cadres réglementaires juridiques spécifiques à l'exception, sans doute, de l'ordonnance n° 2016-1057 du 3 août 2016 relative à l'expérimentation de véhicules à délégation de conduite sur les voies publiques à l'échelle nationale.

[8] Author of Infonomics & Data Juice

Pour apprécier les risques liés à l'utilisation de techniques innovantes de l'IA, il faut compter sur des pans entiers du droit. Concrètement sur des textes qui régissent la protection des droits fondamentaux des utilisateurs dans un contexte d'IA, qui sont :

- Le décret n° 2017-330 du 14 mars 2017 relatif aux droits des personnes faisant l'objet de décisions individuelles prises sur le fondement d'un traitement ;
- Le Règlement 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, dit règlement général sur la protection des données (RGPD) (CNIL, 2022) ;
- La Convention 108+, la France est le 22ième pays à avoir ratifié le 27 mars 2023 (CNIL, 2023), le Protocole d'amendement (dit 108+) modernisant ainsi la Convention 108[9]. Pour permettre son entrée en vigueur, 16 ratifications sont encore nécessaires parmi les pays faisant parties de la Convention 108. La Chine et les États -Unis ne sont pas parties de la Convention 108.

Faisant suite à la ratification de la France, l'Argentine a également ratifié le protocole le 17 avril 2023, à l'occasion du Privacy Symposium[10] à Venise (Conseil de l'europe , 2023). Le Rapporteur spécial des Nations Unies sur le droit à la vie privée, encourage tous les pays membres de l'ONU à adhérer à la Convention 108+, instrument juridique contraignant à l'échelle internationale permettant le flux des données et le respect de la vie privée à l'ère numérique (Tarpin, 2023).

5.4 Régulation des algorithmes d'apprentissage automatique

Sans attendre l'adoption du nouveau Règlement européen sur l'intelligence artificielle, la Loi du 7 octobre 2016 pour une République numérique a été la première à poser le principe d'une régulation sur les algorithmes dans le secteur public (Légifrance, 2016).

À l'ère du numérique, caractérisée par la collecte et l'agrégation de grands jeux de données, le décret n° 2017–330 du 14 mars 2017 relatif aux droits des personnes faisant l'objet de décisions individuelles prises sur le fondement d'un traitement algorithmique est entré en vigueur le 1er septembre 2017, en application de son article 3 (Légifrance, 2017).

Le décret d'application de l'article 4 de la Loi n° 2016–1321 du 7 octobre 2016 pour une République numérique, précise les modalités de la communication des règles définissant le traitement algorithmique ainsi que les principales caractéristiques de sa mise en œuvre, et ce conformément à l'art. R. 311-3-1-1 du Code des Relations entre le Public et l'Administration (CRPA). Ceci constitue une avancée majeure, ce dispositif du 14 mars 2017 reconnait aux administrés le droit à une communication des actes administratifs individuels pris sur une décision algorithmique.

La Loi du 7 octobre 2016 pour une République numérique permet d'éclairer le contexte du décret du 14 mars 2017 sous un angle plus large en introduisant le régime du droit à la transparence des algorithmes publics qui servent à fonder des décisions administratives individuelles.

Doivent désormais être communiqués à toute personne faisant l'objet d'une décision individuelle les données traitées et la source des données utilisées par le traitement algorithmique, les paramètres de traitement et les opérations effectuées par le traitement, ainsi que le degré et le mode de contribution du traitement algorithmique à la prise de décision (Art. R. 311-3-1-2. du CRPA).

[9] Convention n° 108 du Conseil de l'Europe sur la protection des données, ouverture à la signature le 28 janvier 1981, et entrée en vigueur le 1er octobre 1985

[10] Edition 2023 du colloque sur la protection de la vie privée

La question des données sur lesquelles les algorithmes d'apprentissage ont été entraînés n'a pas été prise en considération au moment de légiférer sur les textes de la Loi du 7 octobre 2016, pourtant ces données se situent en amont des décisions algorithmiques. Le législateur s'est principalement concentré sur le principe de la transparence des traitements algorithmiques.

5.5 Régulation actuelle applicable et/ou en vigueur dans un contexte d'IA

Les systèmes d'IA reposant sur le traitement automatisé des données, doivent être fondés sur un certain nombre de principes figurant notamment dans le RGPD et la Convention 108+.

Comme nous l'avons déjà évoqué, la Convention 108+ est le seul instrument international juridiquement contraignant pour la protection des personnes à l'égard du traitement automatisé des données personnelles. Elle offre un niveau élevé de protection des données, analogue à celui du RGPD. Elle permet que le développement et l'utilisation de l'IA respectent le droit à la vie privée et à la protection des données *(article 8 de la Convention européenne des droits de l'homme)*[11], et ainsi renforcer les droits de l'homme et les libertés fondamentales.

Le fonctionnement des algorithmes auto - apprenants et les données utilisées par un système d'IA soulèvent des problématiques bien spécifiques au regard des règles visant à protéger la vie privée d'un utilisateur (Donnat , 2019).

5.5.1 Exigence d'exactitude des données utilisées

La constitution de jeux de données de qualité joue un rôle déterminant au fonctionnement de l'IA, puisque les résultats produits dépendront de la base de connaissances sur laquelle les algorithmes sont entraînés. Des jeux de données biaisées produisent automatiquement des résultats biaisés.

5.5.2 Exigence de pertinence des données utilisées

La question de la quantité des données exploitées par un système d'IA est également délicate. D'un côté, le principe de minimisation, consacré par le RGPD et/ou la Convention 108+, implique que la collecte des données soit adéquate, pertinente et limitée à ce qui est strictement nécessaire au regard des finalités pour lesquelles les données sont collectées et traitées. Mais, de l'autre côté, les algorithmes d'apprentissage automatique requièrent très souvent de grande quantité de données lors de la phase d'apprentissage.

Cette exigence de minimisation de la collecte des données peut être atteinte si lors de la conception des algorithmes, un mécanisme d'effacement progressive des données redondantes ou non essentielles est prévu, tout en augmentant graduellement la taille du jeu de données d'apprentissage (Gama et al. 2013) (Council of europe , 2018). Cependant, cette option risque de conduire à une explication erronée ex-post des décisions basées sur l'IA (Doshi – Velez et al. 2017) (Council of europe , 2018).

En 2018, l'autorité norvégienne de protection des données répond à cette problématique, en préconisant de commencer par un volume restreint de données d'apprentissage, puis de vérifier l'exactitude du modèle au fur et à mesure que le modèle est alimenté par de nouvelles données (Council of europe , 2018).

Afin, de s'affranchir de cette contrainte de minimisation de la collecte des données, la CNIL conseille quant à elle, d'avoir recours à des données anonymisées, à condition que cette technique rende

[11] Convention européenne des droits de l'homme, a été signée à Rome (Italie) le 4 novembre 1950 par douze États membres du Conseil de l'Europe, et est entrée en vigueur le 3 septembre 1953.

impossible la réidentification d'un utilisateur contournant ainsi la protection de la vie privée (CNIL, 2022). Les techniques d'anonymisation des données sont examinées plus loin dans l'étude au travers du chapitre intitulé « Le traitement des données au service de l'IA/ML – Approche technique ».

5.5.3 Exigence de détermination de la finalité des traitements des données

Le principe de finalité protège les utilisateurs contre un éventuel détournement de leurs données. Il n'interdit pas le développement de systèmes d'IA, mais il impose à leurs concepteurs de déterminer en amont du traitement des données leurs finalités. Ils doivent également tenir compte le cas échéant, des finalités ultérieures, c'est-à-dire celles qui sont susceptibles d'apparaître au fur et à mesure de l'entraînement de l'IA. Ces finalités dérivées doivent être compatibles avec les finalités pour lesquelles les données ont été collectées initialement. Il est précisé dans l'article 6-4 du RGPD que le fournisseur doit établir l'existence éventuelle d'un lien suffisant avec les finalités initiales pour que ces nouvelles finalités soient légitimes.

Les finalités doivent être décrites avec suffisamment de précision et sans aucune ambiguïté pour permettre à l'utilisateur d'un système d'IA de comprendre en quoi consiste exactement le traitement des algorithmes. Par ailleurs, la détermination d'une finalité définie avec suffisamment de précision permet de limiter les risques d'un usage illicite des données personnelles qui seraient susceptibles de porter atteinte à la vie privée d'un utilisateur d'IA. Ce principe de finalité a été caractérisé par le Conseil constitutionnel comme une garantie légale du droit à la vie privée[12]. Il a aussi été consacré par l'article 8 de la Charte des droits fondamentaux de l'Union européenne selon lequel :

« toute personne a droit à la protection des données à caractère personnel la concernant. Ces données doivent être traitées loyalement, à des fins déterminées et sur la base du consentement de la personne concernée ou en vertu d'un autre fondement légitime prévu par la loi » (Barraud, 2021)

5.5.4 Quant est-il des données non personnelles ?

Les règlements existants en matière de protection des données ne visent que les données à caractère personnel, et ne précisent en aucun cas les autres catégories de données exploitées par certains algorithmes, c'est-à-dire celles qui ne revêtent pas un caractère personnel comme les données boursières, financières, économiques, commerciales, etc. En d'autres termes, toutes informations qui ne permettent pas d'identifier directement ou indirectement une personne physique.

Plusieurs pistes de réflexion sont envisageables afin de pallier à ce vide juridique, la première piste serait de légiférer sur les données non personnelles et plus particulièrement sur celles qui sont utilisées pour nourrir les algorithmes auto – apprenants. Autrement dit, une proposition qui serait semblable aux textes actuels qui régissent la protection des données personnelles. Cette proposition de règlement relative à l'exploitation des données non personnelles pourrait consister à ce que ces données non personnelles soient de qualité et pertinentes afin d'éviter que les algorithmes produisent des résultats biaisés ou discriminatoires. La deuxième piste serait de prévoir une traçabilité des données fournies aux algorithmes.

5.5.5 Exigence de traçabilité des données utilisées

Le traçage du jeu de données personnelles ou non est primordial, en vue d'être en mesure d'assurer la légalité. Ce qui signifie de documenter la source et les conditions de traitement des données utilisées pour entrainer les algorithmes (CNIL, 2022). Il s'agit d'appliquer une démarche de qualité et de promouvoir un système d'IA digne de confiance.

[12] Cons. const., déc. n° 2008-562 DC, 21 févr. 2008.

5.5.6 Exigence de transparence des algorithmes

Le traitement de données personnelles est défini dans l'article 4 du RGPD, comme :

« toute opération ou tout ensemble d'opérations effectuées ou non à l'aide de procédés automatisés et appliquées à des données ou des ensembles de données à caractère personnel »

Cette définition indique que les algorithmes doivent être considérés comme un traitement automatisé de données. Ce qui implique, que le traitement de données à caractère personnel doit être loyal et transparent, à savoir que les informations relatives à un traitement soient facilement accessibles et aisément intelligibles. En d'autres termes, tout utilisateur d'un système d'IA/ML a le droit d'obtenir des informations sur la façon dont l'algorithme fonctionne.

Cette obligation de transparence de conception des systèmes d'IA/ML peut consister soit à fournir une description logique des systèmes ou soit à donner à l'utilisateur un accès à la structure des algorithmes, et par conséquence un accès à la base d'apprentissage de données sur laquelle les algorithmes se sont entrainés. Même si, cet accès à la structure des algorithmes offre la possibilité de détecter un biais potentiel, les droits de propriété intellectuelle et les secrets commerciaux peuvent néanmoins restreindre cet accès (Council of europe , 2018).

En tout état de cause, la cognition humaine est-elle en capacité d'appréhender la complexité de la structure des algorithmes ? Nous devons nous questionner sur la pertinence de permettre à l'humain d'intervenir pendant la conception du système d'IA afin de contribuer à une IA plus accessible et plus responsable.

La solution axée sur la divulgation de la logique des algorithmes serait potentiellement celle à privilégier. Pour respecter les règles relatives à la transparence, le concepteur doit mettre à disposition de l'utilisateur une notice d'information et l'informer qu'il a affaire à un système d'IA et non à un être humain. Cette documentation pourrait fournir des informations sur les catégories de données utilisées au départ et les résultats attendus, sur le mode opératoire des algorithmes, sur les risques afférents à un algorithme, etc.

Cette question de la transparence est nettement plus critique quant au fonctionnement de certains algorithmes, c'est le cas notamment pour les algorithmes auto – apprenants qui s'appuient sur une base de données volumineuses pour développer leur propre mécanisme de « raisonnement ».

Ces algorithmes qui s'alimentent d'un flux de données, peuvent être mis à jour en continu. Ce qui revient à dire que les algorithmes d'apprentissage automatique sont en perpétuelle mouvement tandis que la transparence est au contraire figée. Ce qui désigne que la transparence ne peut concerner qu'un algorithme tel qu'il est utilisé à un moment donné. En raison de cette complexité, il est opportun de communiquer de manière claire et compréhensible sur la façon dont fonctionne un algorithme auto – apprenant, les risques intrinsèques à l'utilisation d'un système d'IA autonome et les catégories de données en entrée et les résultats du modèle de calcul entre autres.

Ce principe de transparence qui n'est qu'une partie de la réponse aux enjeux de l'IA, est défini dans la proposition de règlement sur l'IA, qui est en cours de discussion au Parlement européen.

5.5.7 Exigence du recueil du consentement à une décision individuelle automatisée

L'article 22 du RGPD consacre le droit à l'utilisateur d'un système d'IA à ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, à moins que l'utilisateur ait donné son consentement de manière explicite (article 4 du RGPD). Ce qui veut dire qu'un utilisateur a ainsi le droit de refuser que des décisions importantes le concernant ne soient prises par des algorithmes.

En outre, cette question du recueil du consentement est d'autant plus délicate que certains algorithmes ont besoin pour s'entrainer d'une base de données en quantité gigantesque, une base de données qui est en perpétuelle circulation, ce qui vient complexifier le recueil du consentement.

En d'autres termes, le recueil du consentement de chaque utilisateur se trouve être extrêmement difficile à enregistrer puisque les législations en matière de protection des données personnelles reposent sur une approche statique des données qui peut se retrouver en contradiction directe avec le caractère parfois dynamique des données utilisées.

Nonobstant, cette problématique relative au recueil du consentement dans un contexte big data ou non, le fournisseur du système d'IA ne pourra exploiter les données de l'utilisateur qu'à la condition que ce dernier y est consenti en acceptant de manière explicite les Conditions Générales d'Utilisation (CGU) ainsi que la politique de confidentialité du système. Cette acceptation se matérialise par un acte positif, à savoir une case à cocher.

5.6 Artificial Intelligence Act européen, un encadrement en fonction des risques

L'évolution rapide de l'intelligence Artificielle dans les différentes strates de notre société a conduit la Commission européenne à publier en 2021 une proposition de règlement sur l'IA, appelée « Artificial Intelligence Act ». Cette Loi européenne sur l'intelligence artificielle pourrait devenir la première norme internationale sur le sujet. Approuvée par les états membres de l'UE fin 2022, sa mise en application est prévue en 2024.

L'Artificial Intelligence Act présente une approche réglementaire uniforme et horizontale qui vise à tenir compte des risques inhérents aux systèmes d'IA, sans entraver le développement de technologies innovantes. Cette proposition législative vise à promouvoir l'adoption d'une IA digne de confiance tout en fournissant des exigences strictes en matière de gestion des risques cyber, de monitoring, de gouvernance et de protection des données.

En faisant le choix d'une règlementation horizontale, la Commission européenne cherche à ce que le texte soit pérenne et dynamique, sans que les technologies ne puissent avoir d'impact sur son efficacité.

L'IA Act poursuit quatre objectifs :

- Veiller à ce que les systèmes d'IA mis sur le marché européen soient sûrs, licites et respectent les droits fondamentaux des utilisateurs ;
- Garantir la sécurité juridique pour faciliter l'investissement et l'innovation dans l'IA ;
- Renforcer la gouvernance et l'application effective de la législation existante sur les droits fondamentaux et les exigences de sécurité applicables aux systèmes d'IA ;
- Faciliter le développement d'un marché unique pour des applications d'IA légales, sûres et dignes de confiance (Conseil de l'UE , 2022).

5.6.1 Définition de l'IA émanant du projet de règlement sur l'IA

La proposition d'Artificial Intelligence Act s'applique à un « système d'IA ». L'article 3.1 de la proposition définit le système d'IA comme :

« U*n logiciel qui est développé au moyen d'une ou plusieurs des techniques et approches énumérées à l'annexe I et qui peut, pour un ensemble donné d'objectifs définis par l'homme, générer des résultats tels que des contenus, des prédictions, des recommandations ou des décisions influençant les environnements avec lesquels il interagit* ».

En complément de cette définition, le lecteur intéressé par l'annexe I de l'Artificial Intelligence Act peut se reporter à l'annexe « Proposition de règlement du Parlement européen et du Conseil » en chapitre 12.

Cette définition permet à la Commission européenne de distinguer les applications relevant du domaine de l'IA des systèmes logiciels dits classiques en fournissant certains critères. En outre, la Commission restreint la définition à des systèmes développés au moyen d'approches d'apprentissage automatique incluant l'apprentissage profond, d'approches fondées sur la logique et la connaissance, des méthodes qui sont listées à l'annexe 1 de la proposition législative (cf. chapitre 12). Toutefois cette liste pourrait être actualisée en fonction de l'évolution des systèmes d'IA mis sur le marché en Europe, qui seront susceptibles de créer des violations graves au regard des droits fondamentaux des utilisateurs entre autres.

5.6.2 Acteurs de la chaine de valeur de l'IA

La proposition de l'Artificial Intelligence Act impacte tous les acteurs de la chaîne de commercialisation d'un système d'IA. Le texte s'applique au **fournisseur du système d'IA** (article 3.2 de la proposition), à savoir la personne qui distribue ou propose le système sur le marché ou en service dans l'Union (article 3.9 et 11 de la proposition). Et peu importe, que le fournisseur soit établi dans l'Union ou dans un pays tiers ou bien à son **importateur** (article 3.6 de la proposition). Il s'agit de la personne qui distribue ou propose sur le marché ou en service le système d'IA pour le compte d'une tierce personne établie en dehors de l'Union européenne.

En outre, ce texte s'applique également selon son article 3.7 au **distributeur du système d'IA**. Il s'agit d'une tierce personne, qui n'est ni le fournisseur ni l'importateur, appartenant à la chaîne d'approvisionnement du système d'IA et mettant le système d'IA à « *disposition sur le marché de l'Union sans altérer ses propriétés* » (article 3.7. de la proposition de règlement).

Et enfin, la proposition de règlement sur l'IA s'applique aussi selon son article 3.4. à **l'utilisateur du système d'IA**, c'est-à-dire à la personne utilisatrice du système d'IA située dans l'Union européenne ou dans les pays tiers, où la solution produite par l'IA est utilisée dans l'Union (Petel, 2023).

Nous observons que la proposition de règlement sur l'IA présente une proximité avec le RGPD dans sa structure et sa philosophie. Le RGPD comme la proposition de la loi sur l'IA s'applique à un ensemble d'acteurs, parmi lesquels nous trouvons le responsable de traitement soit le fournisseur dans l'IA Act ; les sous-traitants au sens du RGPD à savoir l'importateur et le distributeur dans la proposition législative sur l'IA ; et les personnes concernées qui sont les utilisateurs des systèmes d'IA.

5.6.3 Approche fondée sur les risques associés à l'IA

La proposition de l'Artificial Intelligence Act consacre et règlemente quatre niveaux de risques, associés à une contrainte juridique proportionnée (Crichton (Dalloz), 2023) :

- Les systèmes d'IA à risque inacceptable qui sont prohibés (article 5) ;
- Les pratiques à haut risque qui sont soumises à un régime de compliance (article 6 à 51) ;
- Les systèmes d'IA présentant des risques acceptables ou spécifiques qui font l'objet d'obligations de transparence (article 52) ;
- Les systèmes d'IA présentant des risques minimes qui ne sont pas règlementés.

Le premier niveau concerne **l'interdiction de certaines pratiques d'IA qui créent un risque inacceptable** au regard des valeurs de l'Union européenne, notamment des droits fondamentaux.

Quatre usages de l'IA sont interdits sur le marché unique (article 5.1) :

- Les systèmes d'IA qui influencent de manière inconsciente les comportements d'une personne en vue de lui causer ou de causer à un tiers un dommage ;
- Ceux qui exploitent la vulnérabilité d'une personne ou d'un groupe de personnes pour influencer leur comportement en raison de leur situation sociale ou économique ;
- les systèmes de notation sociale d'origine publique (le système de crédit social mis en place en Chine par exemple) ;
- les systèmes d'identification biométrique à distance en « temps réel » dans des espaces accessibles au public par les autorités répressives comme la reconnaissance faciale (sauf exception).

Le deuxième niveau concerne la mise sur le marché **de systèmes d'IA à haut risque.** Ce sont les systèmes qui sont susceptibles d'engendrer un risque pour la santé, la sécurité ou les droits fondamentaux des citoyens européens.

Concernant l'application ChatGPT, elle pourrait être considérée par le Parlement européen comme **un système d'IA à haut risque.** Le recours à cet outil pose des questions au regard du respect des droits fondamentaux des utilisateurs et de leur santé. ChatGPT pourrait créer de la dépendance auprès de certains utilisateurs.

En outre, cette technologie innovante pose également des questions de sécurité, elle peut être utilisée par des acteurs malveillants *« pour créer des messages de phishing (hameçonnage), ou encore pour désanonymiser une base de données et ainsi retracer l'identité d'un utilisateur »* souligne Bertrand Pailhès à l'AFP, qui dirige la nouvelle cellule IA de la CNIL (AFP, 2023).

Si le Parlement européen prend la décision d'y inclure les systèmes d'IA générative, la proposition législative sur l'IA, en cours de négociation risque d'être modifiée avant son adoption finale.

Suite à l'interdiction de ChatGPT en Italie et le dépôt de plusieurs plaintes devant la CNIL (autorité de régulation française), les législateurs européens ont décidé de réguler cette nouvelle technologie. Si cette proposition législative sur l'IA générative est validée, ces modèles d'IA seront classés entre autres en fonction de leur niveau de risque, **d'inacceptable à minime.** Les critères retenus pour évaluer le niveau de risque pourraient être la surveillance biométrique, la diffusion de fausses informations ou les propos discriminatoires.

Les IA génératives comme l'interface ChatGPT, classées au sein des systèmes à haut risque ne seront pas interdits mais soumis à des exigences de transparence plus strictes. Ce qui implique que les fournisseurs devront divulguer tout matériel (notamment les textes, les images ou les musiques) protégé par le droit d'auteur, utilisé pour entrainer les algorithmes. Cette règle doit permettre aux détenteurs des droits de saisir la justice *« pour être rémunérés pour ce qui a été utilisé sans leur consentement »* selon l'eurodéputé italien Brando Bonifie, superviseur des négociations.

En outre, la proposition législative sur l'IA générative stipule que seuls les fournisseurs des systèmes d'IA et non les utilisateurs seront responsables de potentielles violations des droits de la propriété intellectuelle (Marin, 2023).

La proposition de l'Artificial Intelligence Act est principalement consacrée à la mise en conformité **des systèmes d'IA à haut risque** dévoilés sur le marché européen. Cette approche fondée dans une logique de compliance est étudiée plus loin dans l'étude au travers des obligations spécifiques qui incombent aux fournisseurs d'IA à hauts risque entre autres (Crichton (Dalloz), 2023).

Le troisième niveau concerne **les systèmes d'IA présentant des risques acceptables/spécifiques.** Ils sont assortis d'obligations d'information et de transparence renforcées à destination des utilisateurs (article 52). Sont visés ici les systèmes destinés à interagir avec des personnes physiques (Chatbots), les systèmes de reconnaissance des émotions et les systèmes dits de « *Deep fake* » qui manipulent des contenus vidéo, audio, ou autres. Ces obligations visent à s'assurer que la personne physique ait pleinement conscience qu'elle a affaire à une IA et non à un à être humain.

Autrement dit, le fournisseur doit concevoir des systèmes transparents et fournir le mode emploi, le distributeur quant à lui, doit s'assurer que le système d'IA soit accompagné de la documentation technique requise. Et enfin, l'utilisateur a une obligation d'utiliser et de surveiller les systèmes en suivant les instructions d'utilisation jointes.

Les systèmes d'IA sans risque élevé/minime comme les filtres anti-spam ne sont pas régulés par le texte, mais la proposition législative incite le fournisseur à l'application volontaire de codes de conduite. Sont visés ici, notamment les algorithmes de recommandation.

5.6.4 Approche fondée dans une logique de compliance d'un système d'IA à haut risque

Le règlement sur l'IA impose que tout **système d'IA à haut risque** mis sur le marché en Europe, soit soumis à une évaluation de conformité ex-ante et de disposer du marquage de conformité « CE ». Ce marquage permettra de signifier à l'utilisateur que le système d'IA mis sur le marché européen par le fournisseur respecte ses droits fondamentaux.

Il s'agit notamment pour chaque système d'IA à risque élevé :

- D'assurer une gouvernance de l'ensemble des données d'entraînement, de validation et de test répondant ainsi aux critères de qualité des données ;
- De réaliser un processus de gestion continu, c'est-à-dire qu'il soit exécuté tout au long du cycle de vie du système d'IA (Identification, évaluation des risques, et adoption des mesures de remédiation des risques) ;
- De concevoir une documentation technique pour démontrer que les systèmes d'IA à haut risque sont conformes aux exigences attendues ;
- De satisfaire à des obligations de cybersécurité, à savoir la mise en place de mesures de robustesse, d'exactitude et de sécurité ;
- De garantir une obligation d'information et de transparence à destination de l'utilisateur ;
- D'assurer une surveillance humaine efficace, c'est-à-dire une surveillance par des personnes physiques pendant toute la période d'utilisation du système d'IA, notamment au moyen d'interfaces homme-machine ;
- De conserver les journaux de logs générés automatiquement par le système permettant ainsi son auditabilité, le cas échéant (Commission européenne , 2021).

Comme nous l'avons observé précédemment, ce texte s'inscrit dans la continuité du RGPD, il prévoit une analyse d'impact relative à la protection des données personnelles si le traitement des données est susceptible d'engendrer des risques élevés pour les droits et libertés des utilisateurs. Cependant, nous regrettons que le texte n'aille pas jusqu'au bout de sa logique car il ne prévoit pas de publication de l'analyse d'impact pour permettre à chaque utilisateur de connaitre les risques éventuels, et les actions mises en place pour remédier aux risques identifiés.

5.7 Quelle éthique pour une IA digne de confiance ?

5.7.1 Cas du robot conversationnel développé par OpenIA : ChatGPT

La Commissaire aux droits de l'homme Dunja Mijatović déclarait en 2018 :

« L'IA peut aider les êtres humains à être plus libres et à s'épanouir. Mais elle risque aussi de nous entraîner vers une société dystopique. Il est donc urgent de trouver le juste équilibre entre les progrès technologiques et la protection des droits de l'homme. C'est un choix de société dont dépend notre avenir »[13] (Barraud, 2021).

Cinq (5) ans après, ChatGPT suscite de vives inquiétudes auprès du public. Dans une lettre ouverte, des signataires experts en IA ainsi que des anonymes réclament un moratoire de six mois sur tout développement dans le domaine de l'IA en raison *« des risques majeurs pour l'humanité »*.

Ces signataires appellent à la mise en place de systèmes de sécurité dont notamment de nouvelles règlementations afin de réguler les systèmes d'IA mis sur le marché, la surveillance des systèmes d'IA, des techniques pour permettre à l'utilisateur de l'informer qu'il a affaire à une IA et non à un être humain, etc. Ces exigences rappellent celles initiées dans la proposition de règlementation européenne sur l'IA (vu plus haut) pour une IA digne de confiance. Ils estiment que les systèmes d'IA non contrôlés peuvent mettre en danger notre démocratie. Ils considèrent que les systèmes d'IA peuvent présenter un risque de détournement de leurs usages initiaux parmi lesquels la désinformation à grande échelle, les cyberattaques entre autres (France 5, C dans l'air , 2023) .

Quelles sont les réelles motivations d'Elon Musk, à l'origine du moratoire de six mois dans la recherche sur l'intelligence artificielle générative ? En dépit de sa signature, demandant ce moratoire temporaire sur le développement de l'IA, nous apprenons qu'il développe en parallèle un nouveau système d'IA générative en vue de concurrencer ChatGPT de la société OpenAI. La nouvelle startup basée dans le Nevada se nomme « X.AI », (Fabrion, 2023).

Le groupe d'éthique technologique (Center for AI and digital Policy)[14] a déposé plainte contre la société américaine OpenAI, auprès de la Federal Trade Commission (FTC)[15], l'autorité américaine de la concurrence. Cette plainte porte sur la mise sur le marché de la dernière version du robot conversationnel, ChatGPT-4, tout en reconnaissant qu'il y avait des risques de résultats biaisés ou discriminatoires (Cherif, 2023).

Les principales critiques sur ChatGPT-4 portent sur l'absence de mise en place de mesures de sécurité adéquates pour prévenir contre certains risques cyber. L'ONG évoque les usages détournés de l'intelligence artificielle, parmi lesquels la désinformation, la cybercriminalité entre autres. Il s'agit de sujets de préoccupations analogues à ceux identifiés par les experts d'Europol[16] (Europol, 2023). La CAIPD note aussi le manque de transparence dans la conception de l'outil ChatGPT. L'ONG appelle à la création d'une commission nationale pour évaluer l'impact de l'intelligence artificielle sur la société américaine (Cherif, 2023).

En Europe, l'autorité de la protection des données italienne interdit l'utilisation de l'outil ChatGPT sur le territoire italien depuis le 30 mars 2023.

[13] D. Mijatović, « Protéger les droits de l'homme à l'ère de l'intelligence artificielle », Strasbourg, 3 juill. 2018

[14] Center for AI and Digital Policy (CAIDP) : une organisation de recherche à but non lucratif basée à Washington D.C. aux États-Unis.

[15] La Federal Trade Commission (FTC) : une organisation fédérale américaine en charge de la protection des consommateurs

[16] Europol : une agence européenne de police criminelle qui contribue à garantir la sécurité des Européens.

Cette interdiction intervient à la suite de plusieurs manquements aux principes de sécurité des données en vertu du RGPD, des exigences que nous avons abordées précédemment dans ce chapitre. Elle estime que l'entreprise américaine OpenAI aurait traité illégalement les données personnelles des citoyens italiens sans la prise en compte du respect de leurs droits fondamentaux (Dimeglio, Avocat , 2023).

Faisant suite à la décision de l'Italie d'interdire ChatGPT, deux plaintes ont été portées devant la CNIL pour des manquements de l'entreprise américaine d'OpenIA au titre du règlement européen sur la protection des données personnelles. Ces manquements observés consistent d'une part, à une absence d'acceptation à la politique de confidentialité de l'outil ChatGPT – 4, qui doit être recueillie par le biais d'un acte positif (case à cocher) et d'autre part à une demande d'exercice du droit d'accès aux données personnelles restée sans réponse. La première plainte a été déposée par David Libeau, un développeur concerné par la protection des données personnelles (Libeau, 2023). La seconde plainte a été déposée par l'avocate Zoé Vilain, présidente de l'association de sensibilisation aux enjeux du numérique Janus International (Vilain, 2023), (ZED.NET, 2023). Depuis, d'autres plaintes ont été portées devant la CNIL contre le robot conversationnel ChatGPT.

ChatGPT serait conforme au California Privacy Rights Act (CPRA), mais l'applicatif ne respecterait aucune des dispositions issues du RGPD européen (Cohen, Avocate, 2023). Dans la politique de confidentialité publiée via la plateforme, OpenIA énumère les droits que peuvent exercer les utilisateurs résidents dans l'État de Californie[17]. Cependant, Il n'y a aucune mention d'information portée à la connaissance des résidents européens, dans les conditions conformes au RGPD européen.

En Europe, la société OpenAI est surtout suspectée d'utiliser les données des utilisateurs à leur insu pour l'entrainement des algorithmes. Les données des utilisateurs ne peuvent être collectées et traitées, si seulement ils ont consenti de manière explicite au traitement de leurs données personnelles.

Nous en sommes seulement aux prémices des dépôts de plainte devant les autorités de régulation des pays européens et autres. À ce jour, il est à noter qu'il n'y a pas encore eu de décision de justice ni de jurisprudence en qui concerne ChatGPT puisque sa mise sur le marché est très récente.

Cela étant dit, l'interdiction ne va pas sans risque et constitue souvent une mauvaise solution. Ainsi l'apparition de l'imprimerie dans le monde musulman a connu un retard de trois ou quatre siècles sur l'Europe. Les technologies de l'imprimerie avaient été interdites pour des questions d'ordre idéologique (la langue arabe ne devant être retranscrite qu'à la main comme au temps du prophète) mais sans doute également politique. Et ce n'est qu'à partir du XVIII^{ème} siècle que l'Empire ottoman prend réellement conscience de son retard constatant les progrès en Europe occidentale.

Ces moratoires ou interdictions même partielles, en raison de risques majeurs pour l'humanité n'auront-t-ils pas le même résultat que celui subi par le monde musulman au nom du respect d'une certaine idéologie et conduire les états concernés à connaitre un retard technologique ?

5.7.2 Cadre normatif pour une IA digne de confiance

Comme nous l'avons vu précédemment, une IA éthique suppose entre autres la qualité et la pertinence des données en entrée d'un système d'IA/ML, la transparence de l'information et la traçabilité des données utilisées pour l'entraînement des algorithmes. Pour faire face aux enjeux d'éthique, une des réponses qui a été choisie, est celle de l'implication juridique sous la forme de textes nouveaux et/ou existants réaffirmés en matière de protection des droits fondamentaux des utilisateurs.

[17] La Californie a adopté, le 28 juin 2018, une loi en partie inspirée du RGPD européen

Dès 2014, le Conseil d'État au travers de son étude sur le numérique et les droits fondamentaux invite les experts, principalement les juristes à réfléchir sur les modalités de régulation des algorithmes en repensant les grands principes de la protection des droit fondamentaux (Conseil d'Etat , 2014). Le Conseil d'État a formulé la nécessité de favoriser la transparence du fonctionnement des algorithmes, l'intervention humaine dans la prise de décision et d'un contrôle normé des résultats algorithmiques en vue de promouvoir une IA responsable.

En 2016, les Etats-Unis ont aussi publié un rapport de la Maison-Blanche sur la régulation des algorithmes appelé *« On Algorithmic Systems, Opportunity and civil Rights »* (Thibodeau , 2016).

Face aux craintes grandissantes que soulèvent les systèmes d'IA, les concepteurs doivent proposer des systèmes responsables, c'est-à-dire des systèmes qui seront sans risques pour les droits fondamentaux des utilisateurs. Pour répondre à ces inquiétudes, la Commission européenne propose dans ses lignes directrices une définition d'une IA digne de confiance, à savoir :

« Une IA digne de confiance suppose d'offrir aux citoyens la maîtrise de leurs données personnelles et d'éviter que ces dernières ne soient utilisées à leur encontre à des fins préjudiciables ou discriminatoires »[18].

La chef de la division de la politique de l'économie numérique de l'OCDE, Anne Carblanc disait qu'une IA digne de confiance est :

« *Une IA centrée sur l'humain et digne de confiance est une clef de la diffusion et de l'adoption de l'IA »* (Demichelis, 2019)

Dans ses lignes directrices en matière d'éthique, la Commission européenne identifie sept exigences essentielles auxquelles les systèmes d'IA/ML doivent répondre pour parvenir à une IA digne de confiance, des éléments quasi – identiques à ceux initiés par l'OCDE[19] (OCDE, 2019). Les recommandations sont les suivantes :

- **Une approche fondée sur l'humain** : nous retrouvons ici, la notion de prise de décision assistée par l'IA et non prise par l'IA. Les systèmes d'IA doivent se mettre au service de l'utilisateur dans le respect de ses droits fondamentaux, sans chercher à restreindre son autonomie.
- **Robustesse technique et sécurité du traitement** : une IA digne de confiance nécessite des algorithmes suffisamment sûrs, fiables et robustes afin d'éviter les dérives tout au long du cycle de vie des systèmes d'IA.
- **Protection de la vie privée et gouvernance des données** : il faut que les utilisateurs puissent disposer d'un contrôle total sur leurs propres données personnelles et que les données les concernant ne soient pas utilisées contre eux à des fins préjudiciables ou discriminatoires. Nous retrouvons ici les principes couverts par le RGPD (comme de la collecte des données nécessitant le plus souvent le consentement à leur traitement pour une finalité bien déterminée, ou encore le droit à l'oubli numérique, etc.).
- **Transparence autour de l'algorithme :** comme nous l'avons vu précédemment**,** cette notion de transparence implique que le fournisseur doit notamment communiquer à tout utilisateur d'un système d'IA, le mode opératoire des algorithmes, les données utilisées au départ, les résultats des algorithmes, etc. L'utilisateur doit également être informé qu'il a affaire à une IA et non à un être humain. En outre, Le fournisseur doit être en mesure d'assurer la traçabilité des systèmes d'IA.

[18] Commission européenne, « Lignes directrices en matière éthique pour une IA digne de confiance », 8 avril 2019
[19] Recommandation du Conseil sur l'IA, adoptée le 22 mai 2019

- **Diversité, non-discrimination et équité** : il convient ici d'éviter que les systèmes d'IA créent des biais discriminatoires. Prenons pour exemple le cas de Facebook, en juin 2022, la compagnie a été condamnée par la justice américaine car son algorithme Lookalike Audience a été jugé discriminatoire. Il ne présentait des offres de location qu'à une certaine catégorie d'utilisateurs, en fonction de critères bien déterminés tels que la religion, l'origine ethnique, le sexe, le statut marital, etc.
- **Responsabilité** : il convient ici, d'identifier les acteurs responsables du bon fonctionnement des systèmes d'IA et d'assurer la mise en place de mécanismes notamment par le biais du contrat. L'objectif est de pouvoir garantir la responsabilité de chacun des acteurs et les soumettre à une obligation de rendre des comptes, en cas de résultats biaisés ou discriminatoires.
- **Développement durable et bien être** : les systèmes d'IA devraient être utilisés pour soutenir des évolutions sociales positives, renforcer la durabilité et la responsabilité écologique. Les entreprises peuvent par exemple, optimiser les données avec lesquelles les algorithmes s'entraînent, choisir les bonnes heures pour les faire tourner, disposer de l'énergie la plus verte pour les entraîner, et choisir la localisation des serveurs avec l'énergie la plus verte. Toutes ses propositions à titre d'exemple, peuvent être des solutions éco-responsables et facilement applicables dans les projets d'IA (Groupe d'experts de haut niveau sur l'IA (GEHN IA), 2019).

5.8 Modèle de régulation chinoise

5.8.1 Règlement sur la protection des informations personnelles : l'équivalent chinois du RGPD européen

Le 1 novembre 2021, la nouvelle loi chinoise sur la protection des informations personnelles (Personal Information Protection Law) est entrée en vigueur. Ce nouveau règlement compte 74 articles répartis sur huit chapitres (Creemers & Webster, 2021).

La PIPL est en de nombreux points similaire au RGPD européen. Il consacre le principe de qualité et de pertinence des données utilisées, celui de la traçabilité et de la transparence des données collectées ainsi que le recueil du consentement. Cette loi introduit également l'obligation de notification de toute violation de données ainsi que l'obligation de désigner un responsable de la protection des informations personnelles, connu sous le nom de DPO au niveau européen. La personne concernée dispose désormais d'un contrôle sur l'ensemble de ses données, elle peut par exemple refuser que ses données soient utilisées pour une prise de décision fondée exclusivement sur l'IA.

Cependant, la loi n'offre pas de garanties similaires à celles proposées par le RGPD. Contrairement au RGPD qui a vocation à s'appliquer à tout organisme traitant des données personnelles d'un résident européen, ce texte semble s'appliquer seulement aux entreprises domiciliées en Chine. En outre, le RGPD s'applique aussi bien aux entreprises privées qu'aux institutions publiques, ce qui permet d'assurer un niveau de protection complet des droits fondamentaux des personnes concernées au contraire de la règlementation chinoise qui exempte les pouvoirs publics. Autrement dit, la loi n'est pas applicable aux services du gouvernement chinois, qui pourront continuer à surveiller les individus sans risque d'être sanctionnés. La surveillance de masse reste la priorité essentielle de l'État chinois pour le maintien de la sécurité nationale et la cohésion sociale.

Tout comme le RGPD, la PIPL prévoit que les entreprises privées puissent être sanctionnées en cas de non-conformité à la règlementation et payer une amende d'une valeur maximum de 50 millions yuans ou 5 % du chiffre d'affaires annuel de l'année précédente (Bocoum & Briot Juristes , 2021).

En outre, tout manquement à la législation est sanctionné par la Cyberspace Administration of China (CAC), l'autorité de contrôle désignée pour la protection des données personnelles, et également par les ministères et services concernés du Conseil d'État et des gouvernements locaux (Deleporte Avocat, 2021).

Étant donné que la Chine héberge la plus grande communauté en ligne au monde, avec plus d'un milliard d'internautes, la PIPL (Personal Information Protection Law) aura vraisemblablement un impact considérable à l'échelle internationale. Cette loi sur la protection des informations personnelles qui vise à garantir la confidentialité des données des utilisateurs, pourrait modifier la vision des entreprises à y investir ou s'y installer (Zhu, 2022).

Précisément, cette loi illustre la volonté de Pékin de mieux contrôler les géants du numérique exerçant sur son territoire. Les entreprises américaines comme Yahoo et LinkedIn ont déjà quitté le marché chinois en raison des contraintes règlementaires devenues de plus en plus difficiles. En effet, les lignes directrices du texte apportent des précisions sur l'obligation de localisation des données personnelles des citoyens chinois. Elles doivent désormais être traitées en Chine et stockées sur le territoire chinois, sauf si le transfert vers un pays tiers est nécessaire et sous réserve d'avoir recueilli le consentement de la personne concernée. Seules certaines données font exception à cette règle, lorsque le transfert est indispensable. Cependant ce transfert est soumis à une obligation de contrôle, les opérateurs étrangers doivent passer un test de sécurité des données organisé par la CAC (Deleporte Avocat, 2021).

Depuis les révélations d'Edward Snowden sur la surveillance de nos données, cette règle qui impose aux entreprises étrangères de stocker les données des citoyens chinois au niveau national et de collaborer avec des partenaires locaux traduit une forme de souveraineté numérique. Il s'agit pour Pékin d'imposer une limitation de circulation des données à travers les frontières.

Il est à noter que la PIPL vient s'ajouter à la loi sur la sécurité des données (la Data Security Law), entrée en vigueur au 1er septembre 2021, qui impose aux entreprises de classer les données collectées en fonction de leur valeur économique et de leur importance (Coëffé, 2021).

5.8.2 Règlementation des algorithmes de recommandation

La Chine a adopté le 04 janvier 2022 une nouvelle réglementation stricte pour encadrer les activités de recommandation algorithmique[20] dans les services d'information sur internet. Ce nouveau texte élaboré par l'Administration du Cyberespace de Chine (CAC) est entré en vigueur le 01 mars 2022.

Cette réglementation vise à protéger davantage les droits des utilisateurs contre des pratiques discriminatoires et la propagation de fausses informations. Elle interdit les algorithmes qui créent de la dépendance chez les utilisateurs. Les fournisseurs de services de recommandations algorithmiques doivent être transparents sur le fonctionnement de leurs algorithmes. Cette loi rappelle la proposition européenne de règlement sur l'IA avec son article 52 qui stipule certaines obligations de transparence pour certains systèmes d'IA. Ce qui veut dire que Pékin impose aux fournisseurs de révéler le fonctionnement de leurs algorithmes.

Le 12 août 2022, la CAC a mis à disposition du public une liste qui explique comment les algorithmes des géants du numérique fonctionnent. Elle explique par exemple que les algorithmes d'Alibaba exploitent l'historique de navigation et de recherche de ses internautes pour leur recommander de nouveaux produits. Cependant, les informations les plus sensibles des systèmes ne sont pas communiqués au public, même si nous pouvons supposer que Pékin puisse les détenir pour asseoir sa surveillance sur les géants de l'internet (Fabrion, 2022).

[20] Les algorithmes de recommandation sont utilisés par les réseaux sociaux pour influencer nos comportements et nos décisions.

Les lignes directrices du règlement offrent la possibilité à tout utilisateur d'un service IA, le droit de modifier à tout moment les données le concernant ou celui de renoncer à toute recommandation algorithmique. Ce qui implique que tout utilisateur dispose de plus de contrôle sur les informations qui lui sont présentées par les algorithmes de recommandation.

Outre la préoccupation de l'utilisateur, les lignes directrices du texte visent aussi à protéger les intérêts de sécurité nationale, à promouvoir un usage éthique des algorithmes entre autres. Le texte impose aux fournisseurs d'algorithmes de prévenir les autorités de réglementation de Pékin dès qu'ils ont une suspicion de diffusion d'informations fausses et/ou contraire à la morale.

La règlementation stipule également que les fournisseurs ont une obligation en matière de sécurité algorithmique. Ils doivent mettre en place des dispositifs de contrôle pour éviter des abus en ligne (pédophilie, discours haineux, propagande, terroriste, etc.). Ainsi, ils doivent examiner dans la durée, le fonctionnement des algorithmes, l'éthique technologique, le contenu mis en ligne, la sécurité des données des utilisateurs, etc. (Koller, 2022) ; (Fasinou, 2021).

5.8.3 Projet de réglementation de l'intelligence artificielle générative

La Cyberspace Administration of China (CAC), a publié le 11 avril 2023, un projet de règlementation relatif au développement de l'intelligence artificielle générative. Ce texte est un premier aperçu des règles visant à recueillir les avis du public chinois et des acteurs de la nouvelle technologie.

Pékin souhaite encadrer cette nouvelle technologie innovante face à un secteur en pleine expansion, en particulier face au succès rencontré par l'application ChatGPT, lancé par la société américaine OpenAI en novembre 2022, un système qui peut conduire à des dérives et être en contradiction directe avec les valeurs que prônent l'État chinois.

Ce texte intervient au moment où les géants chinois de la technologie dévoilent leurs propres outils d'IA générative pour concurrencer l'application ChatGPT d'OpenIA. Le géant du commerce en ligne Alibaba a dévoilé sa propre interface de type ChatGPT, appelée Tongyi Qianwen pour toutes ses applications commerciales. De même, le moteur de recherche Baidu a annoncé quant à lui, la mise sur le marché chinois d'une variante de ChatGPT, Ernie bot qui fonctionne en mandarin.

Ces géants de l'internet devront donc s'adapter pour répondre aux nouvelles exigences de la réglementation de Pékin. Ils devront trouver un juste équilibre entre le projet règlementaire en matière d'IA générative et l'innovation pour préserver leur compétitivité (Malki, 2023).

Si les fournisseurs du numérique venaient à ne pas respecter les dispositions édictées par le règlement, ils seront condamnés à payer des amendes, leurs services seront suspendus ou ils feront l'objet d'enquêtes pénales (Trueman, 2023).

Avant toute mise sur le marché des systèmes d'IA générative, les lignes directrices du projet prévoient que les fournisseurs devront les soumettre à des évaluations de sécurité qui seront effectuées par l'autorité chinoise de régulation du cyberespace (CAC). Les fournisseurs deviennent alors responsables du résultat, de toute fuite de données personnelles ou de toute violation de la propriété intellectuelle. Ils devront donc s'assurer que chaque donnée traitée par l'IA générative est légitime, même si aucun système d'IA générative ne peut garantir l'exactitude des réponses et que les contenus générés respectent le droit de la propriété intellectuelle (Challand, 2023).

Les lignes directrices stipulent aussi que si des contenus contraires aux valeurs du pays, devaient être générés par les outils d'IA, les fournisseurs disposeront d'un délai de trois mois pour mettre à jour

leurs systèmes de sécurité, afin d'éviter que des contenus similaires soient de nouveau générés (Trueman, 2023).

Depuis le mois d'octobre 2022, les concepteurs chinois doivent désormais faire face aux restrictions américaines en matière d'importation de semi-conducteurs. Ces restrictions devraient ralentir leur développement en matière d'IA. En effet, le 7 octobre 2022, l'administration de J. Biden a annoncé au nom de la sécurité nationale, une restriction drastique de leur exportation de semi – conducteurs vers la Chine, craignant que les puces électroniques produites par les entreprises Américaines ou Taïwanaises ne soient trop avancées et ne servent à améliorer la puissance militaire de Pékin. Par exemple, certains composants, tels que la carte graphique (GPU) H100 de NVIDIA, possèdent un taux de transfert et une puissance de calcul trop élevés et sont interdits à l'exportation (Messina, 2023).

5.9 Modèle de régulation américaine (États-Unis)

5.9.1 Proposition de loi fédérale sur la confidentialité et de protection des données : vers un RGPD des États-Unis

Les États-Unis ne disposent pas de loi fédérale sur la protection des données personnelles contrairement à l'Union européenne ou la Chine entre autres. Toutefois, certains États ont promulgué leur propre législation comme l'État de Californie afin de protéger la vie privée de leurs citoyens. Cette loi est connue sous le nom de California Privacy Rights Act (CPRA).

Depuis juillet 2022, les parlementaires des États – Unis travaillent sur une proposition de loi fédérale sur la confidentialité des données (American Data Privacy and Protection Act) qui pourrait devenir le RGPD des États Unis. Cette proposition de loi qui a été déposée par la commission de l'énergie et du commerce de la Chambre des représentants des États-Unis pourrait devenir la première loi nationale aux États Unis en matière de confidentialité et de protection des données des citoyens américains. En raison de la nécessité d'harmoniser la législation à l'échelle fédérale, il est probable que la Maison-Blanche promulguerait la loi si cette dernière devait être approuvée par la Chambre des représentants et le Sénat.

L'American Data Privacy and Protection Act (ADPPA) est analogue à ce que prévoit le RGPD européen sur certains principes mais appliqué au contexte des États-Unis. Le texte consacre le principe de transparence des données collectées, de minimisation et de qualité des données utilisées. Il règlemente également le principe de sécurité des données et le recueil du consentement. L'ADPPA impose aux entités l'obligation de désigner un responsable de protection des données, comme le DPO en Europe. Cette proposition de loi définit aussi clairement les droits individuels à savoir le droit d'accès, de modification ou de suppression. En outre, Les entités auraient l'obligation d'informer les individus si leurs données devaient être traitées dans des pays dits sensibles tels que la Chine, la Russie, l'Iran et la Corée du Nord (Bonfils, 2023). Cette règle semble s'aligner sur celle édictée par le RGPD en cas de transfert des données de résidents européens vers un pays tiers (hors de U.E.), la Commission européenne impose – dans ce cas là - aux organismes de mettre en place des garde-fous juridiques.

Contrairement au RGPD européen, la proposition de loi permet à un individu américain d'intenter une action privée, si une entité venait à violer ses droits en vertu de l'ADPPA. Cependant, des voix s'élèvent contre cette disposition, la chambre de commerce des États-Unis et l'industrie technologique s'opposent à un droit d'action privé, préférant que l'application de ce droit soit de la compétence de la Commission fédérale du commerce (la Federal Trade Commission). Toutefois, pour éviter de submerger les tribunaux de dossiers, les plaignants devront au préalable alerter la FTC pour leur permettre d'examiner le bien-fondé de leur demande. Si la FTC considère que les plaintes sont recevables, les plaignants devront accorder un délai raisonnable à l'organisation accusée afin de

régulariser la situation avant d'intenter une action en justice. La FTC a été désignée par l'ADPPA pour veiller au respect de son application. (Bonfils, 2023)

Pour l'élaboration de cette disposition relative au droit d'action, l'ADPPA s'est inspirée de la loi américaine sur la confidentialité des communications électroniques (ECPA), qui dispose également d'une règle analogue. L'ECPA en vigueur depuis 1986, n'a pourtant pas submergé les tribunaux de dossiers (Bruno, 2022).

En Europe, les personnes concernées par une violation de leurs données déposent une plainte auprès de l'autorité de régulation, le seul organe compétent pour intenter une action contre une entité qui enfreindrait les lignes directrices du RGPD européen.

Cependant, ce texte présente certaines limites. La proposition de loi ne s'applique pas aux organisations publiques, ce qui implique que les autorités de régulation américaine ne disposeront d'aucun pouvoir pour effectuer des contrôles auprès des entités fédérales, étatiques, tribales, territoriales ou locales qui traitent les données des citoyens américains en vertu de l'ADPPA. Cette limite rappelle la règle introduite par la Chine dans son règlement sur la protection des informations personnelles (PIPL). La PIPL n'est pas applicable aux services du gouvernement chinois.

L'une des limites, qui semble la plus importante de la proposition est qu'une fois la loi adoptée, elle abrogerait toutes les lois étatiques existantes en matière de confidentialité des données telles que la CPRA pourtant une des lois les plus restrictives en matière de protection des données. (Bonfils, 2023)

Nous constatons après l'étude des lois en matière de confidentialité et de protection des données, qu'il s'agisse du RGPD européen, de la PIPL ou de l'ADPPA, les données sont le nouveau champ de bataille géopolitique entre les États. Ils adoptent des normes régissant le flux des données en fonction de leurs objectifs respectifs qu'ils soient économiques, politiques ou sociaux.

5.9.2 Proposition de règlementation des systèmes d'AI /ML

Aujourd'hui, les États-Unis ne disposent pas de loi fédérale pour encadrer l'intelligence artificielle. Par conséquent, certains États comme l'Alabama, le Colorado, l'Illinois ou le Mississippi ont adopté des lois régulant l'IA. Les États de l'Alabama et du Colorado ont promulgué une loi pour restreindre l'utilisation de la technologie de reconnaissance faciale, tandis que l'État de Californie a adopté une loi visant à interdire l'usage de la reconnaissance faciale par la police. Les États de l'Illilois et du Mississipi, ont quant à eux, décider de limiter l'utilisation des systèmes automatisés de sélection de candidats. Ce sont des exemples parmi tant d'autres.

D'autres lois étatiques sont en cours de préparation notamment celle de l'État de Californie, qui incitera les fournisseurs d'IA à mettre en place des mécanismes pour minimiser les biais algorithmiques (Calvi, 2022).

Une proposition de loi connue sous le nom de l'Algorithmic Accountability Act a été déposée à la Chambre des représentants et au Sénat au début du mois de février 2022. Cette proposition de régulation de l'IA semble s'aligner sur la proposition européenne (IA Act) introduit au Parlement européen en avril 2021. L'Algorithmic Accountability Act propose un cadre de régulation uniforme qui sera applicable à l'échelle nationale (Ambassade de France aux Etats - Unis , 2022).

En outre, le Bureau de la politique scientifique et technologique de la Maison Blanche a élaboré un projet intitulé « Blueprint for an AI Bill of Rights ». Cette charte n'est pas un projet de loi mais un plan directeur, qui énumère un ensemble de principes visant à aider les organisations dans *« la conception, l'utilisation et le déploiement de systèmes automatisés, dans le but de protéger les droits des Américains à l'ère de l'intelligence artificielle »,* selon la Maison Blanche (Rey, 2022).

Selon Marc Rotenberg[21], la charte des droits de l'IA dévoile une première ébauche dans la mise en application d'une IA sûre, efficace, éthique, centrée sur l'humain et digne de confiance.

« Il s'agit clairement d'un point de départ. Cela ne met pas fin à la discussion sur la façon dont les États-Unis mettent en œuvre une IA centrée sur l'humain et digne de confiance. Mais c'est un très bon point de départ pour amener les États-Unis à un endroit où ils peuvent poursuivre cet engagement »

Les principes introduits par cette charte des droits vis-à-vis de l'IA semblent être communs à ceux édictés par les lignes directrices du Règlement européen pour une IA digne de confiance, mais la mise en œuvre de ces principes n'est pas contraignante. Cependant, les législateurs pourraient proposer la charte à l'adoption de l'Algorithmic Accountability Act, et ainsi son application aurait une force contraignante.

5.9.3 Réglementation des systèmes d'IA générative

Face à l'essor des systèmes d'IA générative, la National Telecommunication and Information Administration (NTIA), une branche du ministère américain du commerce a ouvert le 11 avril 2023 une consultation publique pour s'assurer que les systèmes d'IA mis sur le marché des États-Unis sont dignes de confiance. Cet appel au public américain intervient le même jour où la Cyberspace Administration of China (CAC) publie son projet de loi relatif à l'intelligence artificielle générative.

En parallèle de cette consultation publique, un groupe d'éthique technologique[22] a demandé à la Commission fédérale du commerce des États-Unis de bloquer toute diffusion de nouvelles versions commerciales de ChatGPT de la société OpenIA sur le territoire des États-Unis pour empêcher les applications d'IA générative de porter atteinte à la vie privée des citoyens américains et à la sécurité nationale (Anthony, 2023).

En outre, face aux inquiétudes que suscitent les modèles d'IA générative, la Maison Blanche a convié le 04 mai 2023, les géants du secteur de l'internet et de l'IA pour aborder les enjeux liés à l'IA et les mesures à prendre pour garantir des systèmes d'IA sûrs, efficaces, éthiques et dignes de confiance entre autres. La vice-présidente des États-Unis, Kamala Harris a déclaré aux patrons de ces grands groupes qu'ils avaient un *« devoir moral »* de protéger la société américaine (Figaro avec AFP , 2023).

La Maison Blanche propose de réguler les systèmes d'IA alors que les États-Unis n'ont toujours pas de loi nationale sur la confidentialité des données promulguée. Les géants du numérique conscients des risques générés par les modèles d'IA générative, réclament une régulation tout en craignant que l'innovation soit freinée par des lois contraignantes. Ne vont – ils pas anticiper et payer des lobbies pour empêcher que de nouveaux projets de loi fédérale ne soient introduits au Congrès des États-Unis?

[21] Fondateur et président du Center for AI and Digital Policy, une organisation à but non lucratif
[22] Center for AI and Digital Policy

6 LE TRAITEMENT DES DONNEES AU SERVICE DE L'IA/ML – APPROCHE TECHNIQUE

6.1 L'approche IA/ML en cybersécurité

L'application de l'IA/ML en cybersécurité est un domaine de recherche dynamique dont les résultats sont prometteurs. Il est certain que dans les prochaines années l'usage de ce type de technologie sera le standard *« de facto »* avec des conséquences directes sur la façon de travailler des professionnels de la cybersécurité.

Néanmoins, force est de constater que le développement de l'IA/ML dans le domaine de la cybersécurité est ralenti par le scepticisme de nombre de professionnels de la cybersécurité, lié le plus souvent à des conservatismes classiquement reliés à une résistance au changement face à l'adoption de toute nouvelle technologie.

Il reste que dans un futur proche, les organisations investiront de plus en plus dans des moyens automatisés d'analyse permettant une réponse adéquate et efficace vis-à-vis des menaces. Elles devront également investir dans la formation des équipes de cybersécurité à l'usage/utilisation des technologies de l'IA/ML. De nouveaux profils apparaitront au sein des équipes en charge de la cybersécurité des entreprises. Ces développements devront également s'accompagner d'un strict respect de la législation en vigueur concernant la protection de la vie privée afin de permettre la généralisation de son usage (vu plus haut).

Bien employées, ces nouvelles technologies permettront dans un premier temps de décharger les équipes de cybersécurité des tâches fastidieuses comme la détection et la caractérisation des cas suspects. Les équipes SOC par exemple, pourront se focaliser sur les tâches à plus fortes valeurs ajoutées.

Comme nous l'avons déjà évoqué, le développement de cette technologie est sous-tendu par les données. Le traitement technique des données est en effet à la base de la qualité des résultats obtenus. Même si les algorithmes ont toute leur importance, la qualité et la quantité des données restent essentielles dans la production d'une solution d'IA/ML pertinente. Dans ce chapitre, nous allons succinctement passer en revue quelques techniques incontournables de traitement des données soit pour améliorer les performances des algorithmes soit plus simplement pour se conformer au règlement général sur la protection des données (RGPD européen).

6.2 Apprentissage et Optimisation

L'un des principaux problèmes dans la détection des menaces est la gestion des faux positifs et des faux négatifs.

Les faux négatifs sont les menaces qui ne sont simplement pas détectées par le système et leur nombre est directement lié à la qualité du modèle. La non-détection d'une menace réelle peut être souvent catastrophique.

Les faux positifs pourraient apparaître assez bénins puisqu'ils ne représentent pas de réel danger. Cependant, leur gestion est un réel fardeau pour les équipes en charge de la sécurité des systèmes étant donné la masse considérable d'alertes auxquelles elles doivent faire face.

L'optimisation des algorithmes commence souvent par la sélection, le nettoyage et la transformation des données d'apprentissage. La préservation de la vie privée des personnes et le respect du droit en la matière peuvent également imposer l'anonymisation de certaines catégories de données.

Considérant la quantité de données souvent disponible, ce travail préliminaire consiste à éliminer les éléments inutiles et/ou redondants, à réduire la dimensionnalité du problème et à vérifier ou faire en sorte que chaque classe soit suffisamment représentée. Ces étapes préliminaires permettent de présenter des données de qualité aux algorithmes d'IA/ML. Elles constituent une grande partie du travail des « Data Scientists » et sont une condition importante de la performance ultérieure des algorithmes.

6.3 Quantité et qualité des données

La mise au point d'une méthode de prédiction efficace repose à la fois sur la qualité et sur la quantité des données d'apprentissage. Les données doivent être suffisamment représentatives du problème et elles doivent donc avoir un niveau de variété suffisant.

À titre d'exemple, dans le cas spécifique de la détection des malwares, les data scientists doivent faire face aux deux problèmes suivants :

- Quels sont les types de malwares pouvant être considérés comme représentatifs des risques et menaces les plus probables pour les organisations ?

- Combien d'exemples faut-il collecter pour obtenir un ensemble suffisamment représentatif et être en mesure de construire un modèle efficient ?

En fin de compte, il ne s'agit pas de simplement choisir le meilleur algorithme pour nos objectifs, mais surtout de sélectionner les cas les plus représentatifs (en quantité suffisante) à soumettre à un ensemble d'algorithmes, que nous devrons optimiser en fonction des résultats obtenus.

Il faut non seulement disposer d'une grande quantité de données mais également vérifier que chaque classe est suffisamment présente pour éviter les biais d'apprentissage.

6.4 Transformation des données

En amont des exemples d'utilisation de l'IA pour le domaine de la cybersécurité, nous précisons que la vaste majorité des algorithmes travaillent sur des valeurs numériques. Ainsi les données textuelles doivent faire l'objet d'une transformation afin d'être traitées. Cette transformation se fait selon trois étapes.

- Étape 1 : La normalisation (nettoyage)
  - C'est à cette étape que les données sont nettoyées pour éliminer les entrées non désirées et pour convertir certains caractères/séquences en formes canoniques.

- Étape 2 : La segmentation (tokenisation)
  - Le flux continu de caractères est divisé en entités. C'est probablement l'étape la plus complexe du processus.

- Étape 3 : La numérisation
  - Les entités textuelles sont converties en nombres pour pouvoir alimenter le modèle. Les données sont alors prêtes pour la phase d'apprentissage.

6.5 Anonymisation des données sensibles

Le règlement général sur la protection des données (RGPD européen) ne comporte pas spécifiquement d'obligation générale d'anonymisation. Il s'agit d'une solution, parmi d'autres, pour pouvoir exploiter des données personnelles dans le respect des droits et libertés des personnes.

D'après le site de la CNIL « *l'anonymisation ouvre des potentiels de réutilisation des données initialement interdits du fait du caractère personnel des données exploitées, et permet ainsi aux acteurs d'exploiter et de partager leur « gisement » de données sans porter atteinte à la vie privée des personnes. Elle permet également de conserver des données au-delà de leur durée de conservation* » (CNIL, 2020).

L'anonymisation de données est donc une pratique courante en IA/ML pour protéger la vie privée et la sécurité des individus. Il existe plusieurs méthodes d'anonymisation de données pour les algorithmes d'intelligence artificielle présentant chacun des avantages et des inconvénients. Nous passons en revue, dans cette section, ces quelques techniques couramment utilisées.

La suppression de données sensibles[23] par exemple consiste à supprimer les données personnelles dites sensibles des ensembles de données avant leur utilisation par les algorithmes d'IA/ML. Cette méthode est simple, mais elle réduit considérablement la qualité des données utilisées pour la formation des modèles puisqu'elle induit naturellement une perte d'information.

La suppression partielle, comme son nom l'indique, supprime une partie des données sensibles tout en conservant certaines informations utiles pour la formation du modèle. Cette méthode préserve une partie de la qualité des données tout en protégeant la vie privée.

La généralisation consiste à remplacer les données sensibles par des valeurs moins précises mais plus génériques (comme remplacer l'âge par une tranche d'âge). Ceci est réalisé en modifiant l'échelle des attributs des jeux de données ou leur ordre de grandeur afin de s'assurer qu'ils restent communs à un ensemble de personnes. Cette technique permet d'éviter l'individualisation d'un jeu de données. Elle limite également les possibles corrélations du jeu de données avec d'autres.

L'ajout de bruit aux données sensibles permet de les rendre plus difficiles à identifier tout en préservant leur utilité. Bien utilisée, cette technique peut dans certains cas rendre les algorithmes d'IA/ML plus robustes, en évitant notamment le surapprentissage. Cette méthode est utile pour protéger la vie privée sans sacrifier la qualité des données.

Enfin, la « pseudonymisation » est une technique réversible qui permet de remplacer les données personnelles par des identificateurs uniques (alias, numéro séquentiel, etc.). Cette technique permet ainsi de traiter les données d'individus sans pouvoir identifier ceux-ci de façon directe. En pratique, il est possible de retrouver l'identité de ceux-ci grâce à des données tierces (par exemple une table de correspondance), sauf si ces dernières sont détruites dans le processus. Cette méthode préserve la confidentialité des données tout en conservant leur utilité.

La combinaison de méthodes : il est souvent utile de combiner plusieurs méthodes d'anonymisation pour obtenir des données anonymes de haute qualité. Par exemple, il est possible de combiner la suppression partielle avec la généralisation pour protéger la vie privée tout en conservant une grande partie de la qualité des données.

[23] Les données sensibles au sens du RGPD sont des données personnelles dont le traitement comporte un risque pour une personne concernée. (exemple : données de santé, données relatives à l'orientation sexuelle, données raciale ou ethnique, etc.)

6.6 Les techniques de rééchantillonnage

Le bagging et le boosting sont des techniques de rééchantillonnage qui permettent un meilleur apprentissage des cas les plus difficiles. En effet, ces méthodes permettent la génération d'une multitude de classificateurs entrainés sur différents jeux de données afin de les combiner entre eux. Le schéma ci-dessous illustre la différence entre le bagging et le boosting en termes de procédure d'apprentissage et de prise de décision. Ainsi la technique du bagging utilise l'échantillonnage pour générer des ensembles d'apprentissage. Le boosting consiste à modifier l'ensemble d'apprentissage en augmentant le poids des éléments mal classés.

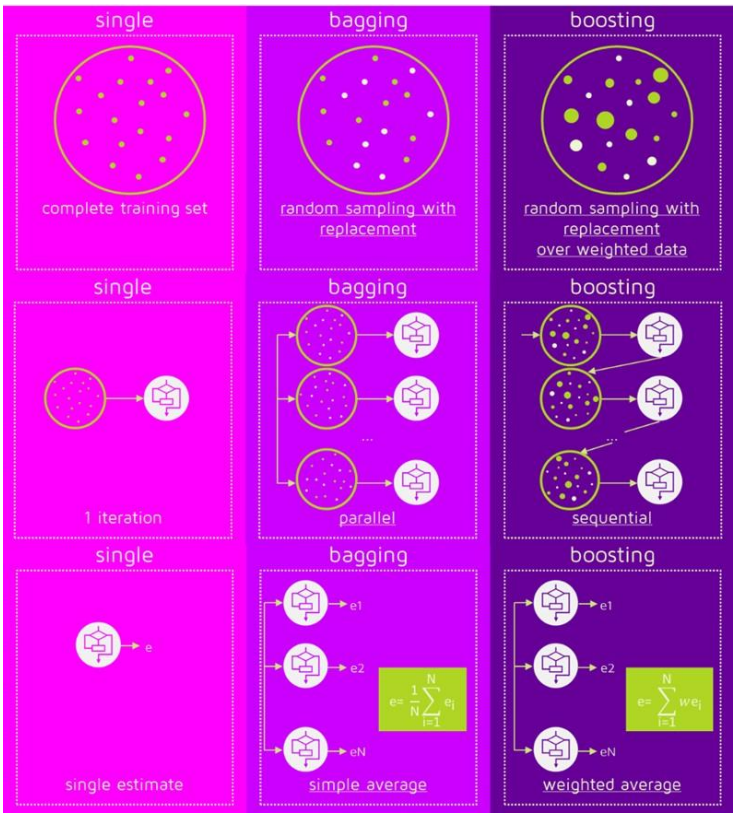


Illustration d'après : https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/

6.7 L'apprentissage à partir d'un ensemble de données non-équilibrées

Les techniques de rééchantillonnage de bagging et/ou de boosting peuvent être utilisées pour obtenir des ensembles de données équilibrées.

Une autre approche consiste à réduire la taille de l'ensemble le plus important en supprimant les observations redondantes ou même aléatoirement. Il est possible de faire l'hypothèse que la réduction de la taille de l'ensemble n'affecte pas la distribution des données. Néanmoins, cette technique présente le risque de supprimer des données importantes.

Une approche qui peut s'avérer intéressante est de faire l'inverse à savoir augmenter artificiellement la taille des ensembles les plus petits en générant des données de synthèse (ou artificielle). Parmi les méthodes de sur-échantillonnage, la technique la plus répandue est le SMOTE (Synthetic Minority Over-Sampling Technique). Dans la pratique, les données synthétiques sont générées depuis des clusters de données calculées à partir d'un algorithme tel que les « K plus proches voisins »[24]. Les

[24] L'algorithme consiste à rechercher les K voisins les plus proches (en utilisant la distance euclidienne, ou autres) et choisir la classe des voisins majoritaires.

exemples synthétiques sont donc générés aléatoirement avec la contrainte d'appartenir au cluster des données à rééchantillonner (par exemple, des données de transactions frauduleuses).

6.8 Combinaison de classificateurs pour améliorer les performances de prédiction

Il est connu depuis très longtemps à travers la littérature que la combinaison de plusieurs modèles de classification améliore de manière assez significative les performances des algorithmes. Cette « combinaison » peut prendre différentes formes telles qu'une simple moyenne, un vote simple, un vote pondéré ou bien l'application d'un classificateur sur les données de sortie des différents modèles.

Dans un processus de classification, chaque modèle dispose de ses forces et de ses faiblesses pour classer de nouvelles données. Compte tenu de la diversité des formes d'entrées (valeurs booléennes, discrètes, continues), un modèle peut bien classer un exemple de données et se tromper sur un autre. Ainsi, la combinaison des modèles permet de pallier, au moins dans une certaine mesure, les erreurs de classification unitaires en partant du principe que des classificateurs différents ne commettront pas systématiquement les mêmes erreurs.

6.9 L'IA/ML dans l'écosystème de la cybersécurité

L'explosion des menaces de cybersécurité, l'augmentation quotidienne du nombre de malwares, la multiplication des groupes malveillants ainsi que la complexité des réseaux à protéger font qu'il devient quasi impossible de conduire les tâches d'analyse et de remédiation de façon manuelle. Il devient nécessaire d'introduire des algorithmes qui permettent à minima d'automatiser le triage des alertes en procédant à une analyse préliminaire et en soumettant à l'attention des experts cybersécurité uniquement les alertes qui nécessitent une analyse plus approfondie. Cette automatisation de l'analyse et du traitement des menaces doit permettre de faire face au nombre toujours grandissant d'attaques.

Les cyber-analystes de demain devront avoir la capacité d'interpréter et évaluer correctement les résultats issus des algorithmes d'IA/ML. Ils devront avoir une compréhension de la logique sous-jacente des algorithmes en place et certains profils devront pouvoir améliorer les solutions en place afin de mieux répondre aux contraintes particulières de leur organisation ainsi qu'à leurs objectifs.

Les tâches reliées à l'IA/ML sont :

- La classification qui est essentielle dans le cadre de la cybersécurité. Elle est utilisée afin d'identifier des ensembles d'attaques présentant certaines similarités comme la détection d'une même famille de malwares présentant des caractéristiques communes même si leurs signatures sont distinctes

- Le clustering permet une classification automatique en identifiant les classes qui émergent « naturellement » dans un ensemble de données. Ces techniques sont particulièrement utiles sur les données non étiquetées. Les analyses de type « forensic » peuvent bénéficier des techniques de clustering.

- L'analyse prédictive permet d'identifier les menaces dès leur apparition en exploitant - par exemple - les réseaux de neurones ou le deep learning. Une approche hautement dynamique est adoptée afin de permettre aux algorithmes d'apprendre automatiquement sur les nouvelles données apparaissant.

- Enfin, l'IA générative est un domaine de l'IA/ML qui se concentre sur la création de modèles qui peuvent générer des données de manière autonome, en utilisant des algorithmes d'apprentissage automatique pour apprendre à imiter des données réelles. Un tel système, entraîné à partir de données de journaux d'événements pour reconnaître les schémas typiques

d'activité malveillante et pour détecter les anomalies dans les données, peut être utilisé pour aider à la résolution d'incidents de cybersécurité. Il est susceptible de fournir une assistance automatisée à l'analyse des données de sécurité et de faire des propositions pour la résolution des incidents.

7 ETUDE DES APPLICATIONS IA/ML POUR LA CYBERDEFENSE

Dans ce chapitre, nous présentons les principales approches IA/ML utilisées pour la détection de menaces ou potentielles attaques (Parisi A., 2019). Il existe une multitude de travaux et d'articles sur le sujet, il est par conséquent très difficile de prétendre à l'exhaustivité.

7.1 Quelques limites

Dans ce chapitre, nous n'aborderons pas les aspects mathématiques des algorithmes, ni les méthodes d'implémentation. Le lecteur intéressé pourra se référer aux ouvrages et site suivants :

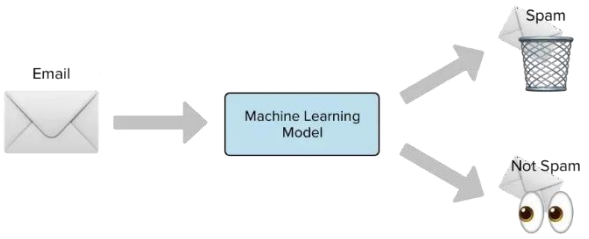
- Le livre intitulé *« Hands-On – Artificial Intelligence for Cybersecurity »* publié aux éditions PACKT en 2019 rédigé par Alessandro Parisi.
- Le livre intitulé *« Machine Learning for Cybersecurity – Cookbook »* publié aux éditions PACKT en 2019 rédigé par Emmanuel Tsukerman.
- Le site : https://scikit-learn.org/stable/

Nous discutons ici des applications et des algorithmes classiquement utilisés, étant entendu qu'à chaque fois, il existe tout une batterie d'algorithmes pouvant être utilisée pour adresser un problème spécifique.

7.2 La détection des « Spams » dans les courriels

Sur les plus de 300 milliards de courriels envoyés quotidiennement, au moins la moitié sont des spams. Les fournisseurs de messagerie ont l'énorme tâche de filtrer les « spams » et de s'assurer que leurs utilisateurs reçoivent les messages qui n'en sont pas (Dickson, 2020).

Historiquement, les premières solutions de filtrage de courriel étaient basées sur des règles de filtrage statiques (systèmes experts) utilisant des expressions régulières afin d'identifier des motifs prédéfinis caractéristiques des « Spams ». Mais face à la capacité d'adaptation des « spammers », il est rapidement apparu que ce type de techniques était relativement limité (Parisi A. 2019). Il a ainsi fallu adopter une approche plus dynamique. Pour cette raison, la détection des « Spams » a ainsi été l'une des toutes première application de l'IA/ML dans le domaine de la cybersécurité à travers la célèbre solution « open source » SpamAssassin. En effet, l'apprentissage automatique s'est avéré être l'approche la plus efficace et la plus privilégiée par les fournisseurs de messagerie. Nous présentons ici les différentes stratégies de détection des spams faisant usage de l'IA/ML.



La détection de spams est un problème d'apprentissage automatique supervisé. Cela signifie qu'il faut fournir à l'algorithme un ensemble d'exemples de spams, un ensemble de messages légitimes et le laisser trouver les modèles pertinents qui séparent les deux catégories différentes.

Le filtrage de « Spams » se base en général sur le calcul des fréquences d'occurrence de certains mots « suspects » ou parfois de couple ou triplet de mots. Sur la base des fréquences d'occurrence, il est alors possible de calculer un score associé au message. Ce score est par la suite comparé à une valeur seuil définie empiriquement afin de classer le courriel en « Ham » ou « Spam ».

Il faut cependant garder à l'esprit que les « spammers » sont parfaitement au courant du filtrage des courriers et qu'ils élaborent constamment des stratégies d'échappement aux solutions de filtrage. Ceci se traduit donc par la mise en place d'un processus itératif d'apprentissage afin d'adapter les algorithmes aux nouvelles stratégies des « spammers ».

Différents algorithmes d'apprentissage automatique peuvent détecter les « Spams », mais celui qui a été sans doute le plus largement utilisé pour ce problème est l'algorithme dit « Naïf Bayes » qui s'appuie sur le "théorème de Bayes" et qui décrit la probabilité d'un événement sur la base de connaissances antérieures.

L'application du théorème de Bayes est dite naïve car elle suppose que les observations des évènements sont indépendantes. Malgré, cette simplification abusive, la méthode est assez performante.

Comme d'autres algorithmes d'apprentissage automatique, la méthode « Naïve Bayes » ne comprend pas le contexte du langage et s'appuie sur des relations statistiques entre les mots pour déterminer si un texte appartient à une certaine classe. Cela signifie que ce type de détecteur de spam peut être trompé si l'expéditeur ajoute simplement des mots « non-spams » à la fin du message ou remplace les termes « Spams » par des synonymes.

Une autre approche relativement simple est l'utilisation de la discrimination linéaire, comme son nom l'indique il s'agit de séparer deux classes d'objets par un hyper-plan (une droite dans le plan) en faisant l'hypothèse que les classes sont bien séparables linéairement. Cette approche aussi simple soit-elle, fonctionne assez bien sur ce type de problème. Le système se base également sur la fréquence d'occurrence de mots ou groupes de mots « suspects » que l'on retrouve dans les « Spams ». Plus récemment, les méthodes d'apprentissage profond ont été utilisées et présentent une très grande sensibilité pour la détection des spams.

Google a porté la détection de spam a un tout autre niveau, notons que Google utilise l'IA depuis 2015 pour la détection des spams à l'aide du deep learning pour sa messagerie Gmail et a annoncé en 2019 avoir atteint une précision de 99,99 % dans le blocage des spams envoyés sur les boîtes mails de ses utilisateurs (Larue, 2022). L'IA est également déployée dans le système « Google Play Protect » qui analyse les menaces dans les programmes et crée des alertes protégeant ainsi le « *Google Play Store* » depuis la même année. (Cette entreprise utilise aussi l'IA sur ses infrastructures de « cloud computing » pour détecter les attaques de dénis de services DDoS et bloquer les connexions suspectes).

7.3 La détection des « Spams d'image » dans les courriels

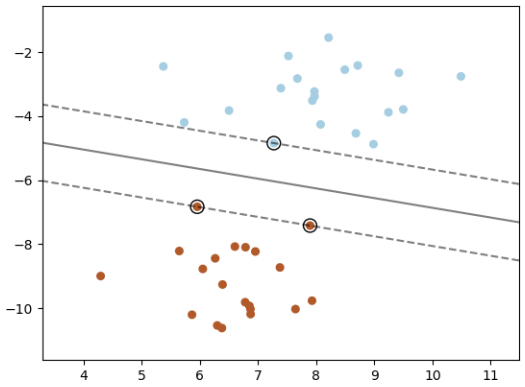
Le spam image est une forme de spam dans laquelle le texte du message est incorporé dans une image, de manière à contourner les systèmes de filtrage tels que nous les avons décrits ci-dessus, ces derniers étant basés sur des algorithmes d'analyse du texte. L'image est - la plupart du temps - insérée directement dans le corps d'un message HTML, et non en pièce jointe, de sorte qu'elle apparaisse directement lorsque le destinataire ouvre le message.

La détection de ce type de spam peut se faire essentiellement selon deux stratégies :

- **Le filtrage basé sur le contenu** : l'approche consiste à essayer d'identifier les mots-clés suspects les plus couramment utilisés dans les spams textuels, même au sein des images. Pour ce faire, des techniques de reconnaissance de formes tirant parti de la technologie de reconnaissance optique de caractères (OCR) sont mises en œuvre afin d'extraire le texte des images. Cette solution est – par exemple - utilisée par SpamAssassin.

- **Le filtrage non-basé sur le contenu** : dans ce cas, il s'agit d'identifier les caractéristiques spécifiques des images de spam (telles que les caractéristiques de couleur, etc.) car les images de spam, étant générées par ordinateur, elles présentent des caractéristiques différentes par rapport aux images naturelles. Pour l'extraction des caractéristiques, des techniques de reconnaissance avancée basées sur des réseaux de neurones et d'apprentissage en profondeur sont utilisées.

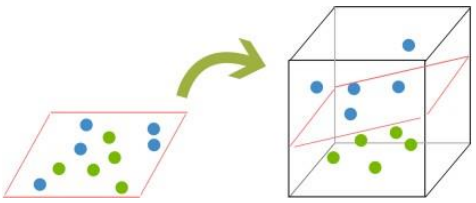
Une fois que les caractéristiques des images sont extraites, les machines à vecteurs de support (SVM) peuvent être utilisées pour résoudre ce type de problème et combattre cette stratégie d'évasion. Pour aller vite, les SVM sont un ensemble de techniques d'apprentissage supervisé qui ont pour objectif de trouver, dans un espace de dimension N>1[25], l'hyperplan qui divise au mieux un jeu de données en deux.



Les SVM sont des séparateurs linéaires, c'est-à-dire que la frontière séparant les classes est une droite.

Le but est d'identifier la frontière la plus éloignée de tous les points d'entraînement qui est optimale, et qui présente donc la meilleure capacité de généralisation (cf. illustration ci-contre).

Les « Support Vector Machines » font appel à des « noyaux », c'est-à-dire des fonctions mathématiques permettant de projeter les données dans un espace de plus grande dimension afin de rendre les données linéairement séparables (cf. illustration contre). Les "vecteurs de support" sont les données les plus proches de la frontière (cf. illustration dessus).



7.4 Détection des URL de « phishing »

Tout d'abord, précisons la différence entre le « Spam » et le « Phishing ». Le « Spam » est simplement un courrier indésirable. C'est par exemple un courriel publicitaire dit « envahissant ».

Les attaques de phishing des « hameçonneurs » visent à voler les données des utilisateurs et les utiliser à leur dépens. La principale différence entre le spam et le phishing est que les « spammeurs » n'ont pas d'intention malveillante.

L'approche classique de détection d'URL de « Phishing » est basée sur la création et le maintien d'une liste noire (l'ensemble des URL dangereuses identifiées) obtenue par des rapports d'utilisateurs ou des avis manuels. Pour rester pertinente, cette base de données des URL de « Phishing » doit être mise à jour fréquemment. Cependant, le nombre d'URL malveillantes ne figurant pas sur la liste noire augmente régulièrement. L'usage de l'IA/ML apparaît comme une bonne solution à l'explosion du nombre de sites de « Phishing ».

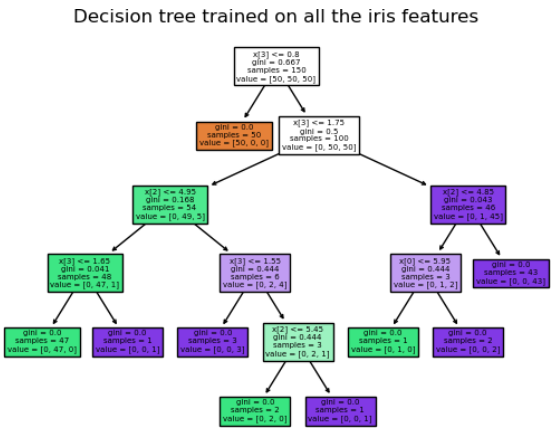
[25] Désigne d'abord chacune des grandeurs d'un objet. Par exemple, notre espace comporte les trois dimensions : longueur, largeur et la profondeur (nous oublions le temps volontairement). Par extension, il est possible de définir des espaces de dimension N arbitrairement grande.

Le problème de la détection des courriels/URL de « Phishing » est équivalent à celui de la détection des « Spams ». Il s'agit d'un problème de classification qui peut être adressé en utilisant par exemple la régression logistique ou bien les arbres de décision. La description du problème peut se faire à l'aide de l'analyse des URL et des contenus des pages HTML associées. Par exemple, l'existence de caractères « // » dans le chemin d'accès de l'URL signifie que l'utilisateur sera redirigé vers un autre site Web. Le symbole du tiret est rarement utilisé dans les URL légitimes. Les « hameçonneurs » ont tendance à ajouter des préfixes ou des suffixes séparés par (-) au nom de domaine afin que les utilisateurs aient l'impression d'avoir affaire à une page Web légitime. La présence d'un @ dans l'URL (permettant d'ignorer ce qui précède) est également le signe d'une URL suspecte. Sur la base du fait qu'un site Web de phishing vit pendant une courte période, les domaines dignes de confiance sont régulièrement payés plusieurs années à l'avance, etc.

Il est ainsi possible de sélectionner plusieurs dizaines de variables numériques ou numérisables pouvant alimenter un algorithme d'IA/ML. Ce travail a par exemple été réalisé par Mohammad R.[26] et ses collaborateurs (Mohammad, et al., 2015).

La régression logistique (encore appelé linéaire) peut être utilisée pour estimer la probabilité qu'une observation appartienne à une classe particulière. Le principe est que si la probabilité estimée est supérieure à 50 %, alors le modèle prédit que l'observation appartient à cette classe, dans le cas contraire, il prédit qu'elle appartient à l'autre classe. À l'instar des méthodes bayésiennes, l'algorithme de régression logistique a une interprétation probabiliste. Elle donne de bon résultat sur ce type de problème.

Une autre approche qui donne de très bons résultats consiste à utiliser les arbres de décision (binaires). Une illustration d'arbre est présentée ci-contre.



La méthode générale se décline comme suit :

- Déterminer la meilleure caractéristique dans l'ensemble de données d'entrainement.
- Diviser les données d'entrainement en sous-ensembles contenant les valeurs possibles de la meilleure caractéristique.
- Générer de manière récursive de nouveaux arbres de décision en utilisant les sous-ensembles de données créés.
- Lorsque l'algorithme ne peut plus classer les données, il s'arrête.

Parmi les avantages des arbres de décision, nous pouvons citer :

- La simplicité de compréhension et d'interprétation.
- L'explicabilité des résultats.
- Le faible niveau de préparation des données.
- L'utilisation de données numériques et/ou nominales.

[26] Mohammed R. est chercheur à l'Université d'Huddersfield en Grande Bretagne.

- Les bonnes performances et la robustesse.

Les limites :

- Le risque de surapprentissage induisant un manque de généralisation de l'algorithme.
- Les problèmes de stabilité parfois.
- Les biais d'apprentissage par rapport aux classes dominantes.

7.5 Détection des malwares

Il existe essentiellement trois techniques pour détecter un malware : les bases de signatures, l'analyse statique et l'analyse dynamique. Elles ont toutes leurs limites mais l'augmentation exponentielle du nombre de malwares rend l'utilisation des signatures de plus en plus difficile. **L'annexe « Typologie des malwares » décrit les principales familles de malwares**.

7.5.1 L'évolution des malwares

La généralisation des malwares polymorphes et/ou métamorphes rend les systèmes antivirus traditionnels de plus en plus perméables à ces menaces. Nous définissons ci-dessous rapidement les concepts de malwares polymorphes et métamorphes :

- Les malwares polymorphes évoluent constamment en modifiant légèrement leurs codes. Ils se composent généralement de deux éléments. L'un d'entre eux reste inchangé, tandis que l'autre est légèrement modifié. Ces changements peuvent s'opérer grâce à une compression ou un chiffrement du code au moyen de différentes clés.

- Les malwares métamorphiques réécrivent automatiquement et entièrement leurs codes à chaque fois qu'ils créent une nouvelle variante d'eux-mêmes. Ce mécanisme de récriture du code se fait principalement en supprimant les signatures que les systèmes traditionnels recherchent. Ce virus a recours à différents types de méthodes de transformation de codes, notamment :

    - Le renommage des registres
    - La permutation du code
    - L'extension du code
    - La compression du code
    - L'insertion d'un faux code

7.5.2 Stratégies de détection des malwares

Parmi les activités les plus communes concernant la détection des malwares, nous pouvons par exemple citer :

- Le calcul du « hash » de fichier qui permet d'identifier les menaces répertoriées à partir d'une base de connaissances.

- La supervision système pour identifier des comportements anormaux à la fois au niveau « hardware » et au niveau du système d'exploitation (par exemple une augmentation de la charge CPU, une augmentation anormale des écritures sur disque, des changements au niveau de la base des registres, etc.).

- La supervision du réseau afin par exemple d'identifier l'établissement de connexions inhabituelles depuis les hôtes du réseau vers des destinations extérieures.

L'ensemble de ces activités sont assez aisément automatisables en utilisant des algorithmes spécifiques.

7.5.3 Analyse statique des malwares

L'analyse statique consiste à évaluer la présence d'artefacts malveillants dans les fichiers binaires en évaluant le code sans l'exécuter. La stratégie consiste le plus souvent à :

- Analyser les instructions contenues dans le fichier binaire.
- Étudier les instructions pour déterminer la présence d'appels système jugés dangereux, en fonction de la séquence dans laquelle ces appels sont invoqués.
- Étudier l'absence d'appel à certaines API (comme l'appel aux API réseaux).
- Rechercher des informations sensibles (comme des IP et/ou des noms de domaine).

L'analyse statique est intéressante et rapide à conduire et se prête bien à l'automatisation. Cependant, elle présente certaines limites particulièrement lorsqu'il s'agit d'analyser un malware polymorphique/métamorphique.

7.5.4 Analyse dynamique des malwares

L'analyse dynamique permet de dépasser les limitations de l'analyse statique des fichiers au moins dans une certaine mesure. Il s'agit d'exécuter le binaire et d'analyser le comportement de l'exécutable dans un environnement sécurisé (une machine virtuelle découplée du réseau ou un environnement bac à sable isolé). L'analyste vérifie par exemple que l'exécutable ne télécharge pas de librairies suspectes ou du code depuis internet, ou bien encore s'il cherche à modifier ses instructions à chaque exécution.

7.5.5 Contre-mesures vis-à-vis des analyses statiques ou dynamiques

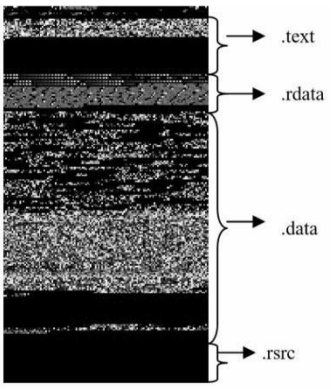
Les pirates adoptent depuis longtemps des contre-mesures pour échapper aux systèmes de détection, puisqu'il s'agit d'une course sans fin entre les pirates et les équipes en charge de la cybersécurité des organisations.

Les contre-mesures adoptées par les développeurs de malwares, permettant de réduire les possibilités de détection au cours d'une analyse, reposent notamment sur le chiffrement du code malveillant, la partie de téléchargement peut être encapsulée dans du code spécifique pour ne pas être détectée. Bien que beaucoup des contre-mesures soient détectables par une analyse dynamique, cette dernière présente également des limitations liées à l'usage de machines virtuelles. En effet, un malware peut assez simplement détecter qu'il est exécuté sur une machine virtuelle et modifier son comportement en conséquence (en exécutant par exemple certains appels qui s'exécutent directement au niveau hardware, en accédant à certaines bases de registres ou encore en mesurant les temps d'exécution). La détection d'un environnement virtuel peut entrainer par exemple l'arrêt d'exécution du malware afin d'échapper à l'analyse.

7.5.6 Détection simple des malwares

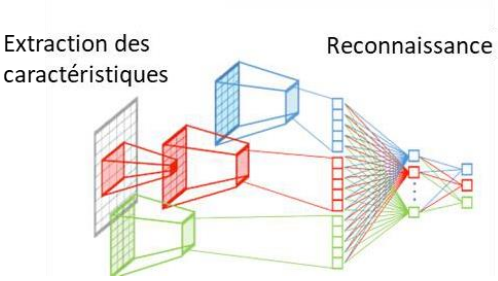
Parmi les algorithmes pouvant être utilisés, les arbres de décision présentent d'assez bonnes performances. **Pour nous en rendre compte, nous avons mené une petite expérimentation dans le cadre de ce projet, afin d'étudier les performances de plusieurs implémentations d'arbres de décision et en mesurer les difficultés d'implémentation. Le lecteur intéressé peut se reporter à l'annexe intitulée « Apprentissage pour la détection statique de malwares, méthodologie et implémentation »**. Cette expérimentation a pour objectifs d'illustrer techniquement les concepts mis en œuvre et de montrer que même des non-experts peuvent s'approprier ce type de technologie.

7.5.7 Détection avancée des malwares avec l'usage du « deep learning »



En 2011, une nouvelle approche d'analyse statique des malwares sous forme d'images voit le jour à l'université de Californie - Santa Barbara (Nataraj, et al., 2011). La méthode consiste à transformer les exécutables en fichiers d'images, en associant chaque octet à un pixel monochrome dont l'intensité est comprise entre 0 et 255 (cf. illustration ci-contre). Par la suite les images sont utilisées pour entrainer un réseau de neurones convolutif (CNN).

Les réseaux de neurones convolutifs sont directement inspirés du cortex visuel des vertébrés (cf. illustration ci-dessous). Il s'agit d'un réseau de neurones classiques multi couches, couplé en amont à plusieurs couches de convolution. La convolution est une opération mathématique utilisée dans le cadre de l'analyse d'image. Elle permet de faire de la reconnaissance de forme et d'extraire les caractéristiques d'un ensemble d'apprentissage. La convolution a également pour effet de réduire la dimension de « la carte de caractéristiques » que l'on obtient après convolution et donc de réduire la dimensionnalité du problème sans perte substantielle d'information.



Sur un ensemble d'apprentissage de 9 458 exemples provenant de 25 familles de malwares, les chercheurs sont arrivés à un taux d'exactitude de 98 %. Par ailleurs, l'avantage est que le taux de détection est insensible à l'obscurcissement et au chiffrement du code.

Sur la base de ces recherches, Microsoft et Intel ont développé le projet « Stamina » (STAtic Malware-as-Image Network Analysis). Les chercheurs ont modifié le dispositif de *deep learning* par transfert, en optimisant la phase d'apprentissage et en ajoutant à la transformation en images des techniques de segmentation et de redimensionnement. Leur modèle a été testé sur un ensemble de 2,2 millions d'exécutables de tailles diverses. Le taux de précision observé sur cet ensemble d'apprentissage est de 99,07 % (Kallenborn, 2020).

Les avantages de ce type de méthodes résident dans les éléments suivants :

- Ces méthodes ont la capacité de reconnaitre des sections spécifiques de codes malveillants comme les parties destinées à créer différents variants à partir du code original et reconnaitre plus aisément les malwares polymorphiques ou métamorphiques.

- Il est également possible d'identifier des modifications mineures car les altérations ne vont concerner qu'une partie de l'image. La structure globale de l'image ne sera pas affectée, ce qui rend l'échappement des malwares polymorphiques ou métamorphiques plus difficile.

- Les familles de malwares présentent des similarités au niveau des images qui sont calculées, ce qui rend également les variétés d'un même malware fortement reconnaissable.

7.6 Détection d'intrusion Réseau (IDS)

7.6.1 Les différents types d'IDS

La détection d'intrusion (IDS) permet de détecter les activités anormales/suspectes sur un réseau et/ou sur un hôte (ordinateur). Il existe trois familles distinctes d'IDS :

- Les NIDS (Network Based Intrusion Detection System), qui surveillent l'état de la sécurité au niveau du réseau.

- Les HIDS (Host Based Intrusion Detection System), qui surveillent l'état de la sécurité au niveau des hôtes.

- Les IDS hybrides, qui utilisent les NIDS et HIDS pour avoir des alertes plus pertinentes.

7.6.2 Les bases de signatures

La détection d'intrusion a longtemps reposé comme pour les anti-malwares sur la détection de signatures d'attaques. Pour ce faire, une base de données de signatures doit être mise en place, maintenue et régulièrement mise à jour.

Par exemple, l'outil « open source » SNORT est un NIDS largement utilisé et qui permet de détecter d'éventuelles intrusions sur le réseau où il est installé (cf. https://www.snort.org/). Il s'agit probablement du système IDS « open source » le plus utilisé au monde. SNORT se base sur des règles pré-enregistrées pour savoir à quel type de paquet s'intéresser et comment les interpréter. SNORT a plus tendance que d'autres IDS à fournir de fausses alertes car en moyenne, 70 % des alertes remontées sont fausses, notamment à cause des petites signatures. Ce type d'outil ne permet par ailleurs pas de traiter les flux chiffrés.

7.6.3 La détection d'anomalie

Une autre approche consiste à faire de la détection d'anomalie. Cette méthode consiste à identifier le trafic réseau « normal » afin d'être en mesure de détecter les activités suspectes. Ce type d'approche devrait permettre de mieux adresser les attaques nouvelles ou inconnues.

Dans le cas des NIDS, la détection d'un comportement anormal peut être étudiée notamment en supervisant les éléments suivants :

- Le nombre de connexions effectuées depuis ou à destination d'un hôte spécifique.
- Des communications sur des ports inhabituels.
- Des pics de trafic se produisant à des heures précises (par exemple la nuit).
- Une augmentation de la consommation de la bande passante générée par des hôtes spécifiques.
- La latence du réseau.
- Les débits réseaux (entre différents hôtes).

L'étude de ces métriques permet par la suite de définir des seuils de déclenchement d'alarme et de « rerouter » le trafic suspect.

Dans le cas d'un système HIDS, la détection d'un comportement suspect peut être adressée en monitorant les métriques suivantes :

- Le nombre et le type de processus qui s'exécutent.
- La détection de nouveaux processus.
- Le nombre, le type et la création de comptes utilisateurs.
- Le chargement de modules spécifiques au niveau du noyau (« drivers » inclus).
- L'activité au niveau des fichiers et répertoires.
- L'activité de l'ordonnanceur des tâches.
- La modification des clés de registre.
- L'activité réseau de l'hôte.

7.6.4 IDS basés sur l'IA/ML

L'introduction des techniques d'IA/ML dans le domaine des IDS permet d'avoir des systèmes plus évolués se basant sur les algorithmes d'apprentissage non-supervisé, supervisé et par renforcement ainsi que l'apprentissage profond. Ces techniques permettent d'implémenter des solutions basées sur la détection d'anomalies.

Le développement de solutions IDS basé sur un apprentissage supervisé demande de disposer de données labellisées, c'est-à-dire d'un ensemble d'apprentissage dans lequel les classes sont définies. Cette étape représente une quantité de travail assez colossale, qui est la principale difficulté à résoudre lors du développement d'un tel système. Ici encore, il est possible d'utiliser les algorithmes d'apprentissage tels que les arbres de décision dans toutes leurs variantes, les SVM et les réseaux de neurones.

Les données proviennent du trafic réseau et des fichiers de « logs » systèmes. Elles peuvent être centralisées sur des solutions d'indexation telles qu'ElasticSearch, Logtash ou Kibana (ELK). Signalons également la solution de supervision propriétaire Cisco Netflow qui permet de présenter le trafic réseau de manière très compacte et structurée.

Les systèmes d'apprentissage non supervisé ou d'analyse statistique (clustering) s'avèrent intéressant pour leur rapidité de traitement mais également pour identifier de nouvelles classes de trafic et/ou d'action. Les techniques de « clustering » permettent de regrouper les paquets selon leur « ressemblance ». Ces systèmes de détection d'anomalies reposent sur un système de scores et l'identification de seuils pour classer le trafic (normal versus suspect). Il s'agit d'évaluer une distance entre « un point » (en toute rigueur un vecteur) et le centre d'une classe ou encore de calculer l'écart d'un point par rapport à la distribution des données.

Les approches statistiques ne nécessitent pas la connaissance préalable des activités normales. Elles permettent d'obtenir une détection précise sur de longues périodes si les seuils d'alertes sont correctement paramétrés.

En tout état de cause, les IDS à détection d'anomalie présentent le risque d'avoir un niveau de faux positifs importants rendant leur usage délicat. À ce titre, il semble important que les équipes cybersécurité utilisent ces solutions en tant que filtre de niveau 1[27], afin de sélectionner les anomalies devant faire l'objet d'une analyse subséquente.

[27] Ce sont les opérateurs en charge de relever les alertes et de faire un premier diagnostic. C'est ici que l'opérateur analyste SOC intervient.

En pratique, de nombreux systèmes de détection (et de prévention des intrusions) combinent à la fois la détection des signatures et celle des anomalies. La détection basée sur les anomalies peut potentiellement (et avec un peu de chance) détecter les menaces « Zero Day », mais peut souffrir de taux élevés de faux positifs car elles alertent sur tout ce qui est anormal.

7.6.5 Quelques exemples d'IDS utilisant l'IA/ML

Il est possible de citer au moins deux IDS commerciaux faisant usage de l'IA/ML (Samson Jr , 2022) :

- BluVecteur anciennement connu sous le nom de Cortex et détenu par Comcast, la solution de détection des menaces de BluVector utilise l'IA pour détecter les logiciels malveillants sans fichier et potentiellement certaines menaces « Zero-day ». Il serait conçu pour devenir plus « performant » avec le temps, toujours d'après l'entreprise.

- Vectra Cognito est la plate-forme Cognito IPS de Vectra. Elle utilise également l'IA pour analyser le trafic provenant de sources variées.

7.7 Détection des menaces internes (UBA/UEBA)

Les équipes de cyberdéfense sont souvent focalisés sur les menaces externes alors que les menaces internes tendent à être sous-estimées. Les menaces internes sont le fait d'employés de l'entreprise susceptibles d'avoir accès à des ressources protégées. Les motivations pour une telle action peuvent être d'ordre financier, la vengeance, l'espionnage industriel, etc. Ces menaces concernent :

- Le vol de données.
- L'abus de privilèges.
- Escalade de privilèges.
- Le sabotage.

Ces menaces sont généralement difficiles à contenir et il faut en moyenne 77 jours pour détecter un incident interne, selon l'enquête de l'Office européen de lutte antifraude (OLAF).

Les outils d'analyse comportementale des utilisateurs privilégiés (UBA/UEBA) permettent de comprendre leurs comportements en créant des profils individualisés pour chaque utilisateur et en utilisant des algorithmes d'IA/ML.

Les profils sont réalisés grâce aux traces dans les fichiers de « logs » générés par les utilisateurs (tous les événements recueillis par le système d'exploitation, les applications, etc.), et leur analyse par des outils de surveillance des activités, sur une période allant de 4 à 12 semaines. Il est alors possible de comparer les bases de profils avec les activités en temps réel et ainsi générer des alertes sur les événements inhabituels ou les comportements atypiques. Les sources de données retenues en général sont par exemple :

- Les logs de connexion pour suivre les connexions des profils sur les différentes machines ainsi que la date et l'heure de connexion pour vérifier s'il y a eu des connexions hors des heures standards de travail.

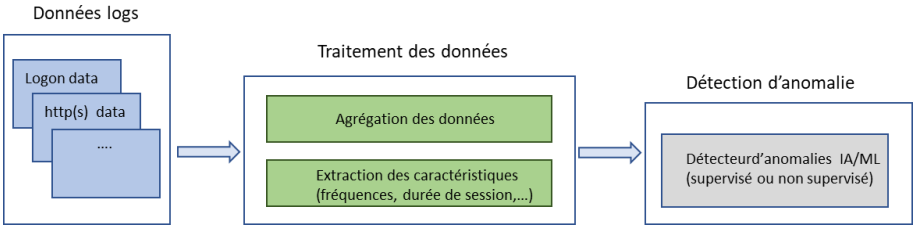
- Les logs d'accès aux fichiers pour tracer les tentatives d'accès aux données confidentielles par exemple ou le suivi des transferts de fichiers.

- Les logs de courriels (logs des serveurs SMTP) pour identifier les communications avec les concurrents ou les tentatives de phishing.

- Les logs DNS et requêtes Http/Https pour analyser les sites web consultés par les employés.

La plupart des approches utilisent des caractéristiques de fréquence (par exemple le nombre d'actions sur une période, le nombre de connexions, le nombre d'ordinateurs utilisés, le nombre de courriels envoyés, le nombre de connexions de périphériques USB inconnus, etc.) ainsi que des caractéristiques statistiques (par exemple, l'heure de la première et de la dernière connexion, la moyenne et écart type de la taille des données transmises, la durée moyenne d'une session, etc.).

Le mode de fonctionnement des solutions de détection d'anomalie commence par la collecte des fichiers logs provenant de plusieurs sources. Le système analyse et agrège les entrées des fichiers logs par types de fonctions pour chaque utilisateur et chaque jour. Les vecteurs de caractéristiques décrivant les jours-utilisateurs sont ensuite traités par un détecteur d'anomalies, qui génère un score d'anomalie pour chaque instance. Souvent, les scores d'anomalies sont normalisés. Les instances sont classées en fonction de leurs scores d'anomalies.



Parmi, les nombreux algorithmes d'IA/ML qui sont assez souvent utilisés, nous retrouvons le iForest (Isolation Forest). Cet algorithme est utilisé pour sa capacité à détecter des anomalies de façon non supervisée. De la même manière que le Random Forest, l'algorithme crée une forêt composée de dizaines ou centaines d'arbres dont les résultats seront combinés afin d'obtenir une plus grande précision dans la détection. En fonction de la distance à la racine, la méthode associe à chaque feuille d'un arbre, un score d'anomalie qui est moyenné sur l'ensemble des arbres qui ont été construits.

Les méthodes des réseaux de neurones profonds sont également utilisées pour ce type d'application dans une approche d'apprentissage non supervisé (Jääskelä, 2020). La description précise du fonctionnement de ces méthodes dépasse le cadre de ce travail.

7.8 Sécurisation de l'authentification des utilisateurs

Les techniques d'IA/ML sont utilisées pour la protection de l'identification et des moyens d'authentification des utilisateurs. Elles sont plus particulièrement utilisées dans les domaines tels que :

- La prévention des fraudes à l''authentification.
- La notation de la réputation du compte.
- L'authentification par l'empreinte de frappe au clavier.
- L'authentification via les données biométriques.

7.8.1 Prévention des fraudes à l'authentification

Avec le développement du tout internet et notamment la multiplication des objets connectés (IoT), les possibilités d'accès à l'aide d'identifiants frauduleux (ou d'identifiants volés) se multiplient plus que jamais. La protection des comptes utilisateurs ou clients concerne non seulement la protection de l'intégrité des données, la protection contre le vol mais aussi la protection de la réputation des organisations ainsi que la maitrise des risques juridiques associés.

Dans ce contexte, il est légitime de se poser la question de l'obsolescence de l'usage des mots de passe pour l'authentification sachant qu'après tout il ne s'agit là que d'une chaine de caractères de complexité variable. L'authentification par mots de passe a depuis longtemps été renforcée par la mise

en place d'une authentification multi-facteurs (par exemple, l'envoi d'un code via SMS sur un numéro unique de téléphone portable ou d'un courriel sur une boite mail secondaire, la génération d'un code spécifique à partir d'une application de génération de code, etc.). Ce type d'authentification forte améliore les choses et réduit la surface d'attaque mais ne constitue nullement une arme absolue contre les cyberattaquants. Elle obéit néanmoins au principe de défense en profondeur puisqu'elle superpose plusieurs couches d'authentification.

La robustesse de l'authentification dite forte repose sur la diversification des supports utilisés. Autrement dit, il est supposé que l'utilisateur ne garde pas toutes ces informations sensibles sur un même support. Il peut se produire que ce principe de diversification ne soit pas suivi dans la réalité, dans un tel cas l'efficacité de l'authentification multi-facteurs est fortement réduite.

Bien évidemment, la protection des comptes utilisateurs/clients ne doit pas se limiter à la simple vérification de la correspondance d'un compte et du mot de passe associé, mais doit aussi s'appuyer sur les activités du compte telles qu'elles sont collectées comme le suivi de la localisation d'accès à partir de l'IP publique. Cette dernière peut par exemple ne pas correspondre à la zone géographique habituelle de connexion. Il est possible de s'appuyer sur les informations liées à l'ordinateur depuis lequel la connexion est initiée (s'agit-il du même « device » qu'habituellement ?). L'authentification peut ainsi être supportée par tout un contexte venant la renforcer. Ce type de vérification ne peut se faire qu'à travers la supervision des activités « normales » des utilisateurs/clients avec la mise en place d'un système de détection d'anomalies de type UBA/UEBA que nous avons évoqué dans la section précédente pour les menaces internes. Parmi les anomalies relatives à la gestion des mots de passe et sans être exhaustif, il est possible de suivre et détecter :

- Les attaques de type « brute force » qui permettent de trouver un mot de passe à partir de multiples tentatives reposant sur l'application d'un dictionnaire. Différents mots de passe sont alors entrés durant une période limitée.

- Les accès simultanés (ou non) à partir d'adresses IP appartenant à des zones géographiques différentes.

- L'utilisation de matériels, de logiciels ou de système d'exploitation qui ne sont pas utilisés de manière usuelle par les collaborateurs/utilisateurs/clients.

- Une fréquence de frappe au clavier incompatible avec les capacités d'un utilisateur humain.

Comme nous le verrons ultérieurement, il est possible sinon de remplacer l'authentification par mots de passe du moins de la renforcer par des procédures d'authentification de type biométriques (empreintes de l'iris, voix, empreintes, reconnaissance faciale). En l'espèce, il convient de ne pas se limiter à une seule empreinte biométrique qui demeure toujours falsifiable, car chaque méthode présente des faiblesses intrinsèques. Le concept de défense en profondeur reste applicable.

7.8.2 Approche réactive versus prédictive

La mise en place d'alarmes sur seuil en réaction à la détection d'un possible accès frauduleux permet de bloquer et/ou suspendre automatiquement le compte impliqué. Cependant cette approche réactive bien que relativement simple à implémenter présente certaines limites. Il existe toujours la possibilité d'être la cible d'une attaque par dénis de service ciblant des utilisateurs légitimes. Le cyberattaquant peut ainsi endommager la réputation de l'organisation en simulant des accès non autorisés afin de bloquer délibérément les comptes des utilisateurs/clients. L'effet est de créer une interruption de service pour l'organisation et ses clients. Certaines attaques sophistiquées sont menées en mode furtif afin de rester sous les seuils de détection. L'attaquant peut rester caché à

l'intérieur du système d'information de l'organisation et mener ses activités. Par exemple Yahoo à travers son service de courrier électronique a fait l'objet d'un des piratages les plus importants de l'histoire qui a mis plusieurs années avant d'être détecté.

La stratégie de lutte contre la compromission des comptes d'utilisateurs devrait tenir compte des changements de contexte et de scénarii susceptibles d'affecter à la fois le comportement de l'utilisateur et de l'attaquant. Ceci nécessite l'adoption d'une approche prédictive de la détection des anomalies, qui commence par l'analyse des données passées afin d'extrapoler les comportements de l'utilisateur et d'identifier à temps d'éventuelles tentatives de compromission ou de fraude.

Le but de l'analyse prédictive est de révéler les « patterns » cachés, d'identifier les tendances latentes à travers l'analyse des données qui peut se faire via une combinaison de techniques issue des statistiques et de l'IA/ML. Les techniques d'apprentissage non supervisé ou de « clustering » sont particulièrement adaptées pour l'exploration des motifs (patterns) de comportement et détecter des comportements qui s'écartent de la « normalité » ou d'une ligne de base. Bien évidement lorsque l'on dispose de données d'apprentissage labélisées, l'apprentissage supervisé peut s'appliquer (par exemple les SVM, les arbres de décision dans toutes leurs versions ainsi que les différentes classes de réseaux de neurones).

7.8.3 Choix des métriques

Le choix des métriques à superviser est un problème très délicat et il dépend du type de menaces adressées. Ainsi dans le cas d'une attaque sur mot de passe par « brute force », il peut être suffisant de suivre le nombre de tentatives de connexion en échec, le suivi peut se faire en suivant le taux de croissance dans le temps des échecs de connexion et ses variations au cours du temps. Il est également possible de suivre les fréquences de changement des mots de passe ou de leur réinitialisation.

Il est bien plus délicat de suivre les attaques « furtives » puisque dans ce cas l'attaquant dispose déjà des paramètres d'authentification, à la suite d'une compromission qui a eu lieu très en amont de l'opération. Dans ce cas, il est plus utile de superviser les adresses IP de connexion associées au paramètre de connexion afin de vérifier que l'accès ne se fait pas depuis une zone géographique inhabituelle ou qu'il existe de multiple connexion en provenance de zones géographiquement éloignées. Bien évidemment, les cyberattaquants passent le plus souvent par de multiples machines de rebond.

7.8.4 Prévenir la création des faux comptes

La création des nouveaux comptes associés à des profils douteux est une activité qui doit également être supervisée. Un indicateur possible de cette activité pourrait être la création de comptes multiples depuis la même IP sur une période courte.

Un indicateur pouvant trahir un faux compte/profil pourrait être un nombre de « post » élevé sur une période assez courte.

7.8.5 Notation de la réputation d'un compte ou d'une entité

La supervision des comptes utilisateurs doit être appliquée à la création des nouveaux comptes mais aussi au suivi des comptes existants. Il est recommandé de mettre en place une notation de la réputation des comptes basée sur leurs activités. L'évolution même faible du score de réputation peut aider à la détection des attaques « furtives ». La création de ce type de notation peut se faire sur la base des éléments suivants :

- Le nombre et la fréquence des postes (pour une plateforme publique).

- Le nombre d'accès via proxy, VPN ou tout autre système d'anonymisation.
- L'utilisation d'agent de connexion.
- La vitesse/fréquence de frappe au clavier.

À noter que ces scores s'appliquent également aux serveurs de mail, aux noms de domaines, aux adresses IP, etc.

7.8.6 Classification des activités utilisateurs

La collecte de l'ensemble des données de supervision permet d'alimenter les algorithmes d'apprentissage d'IA/ML.

Les algorithmes d'apprentissage supervisé peuvent être utilisés à condition d'avoir la capacité de labéliser les données collectées. Ce travail de labellisation est un travail difficile et très fastidieux puisqu'il n'est que partiellement automatisable. La détection des exemples suspicieux et bénins ne peut se faire dans un premier temps que sur la base de règles déjà implémentées par les équipes de cybersécurité, soit de règles utilisées manuellement par ces mêmes équipes. Ainsi la constitution des ensembles ne couvrira en définitive que les connaissances des équipes en cybersécurité et les règles plus ou moins implicites déjà connues. Il est dans ce contexte assez difficile de couvrir de nouveaux types de comportements suspicieux. Toute la difficulté réside dans la constitution des ensembles d'apprentissage ainsi que dans la pondération des classes.

Les algorithmes d'apprentissage non supervisé permettent de créer des groupes/ensembles de comptes/entités homogènes basés sur les données de suivi de leurs activités (fréquence des « posts », temps passé sur la plateforme, fréquence de connexion, typologie des accès, plage d'IP de connexion, noms de domaines, etc.). Ces méthodes de « Clustering » permettent de travailler sur des données non labellisées et de construire des ensembles de données classées par niveau de similarité. La difficulté reste de définir le nombre d'ensembles à construire, par la suite il reste un travail d'analyse par les équipes (data scientists et équipes en cyberdéfense). Le travail consiste à déterminer les groupes caractéristiques des activités suspicieuses ou l'ensemble des groupes relevant d'une activité normale. Une fois cette ligne de base déterminée, tout écart vis-à-vis des ensembles de données « normales » sera considéré comme une anomalie potentielle et redirigée vers les équipes en cybersécurité pour une investigation plus poussée. Il conviendra de tester de multiples algorithmes de « clustering » pour déterminer lesquels sont les plus adaptés au problème.

7.9 Authentification des utilisateurs par la dynamique de frappe au clavier

Compte tenu des faiblesses des méthodes d'authentification et notamment celles reposant uniquement sur un « login/password », les organisations restent à la recherche de méthodes plus efficaces d'authentification. L'authentification par la biométrie concerne la reconnaissance de caractéristiques spécifiques à un utilisateur. Parmi toutes les caractéristiques biométriques, la dynamique de frappe au clavier est particulièrement intéressante. Elle reste difficile à imiter et permet de procéder à une authentification en continu (Giot, et al., 2014).

Sans se substituer à l'authentification par « login/password » ou mieux à l'authentification forte, la dynamique de frappe permet de sécuriser encore plus le processus d'authentification en vérifiant la manière de saisir les informations d'identification. La méthode peut également être utilisée pour sécuriser la session après son ouverture en détectant le changement de comportement de frappe. Dans ce cas, on parle d'authentification continue, l'ordinateur « sait » comment l'utilisateur interagit avec son clavier. Il est capable de reconnaître si un autre individu utilise le clavier, car la manière d'interagir avec lui est différente. De plus, la dynamique de frappe peut également empêcher le vol de données ou l'utilisation non autorisée de l'ordinateur par des attaquants. En effet, ce mécanisme permet au système de détecter le changement d'utilisateur pendant la durée de la session. De cette

façon, l'ordinateur est capable de verrouiller la session, s'il détecte que l'utilisateur est différent de celui qui a été précédemment authentifié (Giot, et al., 2014).

La dynamique de frappe au clavier peut être modélisée notamment par la supervision des métriques suivantes (sachant qu'il existe en existe bien d'autres) :

- **Les durées** : la durée est le temps pendant lequel une touche est enfoncée.

- **Les latences** : il existe différents types de latences. Elles sont calculées en obtenant les différences de temps entre deux événements clés. Par exemple, nous pouvons utiliser les différences de temps entre la pression de chaque touche.

- **Les temps totaux de frappe** : le temps total nécessaire pour taper le texte peut être utilisé. Les informations peuvent être utilisées comme une caractéristique supplémentaire à ajouter aux vecteurs de caractéristiques, ou comme facteur de normalisation.

- **Les demi-périodes** : il s'agit de la différence de temps entre le moment où l'utilisateur saisit le caractère au milieu du mot de passe et le moment du début de la saisie.

- **Les taux d'erreurs** : l'utilisateur commet des fautes qui sont détectées en comptant le nombre de fois où la touche de retour arrière est enfoncée.

- **Les minima/maxima** : il s'agit de calculer la valeur minimale et maximale de chaque type de données (latences et durées).

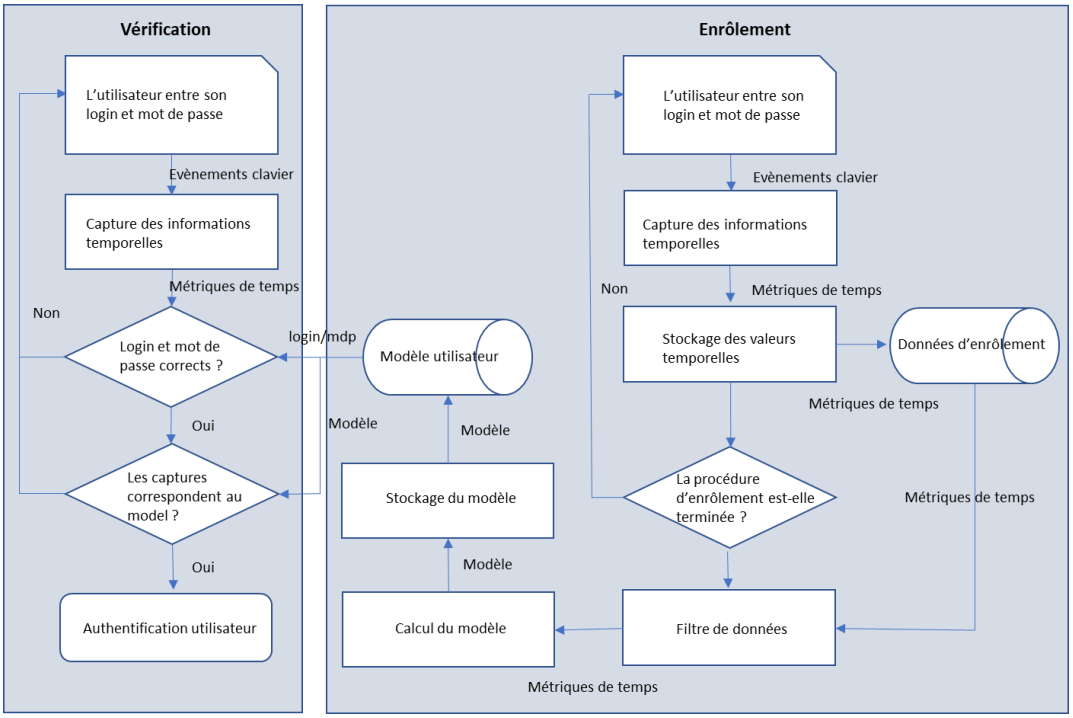
- **Les moyennes et écart types** : c'est le calcul de la valeur moyenne et son écart-type pour chaque type de données (latence et durée).

- **Les pentes (ou dérivée par rapport au temps)** : en utilisant la pente de l'échantillon biométrique, on s'intéresse à la forme globale de la frappe.

L'utilisation des réseaux de neurones (notamment dans leur version « deep learning ») s'avère particulièrement efficace pour ce type de problème.

Le schéma suivant décrit les deux principales étapes pour la réalisation d'un modèle d'authentification par dynamique de frappe au clavier. La première étape consiste à enrôler l'utilisateur. L'étape d'enrôlement permet de créer le modèle de chaque utilisateur, grâce à ses échantillons enrôlés. Cette phase consiste à capturer les informations temporelles relatives à sa dynamique frappe. Les données sont alors filtrées et nettoyées (par exemple suppression des valeurs aberrantes). Un modèle de la dynamique de frappe est alors calculé et stocké. Il servira pour la détection des écarts dans la dynamique de frappe.

Une fois l'enrôlement achevé, la solution a la capacité de procéder à l'authentification et au suivi du comportement utilisateur. C'est la phase de vérification. Les « logins/passwords » sont vérifiés de même que la dynamique de frappe associée. La solution peut superviser en continu le comportement de l'utilisateur.



Source : Giot R. et coll., 2011.

L'utilisation de ce type d'authentification a très tôt été mise en place et prouvée son efficacité. À titre d'exemple, dès 2013, Coursera (la plateforme de formation en ligne) a mis en place ce type d'authentification. Les étudiants peuvent ainsi obtenir des « certificats vérifiés ». La solution a été baptisée « Signature Track » et permet de vérifier l'identité des étudiants effectuant un travail en utilisant un modèle biométrique comprenant une photo et la dynamique de frappe au clavier (Vrankulj, 2013).

7.10 Reconnaissance faciale – Identification et Authentification

La reconnaissance faciale est utilisée pour identifier la structure faciale d'une personne. Ces algorithmes permettent d'identifier les caractéristiques biométriques et faciales uniques (espace entre le nez et la bouche, taille des sourcils, largeur du front, etc.). Ces caractéristiques distinctives sont appelées points nodaux. En moyenne, le visage humain en contient environ 80.

Ces informations analogiques sont converties en code numérique pour former une empreinte faciale. Plus récemment, la biométrie de la peau et du visage s'est développée. Cette méthode permet notamment « d'étudier » la texture de la peau au niveau d'une section particulière. Des mesures sont prises sur les lignes, les textures et les pores de peau. Cette technique permet ainsi de faire la différence entre des vrais jumeaux.

La reconnaissance faciale est utilisée avec succès dans deux contextes différents :

- D'une part, pour faire de **l'identification** : la personne est identifiée parmi d'autres (vérification 1:N[28]). Ses données personnelles sont comparées aux données d'autres personnes contenues dans la même base de données ou dans d'éventuelles bases de données reliées.

- D'autre part dans les procédures **d'authentification**. Il s'agit ici de certifier l'identité d'une personne en comparant les données qu'elle va présenter avec les données préenregistrées de la personne qu'elle prétend être (vérification 1:1[29]). L'authentification par reconnaissance faciale s'est largement répandue ces dernières années. La méthode bénéficie en effet de la diffusion de plus en plus large des réseaux de neurones associés à la généralisation des caméras embarquées sur les ordinateurs, les smartphones et autres tablettes. Le système vérifie si l'identité prétendue est bien la bonne en comparant le modèle du visage présenté au modèle préalablement enregistré. Pour ce faire, un capteur « saisit » le visage, puis le transforme en données numériques par l'opération d'un algorithme et le compare à une base de données.

Il s'agit d'un domaine dans lequel les GAFAM s'illustrent particulièrement (Thales, 2021) :

- L'authentification par reconnaissance faciale sur les « smartphones » (iOS ou Android) est une réalité depuis un certain temps déjà. Le système utilise le visage comme clé de déverrouillage.

- Les premiers développements remontent à 10 ans environ. Ainsi, l'algorithme « GaussianFace » développé dès 2014 par des chercheurs de l'université de Hong Kong affichait des scores d'identification faciale de 98,52 % pour les humains.

- En 2014 également, Facebook annonçait le lancement de son programme « DeepFace », capable de déterminer si deux visages photographiés appartiennent à la même personne, avec une précision de 97,25 %.

- En 2015, Google présentait « FaceNet », un nouveau système de reconnaissance faciale aux scores de 99,63 % de précision au test de référence « Labeled Facebooks in The Wild », 95 % sur la base « YouTube Faces DB ».

- En 2018, Amazon faisait activement la promotion de son service de reconnaissance faciale « Rekognition » basé sur le cloud, auprès des forces de l'ordre. La solution peut reconnaître jusqu'à 100 personnes dans une seule image et peut effectuer des comparaisons avec des bases de données contenant des dizaines de millions de visages.

- Ces dernière années, Thalès a également développé une plateforme de reconnaissance faciale (FRP), la solution de reconnaissance faciale de Thales présenterait un taux d'acquisition de 99,44 % d'un visage en moins de 5 secondes et un taux réel d'identification en moins de 5 secondes de 98 % pour un taux d'erreur de 1 %. La solution a été développée à partir des réseaux neuronaux profonds (réseaux de convolution).

[28] Dans deux tables A et B de relation 1:N, un élément de la table A peut se rapporter à plusieurs éléments de la table B, et un élément de la table B seulement à un élément de la table A.
[29] Dans deux tables A et B de relation 1:1, un élément de la table A se rapporte seulement à un élément de la table B.

7.11 Prévention des fraudes bancaires

L'objectif d'un grand nombre de cyberattaques est le vol de données personnelles et notamment le vol des coordonnées bancaires et des moyens de paiement tels que les cartes bancaires. Uniquement en France, le total de ce type de fraudes pour l'année 2021 s'est élevé à 1,24 milliard d'euros pour un volume de transactions de 42 204 milliards d'euros, concernant 7,5 millions de transactions frauduleuses. Le taux de fraude des cartes était de 0,059 % en 2021 (La finance pour tous, 2022).

Pour lutter contre ce type de fraudes, les institutions financières ont introduit des mécanismes de prévention de la fraude comme la double authentification via l'envoi d'un OTP (One time password) par SMS sur le téléphone portable du client ou bien la demande de validation d'une transaction par internet sur le portable. Ces mesures restent néanmoins insuffisantes pour empêcher les pirates à poursuivre leurs activités.

La détection des fraudes à la carte bancaire est un cas d'usage pour lequel l'IA/ML peut apporter une réelle valeur ajoutée. Le développement de modèles prédictifs à partir de l'analyse « big data » est un axe majeur de développement pour les banques et les sociétés intermédiaires.

Comme d'habitude, différents classificateurs peuvent être utilisés tels que les arbres de décision (RandomForest, XGBoost, AdaBoost) ou encore les réseaux de neurones (dans leurs différentes déclinaisons). Les data scientists testent systématiquement différents algorithmes dans leur phase d'exploration des données.

7.11.1 Les principales difficultés pour l'usage de IA/ML pour la détection de fraudes

Le choix et le développement d'algorithmes de détection de la fraude se heurtent souvent au fait que :

- Les données de transactions frauduleuses ne sont pas disponibles (librement) dans la mesure où les institutions financières ne communiquent pas sur le sujet.

- Du point de vue technique, les données liées aux transactions frauduleuses ne sont pas stationnaires, c'est-à-dire que les propriétés statistiques des fraudes évoluent dans le temps. Ceci s'explique par l'adaptation et les modifications de comportements et habitudes des clients. Le système d'IA/ML doit donc être capable de s'adapter dans le temps aux nouveaux comportements.

- La distribution des ensembles d'apprentissage est par nature déséquilibrée, c'est-à-dire que les transactions frauduleuses ne représentent qu'une infime partie de l'ensemble des transactions. Ceci induit des risques de biais d'apprentissage qui devront être adressés par les techniques d'échantillonnage de données. L'intégration d'experts humains dans la boucle d'apprentissage pour la détection des nouveaux modèles de fraudes est importante pour l'amélioration de la performance de ces algorithmes.

7.11.2 Les principaux scénarii de fraudes

Il existe globalement trois scénarii de fraudes à la carte bancaire que nous pouvons décrire comme suit :

Le vol de la carte bancaire qui est le cas le plus courant. Le voleur vole la carte et dépense autant d'argent que possible sur une période très courte. Ce scénario est très « bruyant » et laisse beaucoup de traces. Les anomalies, qui s'écartent du comportement habituel du détenteur légitime de la carte, sont aisément détectables.

L'utilisation abusive de la carte bancaire. Dans ce scénario le fraudeur ne dispose pas de la carte physique mais possède les informations essentielles associées (numéro de la carte, code PIN éventuellement, code CVC, etc.). Ce type d'attaque est conduit en mode furtif et le propriétaire légitime de la carte n'est pas au courant que ses données personnelles lui ont été dérobées.

L'usurpation d'identité, ce scénario vise à faire éditer une carte bancaire sur la base d'une fausse identité en exploitant des données personnelles volées. Par exemple, avec la simple photocopie d'une pièce d'identité et un justificatif de domicile, l'usurpateur va pouvoir ouvrir des comptes sous le nom de sa victime, en ligne et/ou souscrire à des services financiers particuliers (augmentation des plafonds bancaires, souscriptions à des assurances, etc.).

Il faut garder à l'esprit que les scénarii de fraudes évoluent constamment et s'adaptent aux contre-mesures que les cyberdéfenseurs mettent en place. L'implémentation d'un système de détection et de prévention des fraudes nécessite de distinguer les deux activités suivantes :

- La détection de fraudes qui est l'ensemble des procédures visant à identifier de manière univoque les cas de fraude. Cette détection intervient une fois la fraude réalisée. Le système d'analyse à posteriori doit permettre de classer correctement une fraude à partir des données qui lui sont associées.

- La prévention des fraudes qui est l'ensemble des procédures dont l'objectif est de prévenir la réalisation d'une fraude. Le système d'analyse peut par exemple exploiter un système de règles qui vont déclencher une alarme. Ces règles sont définies par des experts du domaine ou bien automatiquement découvertes/générées par des algorithmes d'IA/ML comme les réseaux de neurones ou les arbres de décision. Il s'agit de découvrir les motifs récurrents et sous-jacents en fonction de la typologie de fraudes.

Par ailleurs, l'implémentation d'un système de détection et de prévention des fraudes, rendue nécessaire par l'immense quantité de transactions, doit minimiser le nombre de faux positifs (c'est-à-dire les transactions légitimes traitées comme frauduleuses), afin de limiter le plus possible le déni de service pour les clients. Les faux négatifs sont les transactions frauduleuses qui ne sont pas détectées, celles -ci doivent également être minimisées.

7.11.3 Les systèmes experts

L'approche de type systèmes experts consiste à implémenter un ensemble de règles de détection développées par des experts du domaine. Elles peuvent être construites avec :

- Les informations des transactions
- L'historique des transactions de la carte bancaire concernée par la transaction
- L'historique de l'ensemble des transactions des cartes bancaires
- L'historique des cartes en opposition
- L'historique des transactions frauduleuses

Les règles sont essentiellement de deux types. Celles qui attribuent une probabilité de fraude à la transaction et celles qui bloquent la transaction immédiatement. Les avantages de cette approche sont :

- La simplicité d'implémentation des alertes
- La compréhension des alertes
- L'explicabilité des alertes

Parmi les inconvénients de ce type d'automate :

- Les experts ne sont pas toujours d'accord entre eux.
- Les variables sur lesquelles s'appuient la prédiction sont en nombre très limité par rapport à la dimensionalité du problème.
- La méthode se base sur des données du passé et les nouvelles menaces ne peuvent pas être identifiées
- Le réglage constant et manuel du système réalisé par les experts pour tenir compte de l'évolution des fraudes. En conséquence ce type de système est difficilement maintenable.

7.11.4 Modèles prédictifs basés sur l'analyse automatique de données

Ce type de modèle repose sur l'apprentissage automatique adaptatif. C'est-à-dire que ce type de système IA/ML adapte ses paramètres dynamiquement lorsque de nouvelles données reflétant de nouveaux comportements ou de nouvelles classes de fraudes sont injectées. L'usage de l'IA/ML présente plusieurs avantages :

- La possibilité d'analyser un ensemble de données présentant un nombre élevé de dimensions ou caractéristiques pouvant suggérer une fraude.

- La possibilité de construire des corrélations entre un nombre élevé de caractéristiques.

- La mise à jour dynamique des modèles pour s'adapter aux changements/évolutions stratégiques des fraudeurs.

- Les techniques d'IA/ML permettent d'analyser une grande volumétrie de données (big data) qui dépassent les capacités humaines.

- Ces algorithmes peuvent être manuellement enrichis par les experts du domaine avec des règles métiers afin d'améliorer leurs performances.

Ces méthodes présentent néanmoins quelques désavantages :

- Les algorithmes sont des boites noires et les classifications ne sont pas interprétables par un humain.

- « L'entrainement » des algorithmes est rendu plus délicat à cause du déséquilibre des ensembles d'apprentissage (les fraudes ne représentant qu'une très faible proportion des transactions).

- La nature non-stationnaire des données implique que les algorithmes sont dans une boucle continuelle d'apprentissage.

- Leur implémentation nécessite une grande volumétrie de données avec pour conséquence la nécessité de disposer d'une très grande capacité de calcul et d'avoir recours au calcul distribué.

7.11.5 La combinaison apprentissage automatique et systèmes experts

Comme nous l'avons évoqué précédemment, la combinaison de plusieurs classificateurs permet presque toujours d'améliorer les prédictions en réduisant les faux négatifs et les faux positifs.

En général, les systèmes experts basés sur des règles du métier permettent de réduire les faux négatifs, bien que cela se fasse au prix d'un accroissement du nombre de faux positifs. L'introduction de méthodes d'apprentissage automatique issues de l'IA/ML permet en retour de réduire les faux positifs.

De leur côté les systèmes experts permettent de compenser les problématiques qui émergent de la non-stationnarité des données et de leur caractère non équilibré, ce qui induit un biais évident dans les prédictions favorisant l'ensemble des transactions régulières.

7.11.6 Les applications actuelles

La détection de la fraude à la carte de paiement à l'aide de l'IA/ML est une réalité depuis plusieurs années déjà, dans le monde bancaire (Grandmontagne, 2018), (Barguisseau, 2019).

L'IA/ML intervient notamment au niveau des réseaux de routage d'autorisation que les grands gestionnaires internationaux de cartes, tels que Visa, MasterCard et American Express exploitent.

Lors d'une transaction par carte bancaire partout dans le monde, une demande d'autorisation est envoyée du point d'acceptation de la carte vers la banque émettrice de la carte. Cette dernière décide alors d'autoriser ou non, la transaction, c'est-à-dire de supporter le préjudice financier en cas de défaillance ou de fraude. Pour répondre aux demandes d'autorisation et réduire leurs risques, les banques émettrices sont équipées de serveurs d'autorisation et de systèmes capables d'évaluer les risques de fraude.

Les exemples de PayPal et Mastercard
Il est intéressant de constater que les systèmes de détection de la fraude de classe industrielle reposent sur convergence de trois technologies :

- Le calcul haute performance (HPC)
- Le Big Data
- L'IA/ML

Par exemple, MasterCard a installé au niveau des centres de routage, un service de « scoring » des transactions. Le résultat de ce « scoring » est ajouté au passage à la demande d'autorisation, et est fourni à la banque émettrice pour l'aider à décider d'accorder l'autorisation on non.

Le système est basé sur l'apprentissage automatique hybride supervisé et non-supervisé en complément des technologies Big Data « traditionnelles » Hadoop et Spark. Le système examine l'emplacement, les habitudes de dépenses et les schémas de voyage des clients avant chaque achat. (Chaque transaction fait l'objet de 1,9 millions de règles distinctes qui l'examinent en quelques millisecondes).

De son côté, PayPal qui gère plus 13 millions de transactions en ligne par jour, utilise un système de détection de la fraude basé sur une solution d'IA/ML qui associe le framework de machine learning H2O. Il s'agit de la solution open source de H2O.ai qui se présente sous la forme d'une plateforme d'apprentissage distribué. La solution repose sur une infrastructure HPC et Big Data qui rassemble quotidiennement plus de 20 TB de données. La solution aurait permis la détection de 700 millions de dollars de transactions frauduleuses (Grandmontagne, 2018).

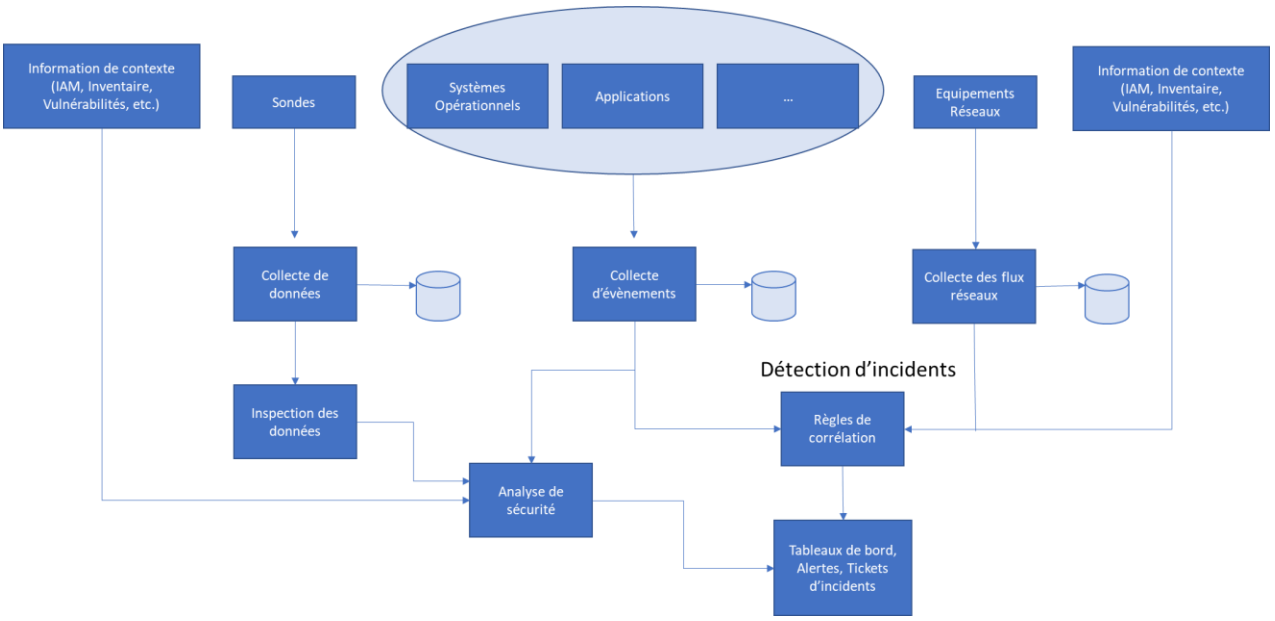
Dans un avenir proche, l'analyse de données pourra aller encore plus loin puisqu'il est déjà également possible de rechercher des informations sur le client via le web et notamment les réseaux sociaux afin d'enrichir les ensembles d'apprentissage. Ainsi, si un client poste une photo de lui en vacances et qu'un paiement à lieu à 1 000 kilomètres de là, le système pourra le détecter.

Ce type d'application pose néanmoins la question du respect de la vie privée des clients versus le complément d'informations apporté à l'IA/ML qui pourrait lui servir à identifier une opération frauduleuse. Il y a là des questions juridiques intéressantes, sachant que les réseaux de paiement sont opérés à l'échelle internationale.

7.12 Les solutions SIEM (Security Information Event Management)

Nombre d'entreprises, qu'elles soient classées opérateurs d'importance vitale, qu'elles opèrent dans des secteurs critiques ou simplement qu'elles soient sensibilisées aux risques cybersécurité sur leurs activités sont pourvues de centres opérationnels de sécurité (SOC pour Security Operation Center en anglais). Cette organisation peut être interne ou externe (service infogéré) à l'entreprise.
Il s'agit d'une organisation humaine en charge de la supervision et de l'administration de la sécurité du système d'information supportée par des solutions de collecte, de corrélation d'évènement et d'intervention à distance.

Le SIEM est l'outil principal d'un SOC qui permet la centralisation et la mise en forme de la collecte des informations. L'objectif est ici de détecter, analyser et remédier aux incidents de cybersécurité. Pour ce faire, les analystes SOC surveillent et analysent l'activité sur les réseaux, les serveurs, les terminaux, les bases de données, les applications, les sites Web et autres systèmes, à la recherche de signaux faibles ou de comportements anormaux qui pourraient être le signe d'un incident ou d'une attaque (De Montes, 2018). Le diagramme ci-dessous illustre un système de collecte simplifié pouvant être mis en place pour un SIEM et exploité par le SOC.



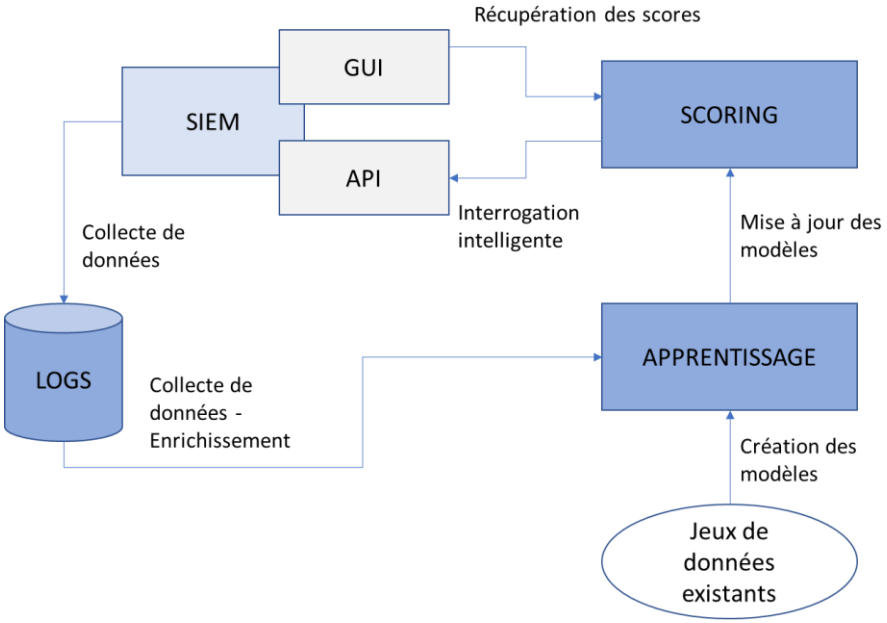
Les experts en charge du SOC reçoivent des alertes à chaque suspicion d'intrusion, mais le nombre de « faux positif » est très élevé. Une enquête menée par FireEye sur des entreprises de grande taille au niveau mondial a montré que 37 % des répondants recevaient plus de 10.000 alertes chaque mois. De ces alertes, 52 % étaient des faux-positifs et 64 % étaient des alertes redondantes (Ibid.).

L'intelligence artificielle permet une approche plus fine des détections d'intrusions. Cette finesse d'analyse permet de réduire le nombre de fausses notifications d'intrusion en filtrant préalablement les alertes et doit donc permettre de focaliser l'attention de l'équipe SOC sur les vraies menaces. Les tâches les plus chronophages et fastidieuses peuvent ainsi être gérées par la machine, permettant aux équipes de se concentrer uniquement sur les actions à plus forte valeur ajoutée.

Ainsi les chercheurs en milieu académique ou dans les entreprises se sont penchés sur la détection de trafic réseau suspect ou anormal. Les universitaires ont montré la faisabilité de solution de détection de comportements suspicieux/anormaux sur des jeux de données limités donc critiquables (Kassem, 2022), (Msika, 2020). En parallèle, les éditeurs (tels qu'IBM ou Splunk) incorporent des modules d'apprentissage automatique dans leurs solutions de supervision des réseaux.

Ces composantes commencent par « observer » le fonctionnement normal (hors attaque) d'un réseau en collectant toutes les données de trafic et métadonnées pour construire un « modèle de normalité ». Après cette phase initiale d'entrainement, les composantes sont en mesure de détecter des séquences d'évènements en déviation par rapport à la « normalité » apprises puis de produire des alertes d'attaques probables (Teboul, 2022).

Le schéma ci-dessous présente une vision haut-niveau de l'intégration d'algorithmes IA/ML au sein d'un SIEM. L'architecture est ainsi décomposée en trois éléments logiques : l'interface utilisateur (GUI), la partie dont le rôle est de calculer un score d'anomalie ou une probabilité de classification (Scoring), et la partie en charge de calculer les modèles.



Ce type d'apprentissage permet dans les cas favorables de détecter des menaces inédites qui ne figurent pas dans les bases historiques d'attaques connues. Cette approche permettrait d'améliorer les solutions efficaces de SIEM (Security Information and Event Management) orientées UEBA (User and Entity Behavior Analytics).

8 ETUDE DES APPLICATIONS IA/ML POUR LA CYBERATTAQUE

Dans ce chapitre, nous présentons quelques approches IA/ML utilisées pour le déploiement de cyberattaques.

8.1 Attaques de type « Spear phishing » augmentées par l'IA/ML

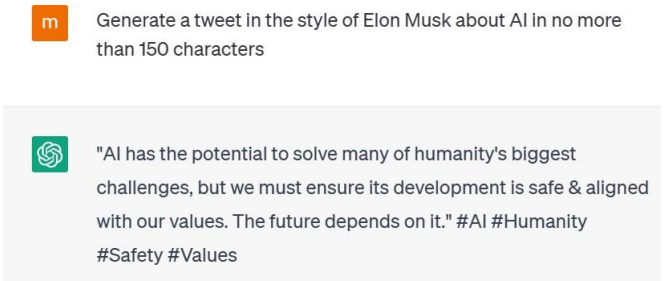
De manière assez simple et à titre d'exemple, il est possible de construire un automate de « Spear phishing ». L'automate utilise l'IA/ML pour imiter les tweets suivis par la cible. Les utilisateurs sont ciblés en fonction de leurs centres d'intérêt déterminés par divers signaux, notamment les comptes que ces derniers suivent, ce qu'ils « retweetent », les contenus sur lesquels ils cliquent, leurs tweets, etc.

Le but est donc de générer un tweet semblant légitime sur un sujet d'intérêt pour la victime, le tweet contiendra un lien de redirection de l'attaque. Le lien est « brouillé » de sorte que la victime ne soit pas en mesure de détecter directement l'URL exacte sur laquelle elle clique.

Ces attaques présentent l'avantage d'avoir un fort taux de réussite pour les attaquants. Pour procéder à ce genre d'attaques, il convient de disposer d'un compte développeur sur Twitter qui permet l'accès aux API du service. L'étape suivante consiste à choisir un utilisateur de twitter dont on imitera les messages afin de pouvoir imiter son style et les sujets qu'il aborde. Ces tweets forment donc un appât probable pour toute personne intéressée par les mêmes sujets que cet utilisateur.

Par exemple, nous pouvons choisir d'imiter les tweets d'une personnalité très suivie telle qu'Elon Musk. Pour ce faire, il faut procéder dans un premier temps à la collecte des tweets publiés par Elon Musk afin de constituer l'ensemble d'apprentissage. Il est par la suite assez simple d'entrainer un modèle d'IA/ML, par exemple de type chaines de Markov[30] ou un réseau de neurones, dans le but de générer des tweets ressemblant au style de Musk. L'introduction du lien d'attaque est sans difficulté et pour masquer la véritable nature du lien, l'URL sera cachée sous un mot ou un texte.

En réalité, dans le cas de personnalités aussi célèbre qu'Elon Musk, l'imitation est même plus simple puisqu'il est possible de générer des tweets dans un style identique au sien en demandant à l'IA générative ChatGPT.



8.2 Camouflage et encapsulation de malware

En 2017, à la conférence BlackHat[31], des chercheurs ont présenté AVPASS, un outil conçu pour déduire les règles de détection d'un moteur d'antivirus. Par la suite, l'outil utilise ses inférences pour rendre

[30] Une chaîne de Markov est un modèle stochastique de description d'une séquence d'événements possibles dans lesquelles la probabilité de chaque événement dépend de seulement de l'état atteint dans le cas précédent.
[31] Les Conférences Black Hat (ou Black Hat Briefings) sont un événement unique qui rassemble officiellement des experts des agences gouvernementales américaines et des industries, américaines ou non, avec les hackers les plus respectés de l'« underground » (Cf. Wikipédia).

indétectable un malware par l'anti-virus ciblé. Cet outil de recherche a ainsi démontré qu'il était possible pour un système d'IA/ML de masquer les malwares pour contourner un antivirus spécifiquement ciblé (Berthier, 2022).

Lors de la conférence Blackhat de 2019, des chercheurs IBM ont démontré qu'il était possible de camoufler un malware en embarquant un réseau de neurones dans son code à l'aide d'un système baptisé DeepLocker. Ce dernier permettait au malware de vérifier l'environnement sur lequel il s'exécutait de manière implicite. En effet, les malwares d'un certain niveau de sophistication effectuent la plupart du temps des vérifications spécifiques pour connaître leur environnement d'exécution. Il s'agit d'une étape de reconnaissance en quelque sorte. Ce type de vérification permet au malware de ne s'exécuter que sur les environnements cibles pour lesquels il a été conçu. Les antimalwares conçus pour effectuer une analyse dynamique (lors de l'exécution en mémoire) peuvent lors de cette étape détecter un comportement suspect. L'adjonction d'un réseau de neurones au malware par DeepLocker permet à ce dernier de pouvoir détecter l'environnement de manière indirecte ou implicite et ainsi d'être « furtif ». Ce travail a permis de démontrer que l'IA permettait d'amplifier les capacités de camouflage des malwares (Berthier, 2022).

8.3 Compromission du mot de passe

Les outils traditionnels de découverte de mot de passe (tels que HashCat et John the Ripper) fonctionnent en comparant de nombreuses variantes du hachage du mot de passe à partir d'un dictionnaire. Depuis quelques années déjà, l'IA s'invite avec succès dans ce domaine. Ainsi, l'outil PassGAN qui existe depuis 2019, s'appuie sur une technique de réseaux de neurones antagonistes pour apprendre la distribution statistique des mots de passe à partir des fuites et générer des propositions de mots de passe probables. En termes de performances, ce système a permis de découvrir entre 51 % à 73 % de mots de passe en plus par rapport à la solution classique HashCat.

Le code de PassGAN est disponible sur l'entrepôt Github au niveau de l'URL suivante : https://github.com/brannondorsey/PassGAN.

8.4 Compromission des systèmes de sécurité CAPTCHA

L'IA peut être utilisée pour tromper les systèmes de sécurité CAPTCHA, exploités pour protéger certains services. Par exemple, la solution XEvil peut casser les systèmes de reconnaissance humaine à l'aide de CAPTCHA sur les pages Yandex et est louée sur le dark web pour environ 54 dollars à la semaine ou 136 dollars au mois (Berthier, 2022).

8.5 Attaques par imitation de la voix

D'après la revue Forbes, en 2020 des hameçonneurs ont mené une attaque contre la succursale de Hong Kong d'une banque Emiratie en clonant la voix d'un de leur client. Un directeur de banque à Hong Kong a reçu un appel d'un homme dont il a reconnu la voix - un directeur d'une entreprise (client de la banque) avec qui il avait déjà parlé. Le directeur avait une bonne nouvelle : son entreprise était sur le point de faire une acquisition, il avait donc besoin que la banque autorise des transferts à hauteur de 35 millions de dollars. Un avocat du nom de Martin Zelner avait été embauché pour coordonner les procédures. En outre, le directeur de la banque pouvait voir dans sa boîte de réception les courriels du directeur et de Zelner, confirmant les montants à transférer. Le directeur de la banque, estimant que tout semblait légitime, a commencé à effectuer les virements.

En fait, il s'agissait d'une escroquerie, la banque avait été dupée dans le cadre d'une escroquerie élaborée, dans laquelle des fraudeurs avaient utilisé la technologie d'imitation automatique de la voix (Audio Deep fake) permettant de se faire passer pour qui l'on souhaite (Brewster, 2021).

Les technologies de type « deep fake » audio et/ou vidéo sont basées sur les réseaux de neurones (notamment les réseaux de neurones convolutifs). Il est possible d'entrainer ces systèmes à partir d'enregistrement audio /video. À partir de ces enregistrements, il est assez simple d'imiter la voix d'une personne et de lui faire prononcer le message que l'on souhaite.

À partir de là, les applications malveillantes sont innombrables, comme dans notre exemple il peut s'agir d'une simple escroquerie, mais également d'actions de désinformation ou d'attaques à la réputation de personnalités.

8.6 Tests de pénétration - système augmenté par IA/ML

Depuis longtemps, des outils d'automatisation sont utilisés par les cyberattaquants (ou les cyberdéfenseurs). Parmi ces outils, MetaSploit est une solution incontournable qui permet soit de détecter des vulnérabilités dans le cadre de campagnes légitimes de tests de pénétration soit de mener des cyberattaques contre une cible désignée. Ce logiciel est aujourd'hui la propriété de la société de cybersécurité Rapid7 dont une édition Open Source est toujours disponible grâce aux larges contributions et développements de la communauté de sécurité. La version Open Source de MetaSpoil est déjà installée et configurée sur la version Kali Linux.

Il existe de nombreux modules que l'on peut ajouter à la solution. Ces modules sont téléchargeables depuis le site www.rapid7.com. Ils sont contenus dans une base interne permettant d'interroger à la fois les bases Common Vulnerabilities and Exposures (CVE), ou Open Sourced Vulnerability Database (OSVDB), le site Bugtraq ou encore le Microsoft Security Bulletin. MetaSpoil ne fait aucunement appel aux techniques de IA/ML.

Le Framework MetaSpoil contient ainsi une grande base de données de modules d'exploitation, de payloads et de scripts de post-exploitation qui peuvent être utilisés pour automatiser les tests de sécurité. Il permet notamment de :

- Scanner et collecter des informations sur une machine cible.
- Détecter et exploiter les vulnérabilités.
- Augmenter les privilèges d'un système d'exploitation.
- Installer une porte dérobée pour maintenir un accès persistant.
- Utiliser la technique de "Fuzzing" pour tester la robustesse d'un logiciel.
- Utiliser des outils d'évasion pour contourner les logiciels de sécurité.
- Utiliser des charges actives ou « payloads » pour exécuter des commandes à distance sur les systèmes compromis.
- Utiliser des outils de pivot pour propager l'accès à d'autres systèmes connectés.
- Effacer les traces et les journaux pour dissimuler les activités malveillantes.

Nous évoquons ici, MetaSploit car ce logiciel a fait l'objet d'améliorations basées sur l'usage de l'IA/ML et plus précisément par les techniques d'apprentissage par renforcement.

Nous rappelons ici, que les algorithmes d'apprentissage automatique par renforcement sont une méthode d'apprentissage qui produit des actions et découvre les erreurs dans son environnement. Cette méthode permet à l'algorithme de « découvrir » automatiquement le comportement dans un contexte spécifique afin de maximiser ses performances. Une rétroaction de récompense est nécessaire pour que l'agent « comprenne » quelle action est la meilleure réponse, et ceci est connu comme un signal de renforcement.

DeepExploit est un outil de test d'intrusion automatique utilisant l'apprentissage par renforcement et qui s'appuie sur MetaSpoil. Le logiciel a été écrit par l'ingénieur en cybersécurité et IA, Isao Takaesu. Il est librement disponible sur GITHUB au niveau de l'URL suivante :
https://github.com/13o-bbr-bbq/machine_learning_security/tree/master/DeepExploit

La dernière mise à jour remonte au 30/07/2022. Elle permet de mener les actions suivantes :

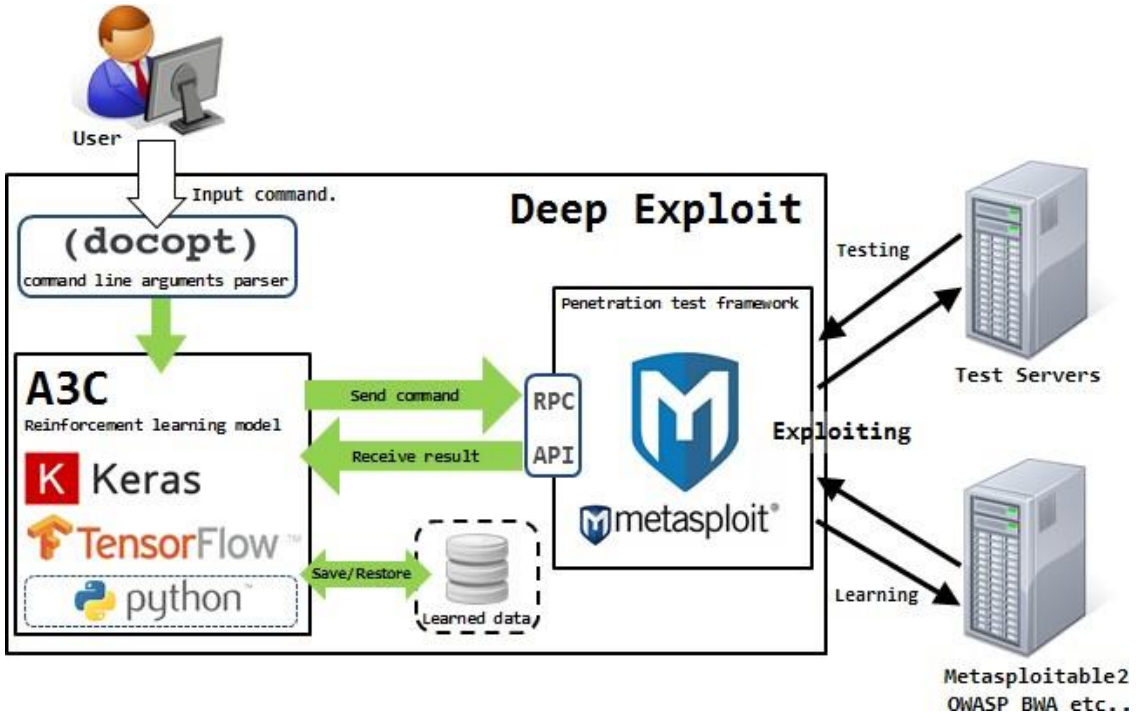
- La collecte de renseignements.
- La modélisation des menaces.
- L'analyse de vulnérabilité.
- L'exploitation.
- La génération de rapports.

La solution fonctionne selon deux modes d'exploitation :

- Le mode « intelligence » : DeepExploit identifie l'état de tous les ports ouverts sur le serveur cible et exécute l'exploit de manière précise en fonction de l'expérience passée (résultats d'apprentissage)

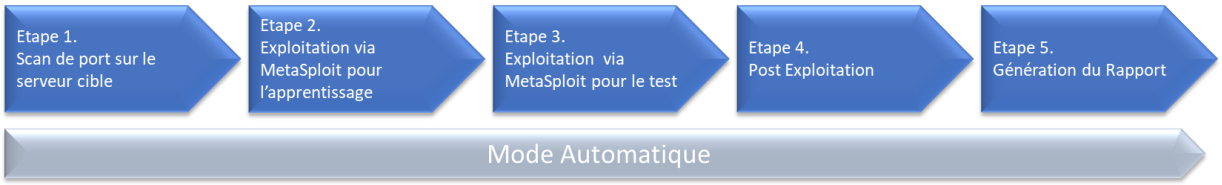
- Le mode « brute force » : DeepExploit exécute des « exploits » en utilisant toutes les combinaisons de « module d'exploitation », « cible » et « charge utile » correspondant au nom de produit et au numéro de port indiqués par l'utilisateur.

D'après la documentation d'architecture générale logicielle est comme suit :



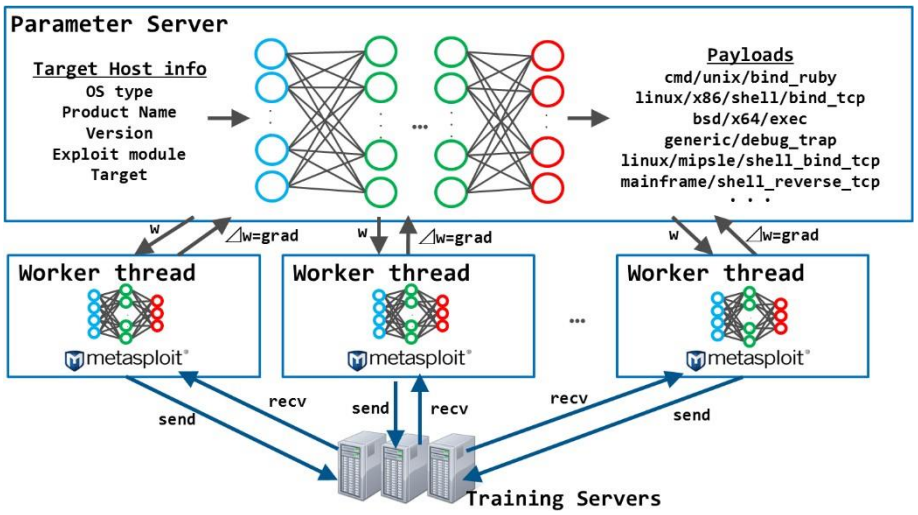
Le mode Intelligence

Nous présentons ci-dessous le « workflow » associé au mode « intelligence ».



Dans une première étape, DeepExploit exécute l'analyse des ports à l'aide de Nmap pour collecter les informations du serveur cible. Ensuite, DeepExploit exécute deux commandes de Metasploit (hosts et services) via l'API RPC.

DeepExploit apprend à exploiter les failles de sécurité par lui-même à l'aide d'un modèle d'apprentissage automatique appelé A3C. Le logiciel utilise des serveurs vulnérables tels que metasploitable2, owaspbwa pour l'apprentissage.



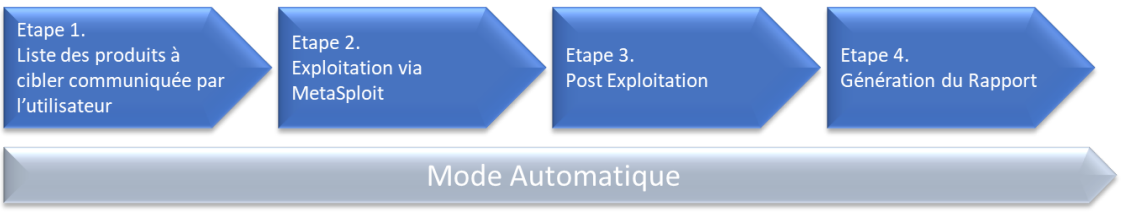
L'A3C se compose de plusieurs réseaux de neurones. Ce modèle reçoit les informations du serveur d'apprentissage telles que le type de système d'exploitation, le nom du produit, la version du produit, etc. en tant qu'entrées du réseau neuronal, et génère la charge utile en fonction des informations d'entrée. Le fait est que l'exploitation est réussie lorsque ce modèle produit une charge utile optimale en fonction des informations d'entrée.

En phase d'apprentissage, ce modèle exécute plus de 10 000 exploits sur les serveurs d'apprentissage via MetaSploit tout en modifiant la combinaison des informations d'entrée. Ce modèle met à jour les poids du réseau de neurones en fonction des résultats d'exploitation (récompenses), ce qui va progressivement optimiser le réseau de neurones.

Afin de raccourcir le temps d'apprentissage, ce dernier est exécuté par multithreading. L'étape 3 consiste à exploiter la/les failles en utilisant MetaSploit pour les tests. L'étape 4 consiste en la publication de l'exploit. Et l'étape 5 est la génération d'un rapport au format html qui résume les vulnérabilités.

Le mode « brut force »

Nous présentons ci-dessous le « workflow » du mode « brute force » de DeepExploit :



Il s'agit d'un « workflow » complètement automatique simple qui exploite toutes les combinaisons possibles liées aux vulnérabilités de la liste des produits/logiciels.

L'avenir des solutions telles que DeepExploit

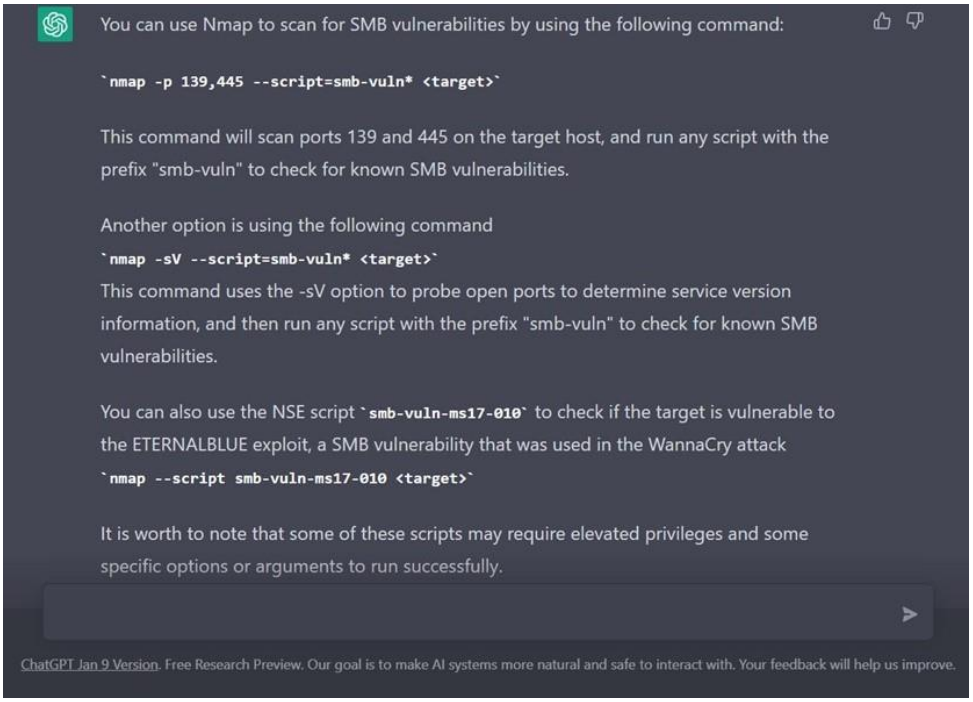
DeepExploit est un logiciel open source dont la principale limitation de notre point de vue est qu'il n'y a qu'un seul contributeur sur le projet (Isao Takaesu). Nous n'avons pas vu de communauté associée au projet. Cela étant dit, le point fondamental est que cette solution démontre la faculté de l'IA/ML à améliorer les capacités des outils de détection des vulnérabilités qui peuvent être appliqués soit dans le cadre de l'automatisation de la cyberdéfense des organisations, soit pour mener des attaques complètement automatisées. Nous constatons une tendance au développement de solutions de détection automatique de vulnérabilités logicielles basées sur le deep learning (Recurrent Neural Networks). À ce sujet, nous pouvons citer les travaux de recherche de Li et ces collaborateurs (Li, et al., 2018). La combinaison de l'IA/ML et des solutions de détection/exploitation de vulnérabilités devrait dans les prochaines années se généraliser pour le meilleur et pour le pire.

8.7 Usage hybride d'un système de type LLM en cybersécurité

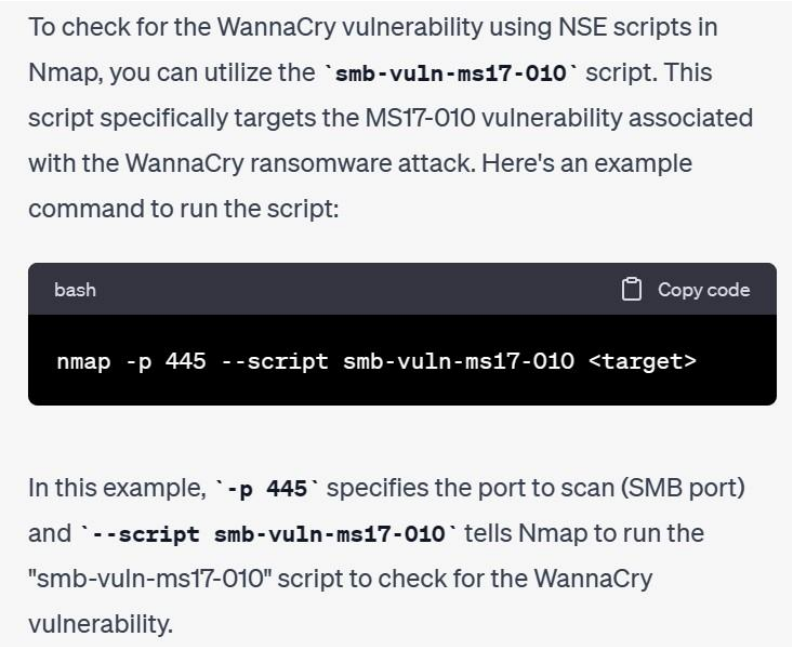
ChatGPT (Generative Pretrained Transformer) est une IA généraliste de type LLM (Large Langage Model). Mais ses applications dans le domaine de la cybersécurité vont de l'écriture de code malveillant (infostealer ou ransomware) à la génération d'un flux de campagnes de phishing. Ceci bien que les conditions d'utilisation d'OpenAI interdisent spécifiquement l'utilisation du programme ChatGPT pour créer des logiciels malveillants de toute sorte, définis par la société comme *« un contenu qui tente de générer des rançongiciels, des enregistreurs de frappe, des virus ou d'autres logiciels destinés à infliger un certain niveau de préjudice »*.

Le Dr Suleyman Ozarslan, a ainsi déclaré qu'il était possible d'utiliser ChatGPT pour une variété de tâches de cybersécurité offensives et défensives. Celles-ci comprennent la rédaction d'un courriel de phishing dans un anglais parfait, la génération de règles de détection pour identifier les anomalies de cybersécurité et la création d'un code d'évasion qui pourrait contourner les règles de détection (Technologies, 2023).

Si l'on se focalise sur les attaques, il est possible de demander à l'automate comment rechercher les vulnérabilités SMB (Server Message Block) avec Nmap. Nous présentons ci-dessous la réponse particulièrement détaillée qui explique comment mener cette attaque et fait des suggestions pour déterminer si la cible est vulnérable à l'exploit EternalBlue (le ransomware WannaCry utilise cette faille de sécurité).



Signalons, que nous avons testé les affirmations sur Nmap, et même si nous n'obtenons pas exactement les mêmes résultats, nous avons globalement obtenu des résultats comparables (cf. copie de la réponse à notre interrogation : « exemple of nse script wannacry check »).



Ce simple exemple suffit à montrer que pour les équipes de cybersécurité Red Team et les cybercriminels, les implications sont énormes. Même si ChatGPT n'a pas la capacité de générer des attaques inconnues, il est clair que ce type d'outil abaisse considérablement le niveau technique requis pour mener à bien une attaque et c'est finalement cela le plus dangereux. Il est d'ores et déjà envisageable d'entrainer un modèle de type LLM (Large Language Model) sur des données de cybersécurité (stratégie d'attaques, stratégie de défense) afin d'en faire une IA spécialisée en cyberattaque et/ou cyberdéfense. L'apprentissage de ce type de modèle est simplifié par les possibilités d'apprentissage par transfert (c'est-à-dire que l'on repart d'une IA pré-entrainée) et

surtout par la possibilité d'entrainer ce modèle avec des spécialistes humains en utilisant l'apprentissage par renforcement (système de récompenses/pénalités) qui semblent particulièrement efficace.

Du point de vue de la cyberdéfense, il est possible d'obtenir des réponses de haut niveau pour savoir par exemple (Ibid.) :

- Rédiger une requête de recherche pour identifier les modifications dans le registre Windows (par exemple, une requête ELK pour détecter les modifications du registre).

- Rédiger une expression régulière pour filtrer les adresses IP dans le SIEM Splunk.

- Identifier les problèmes dans un code PHP avec une vulnérabilité connue. Dans les cas favorable, ChatGPT peut identifier la faille de sécurité, mais également fournir le code pour la corriger.

Alors que les cyberattaques augmentent en volume et en complexité, l'IA/ML peut aider les équipes de sécurité à atténuer les menaces. En organisant des renseignements sur les menaces à partir de sources de recherche, de blogs et d'articles d'actualité. Les technologies d'IA/ML telles que l'apprentissage automatique et le traitement du langage naturel (NLP) fournissent des informations exploitables qui réduisent l'encombrement avec pour effet de réduire considérablement les temps de réponse aux incidents de sécurité.

Les applications au niveau du SOC sont évidentes et les impacts sur les salariés également. Il serait ainsi possible pour un employeur de ne recruter que des analystes SOC avec une expérience moindre et de les « augmenter » avec une IA générative de type LLM couplé au SIEM. Dans ce cas l'IA/ML assisterait les analystes dans la qualification et le traitement des alertes en fournissant une meilleure compréhension du contexte et du type d'attaque.

9 CYBERSECURITE DES SYSTEMES IA/ML

9.1 Une brève taxonomie des attaques sur le système d'information

9.1.1 Les attaques et leurs motivations

Une attaque est l'exploitation d'une faille permettant un accès non autorisé à un système d'exploitation, une application dans le but de causer des dommages. Afin de prévenir ces attaques malveillantes, il est important de connaitre les typologies d'attaque. Les motivations des attaquants peuvent se résumer comme suit :

- Obtenir un accès au système distant
- Récupérer les données stratégiques des acteurs comme les secrets industriels ou propriétés intellectuelles
- Récupérer les données de tel ou tel individu
- Obtenir les données bancaires
- Cartographier l'organisation
- Perturber le fonctionnement d'un service
- Utiliser le système de l'utilisateur, ou accessoirement l'utiliser comme rebond pour des attaques
- Utiliser les ressources des systèmes de l'utilisateur pour profiter de la bande passante

9.1.2 Les composantes principales du SI – cible des attaques

Au niveau le plus général, le système d'information est composé principalement de 3 composantes, à savoir le matériel, les logiciels et l'humain.

- **Le matériel/infrastructure** : il s'agit de l'ensemble de l'environnement lié aux éléments matériels qui participent au système informatique (ordinateurs, réseaux, routeurs, électricité, clés USB, téléphones, tablettes).

- **Les logiciels et applications** : ils englobent chacun des éléments qui peuvent être installés et ou programmés sur les divers équipements.

- **Les collaborateurs** : la composante humaine est primordiale à tout système informatique. Elle regroupe tous les utilisateurs qui interviennent sur les outils via les logiciels.

Les attaques peuvent intervenir sur chacune de ces 3 composantes, pour peu qu'il existe un point faible exploitable.

9.1.3 Catégorisation des principaux risques

Nous pouvons donc catégoriser les risques qui pèsent sur le système d'information, comme suit :

1 - Accès physique :
Quand l'attaquant accède physiquement au matériel :

- L'électricité peut être coupée dans le bâtiment ou en amont.
- Les ordinateurs et/ou les serveurs peuvent être éteints.
- L'infrastructure peut être vandalisée.
- Les ordinateurs et/ou les disques durs peuvent être volés.

2 - Interception des communications :

- Trafic réseau mis sur écoute.
- Vol de session (Session hijacking).
- Usurpation d'identité.
- Détournement ou altération de messages.

3 - Dénis de service :

Il s'agit de rendre indisponible un service, d'empêcher les utilisateurs légitimes d'un service de l'utiliser :

- Via l'exploitation de faiblesses des protocoles TCP/IP.
- Via l'exploitation de vulnérabilités des logiciels serveurs.

4 - Intrusions :

Une intrusion représente l'idée de pénétration au sein d'un système d'information.

- Par un balayage préalable de ports (Port scanning), c'est la recherche de ports ouverts sur un serveur afin d'en exploiter les vulnérabilités.
- Élévation de privilèges, cela consiste à exploiter une vulnérabilité applicative en envoyant un code non prévu par le concepteur et conduisant parfois à un accès au système avec les droits de l'application. Les attaques par débordement de tampon (Buffer overflow) utilisent ce principe.
- Logiciels malveillants, ils englobent les virus, vers et chevaux de Troie, ainsi que d'autres menaces.

5 - Ingénierie sociale :

Communément décrit comme un « piratage psychologique », ces attaques exploitent les faiblesses psychologiques et sociales des individus ou plus largement les faiblesses des organisations, pour obtenir des informations frauduleusement (un mot de passe, la divulgation d'informations confidentielles, l'ouverture de pièce jointe non sollicitée). L'individu ne peut être protégé par un dispositif, la mise en place d'une veille sécuritaire et de formations spécifiques dans l'organisation peut éviter aux collaborateurs de tomber dans ce piège.

9.1.4 Les exemples les plus fréquentes d'attaques

Nous présentons ci-dessous un panorama succinct des formes les plus fréquentes d'attaques cyber (Picciau, 2022) :

- Le phishing et le spear-phishing
  - Il s'agit d'envoyer un courriel qui semble provenir d'une source fiable afin d'obtenir des informations sensibles ou d'inciter le destinataire à entreprendre une action déterminée. Le courriel en question peut contenir une pièce jointe ou un lien cliquable qui servira à recueillir les informations ou à charger un logiciel malveillant sur l'ordinateur utilisé.

- Les attaques par logiciels malveillants
  - Le ransomware (rançongiciel) bloque l'accès aux équipements ou aux données et ne rétablit l'accès qu'en échange d'une somme d'argent. Le risque est de voir ses informations supprimées ou publiées si la rançon n'est pas payée.
  - Le Cheval de Troie est un programme malveillant dissimulé derrière un programme utilisé par l'entreprise. Généralement, il sert à ouvrir des portes d'accès aux attaquants, qui leur permettront – par exemple – d'écouter le trafic.

- Les macro-virus infectent la plupart du temps des applications très classiques, comme Microsoft Word ou Excel. Dans ce cas de figure, un code est dissimulé dans le logiciel qui, en s'exécutant, télécharge un ransomware ou un Cheval de Troie, qui vont mettre en danger d'autres points du système informatique.

- Le déni de service (DDoS)
  - Le (Distributed Denial of Service) est une attaque qui a pour conséquence de rendre silencieuse une machine en la submergeant de trafic inutile. Plusieurs machines peuvent être à l'origine de cette attaque visant à « anéantir » des serveurs ou des sous réseaux entiers. Les attaques de type déni de service peuvent par exemple toucher le service de courrier électronique, d'accès à Internet, de ressources partagées (pages Web). Le déni de service est un type d'attaque qui peut avoir un coût élevé pour une entreprise (exemple : un site web de commerce), en bloquant le cours normal des transactions. Ces attaques sont difficiles à contrer ou à éviter.

- L'attaque par Drive by Download
  - Pour propager un logiciel malveillant, les pirates peuvent repérer les sites web non sécurisés et intégrer un script malveillant dans le code HTTP ou PHP d'une page. Ce script permettra :
    - De rediriger tout utilisateur du site vers un autre portail contrôlé par les pirates
    - D'installer des logiciels malveillants directement sur les équipements de l'utilisateur.

- L'attaque de l'homme au milieu ou MitM
  - Un attaquant ou un serveur se positionne entre deux entités et en intercepte les communications. L'attaque prend alors la forme d'un détournement de session via une usurpation d'IP par exemple.
  - Ce type d'attaque utilise parfois l'IP spoofing. La première utilisation (Blind Spoofing) est de falsifier la source d'une attaque, pour éviter de localiser sa provenance. La seconde utilisation (IP Source Rating) est de profiter d'une relation de confiance entre deux machines pour prendre la main sur l'une d'entre elle.

- Le piratage de compte
  - Il s'agit d'acquérir le mot de passe associé au compte. Le piratage peut viser le compte d'administrateur du site de l'entreprise, le compte bancaire de la société, le compte de messagerie ou les accès aux réseaux sociaux.

- La fraude au président ou Faux Ordre de Virement (FOVI)
  - Dans le cas d'une fraude au président, un individu se fait passer pour le président d'une société-mère afin d'obtenir un virement bancaire auprès d'une société cible, appartenant au même groupe.

9.2 Principes de confiance de l'IA

L'acceptation des système ML/IA par les utilisateurs nécessite de générer de la confiance dans les décisions du système en évitant au maximum les biais et discriminations. Au minimum, de tels systèmes devraient suivre les principes suivants (Ijlal, 2022) :

- **L'intégrité** : il s'agit de s'assurer que le système ne puisse pas être falsifié. Les données collectées pour la phase d'apprentissage ne doivent être utilisées qu'à cet effet.

- **L'explicabilité** : il convient d'éviter dans la mesure du possible l'effet « boite noire » dans le processus de décision. La décision devrait être le plus transparente possible et elle est possiblement argumentée.

- **L'équité** : la décision est équitable et éthique, non fondée sur des biais ou préjugés (par exemple discrimination à l'âge, au genre, à l'ethnicité, etc.)

- **La résilience** : le système doit être sécurisé et être robuste contre les attaques dont il est susceptible de faire l'objet.

Ainsi dans les principes de confiance de l'IA apparaissent à minima deux principes étroitement liés à la cybersécurité du système à savoir l'intégrité et la résilience. Nous discuterons plus loin de l'implémentation de mesures pour essayer de garantir ces principes.

9.3 Vers la mise en place d'un référentiel de sécurité IA/ML

L'identification des menaces passe par la mise en place d'un référentiel de cybersécurité pour les systèmes d'information utilisant l'IA/ML en traitant les points suivants (Ijlal, 2022) :

- La revue de l'environnement juridique dans lequel le système opère
- La mise en place d'une ligne de base sécurité
- Le maintien d'un inventaire à jour de tous les systèmes en production
- L'évaluation périodique des risques techniques
- La création d'un programme de sensibilisation sur les risques IA/ML
- La mise à jour des contrats d'assurance pour intégrer les risques IA/ML (le cas échéant)

9.4 Taxonomie des attaques sur les systèmes IA/ML

9.4.1 Des systèmes particulièrement sensibles aux attaques

L'usage de l'IA/ML en cybersécurité apparaît comme l'avenir pour diverses solutions afin d'y apporter des améliorations substantielles, notamment à travers la possibilité de détecter et d'arrêter des attaques prenant sans cesse de nouvelles formes.

Cela étant dit, l'IA/ML peut faire l'objet de nouvelles attaques qui doivent être prises en compte et traitées spécifiquement. Les cyberattaques contre un système d'IA/ML peuvent prendre essentiellement deux formes :

- Le système IA/ML, comme tout composant du système d'information peut faire l'objet d'une attaque de l'infrastructure sous-jacente. Ce type d'attaque peut exploiter - par exemple - des dysfonctionnements au niveau d'un contrôle d'accès inexistant ou défaillant, des mises à jour de sécurité qui ne sont pas effectuées ou réalisées de manière partielle, des problèmes de filtrages de flux, une acquisition frauduleuse des accès à travers les techniques de « social engineering », etc. Toutes les attaques classiques sont ici envisageables.

- Le cyberattaquant peut s'attaquer directement au système d'IA/ML en manipulant les caractéristiques de la solution. Plusieurs modèles commerciaux ont fait l'objet de telles attaques. Ce type d'attaques ne fera que se développer selon le Gartner et ce risque est d'autant plus sensible que beaucoup de compagnies adoptant ces technologies ne sont pas systématiquement au courant de ces menaces (Ijlal, 2022).

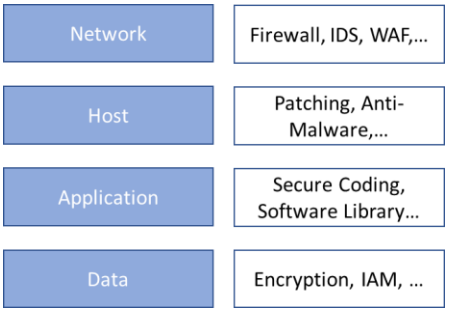
Comme nous l'avons vu, si l'IA/ML est un outil intéressant en cyberdéfense, son usage peut être détourné par les cybercriminels. L'usage de l'IA/ML (en conjonction avec l'automatisation des cyberattaques) permettra d'atteindre des niveaux d'industrialisation des attaques encore inédits à ce jour.

L'usage généralisé de l'IA/ML devrait selon toute vraisemblance permettre aux cyberattaquants de réduire de manière substantielle les coûts d'une cyberattaque ciblée en réduisant le travail humain nécessaire, le niveau d'expertise requis et en simplifiant la phase de collecte d'informations (Ijlal T. page 46, 2022).

9.4.2 cybersécurité de l'IA/ML versus cybersécurité traditionnelle

Les algorithmes d'IA/ML reposent sur le modèle sous-jacent utilisé (le type d'algorithme d'apprentissage) et les données utilisées pour l'apprentissage et la prise de décision. Dans ce contexte, le risque sur la falsification des données pour altérer le processus de décision revêt un caractère critique. Les attaques sur les systèmes IA/ML peuvent se produire à chaque étape du cycle de vie de la mise au point jusqu'à l'exploitation de ces derniers. Nous passerons par la suite en revue les attaques spécifiques sur de tels systèmes (Ijlal, 2022).

Le diagramme ci-contre montre un exemple typique de déploiement d'un modèle de sécurité en profondeur d'une application et des données. Ce modèle doit être naturellement repris pour tous les systèmes IA/ML, mais il doit être complété pour tenir compte des attaques spécifiques auxquelles ils sont soumis. Il est ainsi recommandé d'adopter des contre-mesures pour les attaques « adverses » (adversial attacks). Ces attaques reposent sur le plus souvent sur les réseaux GAN.



Les réseaux GAN (Generative Adversial Networks) ont été théorisés en 2014 par les travaux de Ian Goodfellow et Yoshua (Goodfellow, et al., 2014). Le système consiste essentiellement à mettre en compétition deux réseaux de neurones afin d'améliorer simultanément les performances des deux réseaux. Le résultat final est l'atteinte d'un équilibre dans lequel les performances des deux classificateurs opposés ne peuvent plus être améliorées.

Un système GAN se compose d'un réseau génératif dont la tâche est de créer des données synthétiques qui resteront indétectables. Le second réseau est dit discriminateur, son rôle est de classer correctement les données en fonction du problème traité. Bref, le but du réseau génératif est de « tromper » le réseau « discriminateur ». Il s'agit finalement d'un « jeu à somme nulle » dans lequel un équilibre (Nash equilibrium) entre les deux réseaux sera atteint.

Il est assez évident par conséquent que les réseaux de neurones peuvent être induits en erreur par des données synthétiques via l'usage de réseaux GAN.

Les moyens de défense reposent essentiellement sur trois types d'approche :

- La détection statistique des exemples « adverses » en exploitant les propriétés statistiques des données des ensembles d'apprentissage : l'hypothèse est que les données synthétiques présentent des caractéristiques statistiques qui s'écartent des données réelles (distributions différentes par exemple). Cette hypothèse n'est pas toujours vérifiée.

- L'inclusion d'exemples « adverses » dans la phase d'apprentissage de l'algorithme afin de renforcer la robustesse des prédictions du classificateur : cette méthode a néanmoins un coût

puisqu'elle passe par la complexification du classificateur (augmentation du nombre de paramètres à optimiser).

- Masquer les informations relatives à la méthode d'optimisation du classificateur durant la phase d'apprentissage.

9.4.3 Attaque par empoisonnement de données

Cette attaque vise à altérer les données d'apprentissage. En contaminant les sources de données, l'attaquant peut créer « une porte dérobée », puisque le modèle est construit sur un ensemble de données biaisées et trafiquées. L'empoisonnement de données altère directement le processus de décision.

Ce type d'attaque impose donc la mise en place de processus de contrôle des données et de leur intégrité. Ces contrôles d'intégrité doivent être mis en place afin d'avoir la capacité d'inverser les dommages et de repartir d'une situation saine. Ceci suppose bien évidemment d'avoir implémenté une politique de sauvegarde/restauration pertinente.

Plus généralement, les processus pour assurer la disponibilité, la confidentialité, l'intégrité et la traçabilité des données doivent être mis en place et régulièrement audités.

9.4.4 Attaque par empoisonnement du modèle

Il s'agit d'attaquer un modèle pré-entrainé/préparé pour le compromettre et mettre en place « une porte dérobée » afin d'être en mesure de contourner le processus de décision. En effet, il faut savoir que la plupart des entreprises ne construisent pas leurs modèles à partir d'une feuille blanche, mais utilisent un système pré-entrainé (par exemple ResNet de Microsoft) ou plus généralement externalisent le processus d'apprentissage. En effet, l'entrainement d'un système selon les standards industriels requiert énormément de temps de calcul, parfois plusieurs mois. Ces modèles sont parfois disponibles en « open source ». Il s'agit en quelque sorte d'une attaque de la chaîne d'approvisionnement logicielle dans laquelle un attaquant peut empoisonner les sources pour de nombreux utilisateurs.

Les systèmes « open source » constituent certes une incroyable opportunité pour accélérer le développement d'une entreprise. Cependant des précautions s'imposent, il est recommandé de ne pas réutiliser des modèles directement sans vérification préalable du comportement des systèmes. L'utilisation de modèles commerciaux peuvent s'avérer intéressants à condition de disposer des rapports décrivant les processus de gestion des risques internes pour les systèmes IA/ML.

9.4.5 Attaque par extraction de données

L'attaquant peut en requêtant le système IA/ML inférer les données qui ont été utilisées pour la phase d'apprentissage du système. Cette compréhension de l'ensemble des données d'apprentissage peut conduire à la compromission du système, c'est-à-dire qu'il peut permettre de trouver des modèles de contournement. Ce type d'attaque également appelé attaque par « inférence d'appartenance » ne requiert pas un accès particulier aux fonctionnalités du modèle et peut être mené simplement par observation du modèle de réponses en fonction d'entrées spécifiques.

En tout état de cause, il convient de s'assurer que l'interrogation du système et l'accès aux sources de données sont soumises à autorisation pour y accéder et intégrer des limites sur le nombre de requêtes possibles. Si une source de données d'une tierce partie est utilisée, il faut également s'assurer de son intégrité. Il convient de réduire le niveau de « verbosité » des réponses du système en n'exposant pas d'informations susceptibles d'être utilisées contre le système. Le durcissement des API exposées est

impératif également et ces dernières ne doivent être accessibles qu'à des utilisateurs spécifiques (dans la mesure du possible). Les scores ou probabilités associés aux réponses ne doivent pas être disponibles pour tout le monde.

9.4.6 Attaque par extraction de modèle

L'attaquant peut créer une copie hors ligne du modèle en requêtant et en observant les réponses du système. Disposer d'un jeu d'entrée/sortie significatif et de quelques informations permet tout à fait de reconstruire le système. Ce type d'attaque est simplifié par le fait que la plupart des modèles exposent leur API publiquement et ne gèrent pas correctement leurs sorties car trop d'informations sont communiquées à l'utilisateur. Cette technique permet donc à l'attaquant d'analyser en profondeur la copie hors ligne et de comprendre comment contourner le modèle de production.

Comme précédemment, il convient de réduire le niveau de verbosité du système. La mise en place d'indicateurs de sécurité pour suivre par exemple les pics d'usage anormal du système peuvent permettre la détection d'une attaque en cours.

9.4.7 Attaque par évasion de modèle

Cette attaque vise à tromper le système d'IA/ML en « forgeant » une entrée qui trompera à coup sûr l'algorithme. Ce type d'attaque est réalisé en observant le modèle en action et en comprenant comment contourner le processus de décision. Par exemple, un attaquant peut essayer de contourner un système antimalware basé sur l'IA en trouvant et intégrant des séquences spécifiques à ses malwares qui feront que ces derniers ne seront pas détectés par le système. Plus simplement, ce type de système ne détecte jamais 100 % des malwares existants, il s'agit alors de trouver les quelques cas non détectés par l'algorithme et d'en faire usage.

Dans la phase d'apprentissage du modèle, il est judicieux d'intégrer à l'ensemble d'apprentissage des données « adversials » pour rendre le système plus robuste. Il est même possible de dédier un sous-système spécifiquement entrainé à la détection de ce type d'attaque. La phase de test devra en tout état de cause prendre en compte les exemples « adversials ».

9.4.8 Attaque par compromission de modèle

Un modèle opérationnel en production est compromis au travers de vulnérabilités dans l'application ou au niveau de l'infrastructure sous-jacente. Il s'agit en fait d'une attaque « traditionnelle » (classique) afin de compromettre la solution.

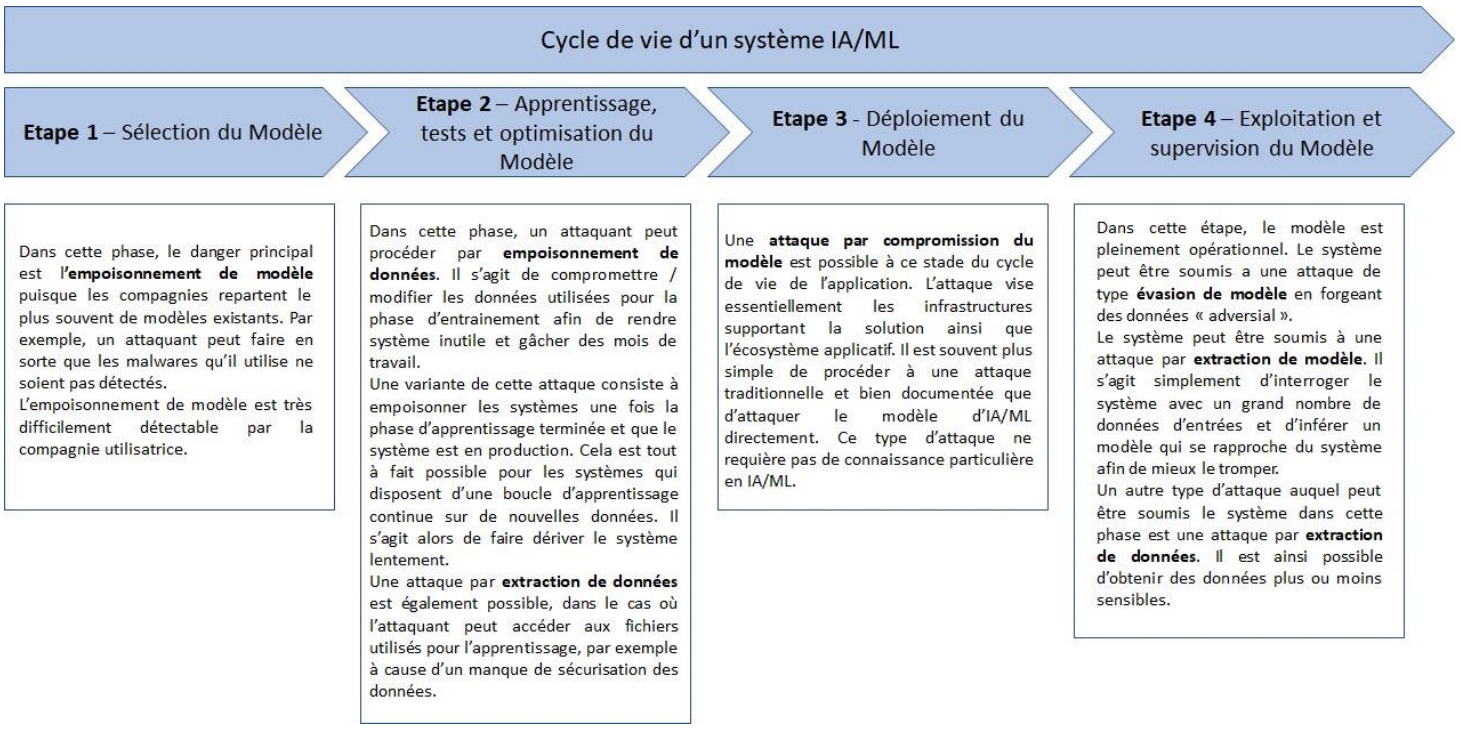
Il convient de vérifier que les librairies tierces utilisées par le système ne possèdent pas de vulnérabilités majeures, de même l'ensemble des infrastructures supportant le système. Une ligne de base du système doit être maintenue et des tests sur l'historique des réponses doivent être menés pour la vérification des dérives du système.

9.5 Typologie des attaques et cycle de vie de l'application

Le schéma ci-dessous présente le cycle de vie d'une application de type IA/ML et décrit à chaque étape du cycle, la typologie des attaques possibles sur le système. Classiquement le cycle de vie d'un tel système se décompose en quatre phases : la sélection du modèle, l'apprentissage / tests / mise au point, le déploiement et enfin la mise en exploitation.

La sécurisation d'un tel système passe par la surveillance renforcée de son usage au cours de l'ensemble de son cycle de vie. Il s'agit d'une supervision à la fois technique sur l'usage des infrastructures et leur sécurisation via des solutions standards de la cybersécurité, mais aussi d'une

supervision applicative. En effet, il est important de surveiller le comportement du système IA/ML par rapport à un jeu d'entrée prédéfini, judicieusement choisi afin de détecter toutes les dérives non contrôlées du système.



9.6 Modélisation des menaces

Il est important d'avoir la capacité de « modéliser » les types d'attaques auxquelles peuvent être soumis un système IA/ML. De façon relativement basique, nous pouvons citer les menaces qui pèsent sur :

- Les données d'apprentissage.
- Les API publiques.
- Le modèle de machine learning.
- Les infrastructures supportant le système.
- Les systèmes d'authentification (clés et « credentials »).

Les menaces peuvent être classées selon la méthode STRIDE de Microsoft :

- L'usurpation d'identité (spoofing).
- La falsification (tampering).
- La répudiation.
- La divulgation d'information.
- Le déni de service.
- L'élévation de privilèges.

Disposant des cibles d'attaques possibles et de leur classification, la modélisation des menaces se fait classiquement à travers l'identification de scénarii possibles, sachant que l'imagination des attaquants est très certainement « supérieure » à celle des équipes en charge de la sécurisation des systèmes.

9.7 Menaces et questions à adresser

Nous présentons ci-dessous une liste de questions non-exhaustive à adresser pour conduire la sécurisation d'un système IA/ML, dans le cadre d'une modélisation des menaces (Ijlal, 2022).

- Les données d'apprentissage sont-elles publiquement disponibles ? Si tel est le cas comment s'assure-t-on de la qualité des données et de leur non-falsification ?

- En cas d'attaque par extraction de données, quels contrôles sont-ils en place et comment les administrateurs sont-ils notifiés ?

- En cas d'empoisonnement des données d'apprentissage, existe-t-il un mécanisme pour s'en rendre compte ? Est-ce que des métriques sur la qualité des données sont-elles en place et susceptibles d'indiquer une falsification des données ?

- En cas d'empoisonnement des données, est-il possible d'entrainer à nouveau le système ? Est-il possible de faire un retour arrière pour repartir d'une situation saine ?

- Est-ce que les données d'apprentissage contiennent - elles des données sensibles pouvant être classées comme des données personnelles ? Si oui, comment le consentement des personnes a-t-il été obtenu ? Est-il possible de retrouver ces données ? Comment sont protégées ces données ?

- Existe-t-il des contrôles en place avant l'utilisation des données par le système ?

- Les données de sortie sont-elles expurgées avant leur envoi afin de réduire la surface d'attaque ?

- Dans le cas d'une dérive du modèle, comme la baisse de l'exactitude des prédictions, est-ce qu'un système d'alerte est en place ?

- Le système a-t-il fait l'objet d'un apprentissage contenant des exemples « de type adverses » afin de durcir le système ?

- Une attaque par déni de service est-elle possible pour les utilisateurs ?

- Quel serait le niveau d'exposition de l'entreprise en cas de vol du modèle ?

- Est-ce que le système peut faire l'objet de requêtes en mode continu sans mécanisme de limitation et ainsi dévoiler ces données d'apprentissage ?

- Quel est le système d'authentification en place et ce dernier est-il révocable ?

- Le système fait-il usage de librairies issues d'une tierce partie ? Si oui comment sont contrôlées les potentielles vulnérabilités ?

- Dans le cas où il s'agit d'un système externalisé, est-ce que l'existence d'une porte dérobée a-t-elle été étudiée ? Comment le risque d'un fournisseur malveillant est-il adressé ?

- Existe-t-il un processus d'évaluation des dispositifs de sécurité en place auprès du fournisseur de solutions IA/ML ?

- L'écosystème et l'infrastructure supportant le système IA/ML ont-ils été validés par l'équipe cybersécurité ?

Il s'agit essentiellement d'une « Check List » visant à améliorer la cybersécurité des systèmes en production. La réponse à ces questions doit donc conduire à améliorer la posture de sécurité de l'organisation.

9.8 Exemples de menaces et réponses possibles

Scénario	Catégorie	Remédiation
Un attaquant tente d'empoisonner la « Supply Chain » des outils appartenant à une tierce partie utilisés par les data scientists	Falsification Élévation de privilèges	Vérification des librairies avant usage Vérifier l'intégrité des packages (signatures, hash).
Un attaquant fait un déni de service en détruisant les données d'apprentissage	Dénis de service	Mise en place de sauvegardes régulières. Contrôle d'accès aux données d'apprentissage.
Un attaquant accède au modèle et tente de le répliquer	Divulgation d'information	Mise en place d'une limite de temps sur l'usage des API exposées pour réduire le nombre de requêtes pouvant être effectuées. Mise en place d'alertes en cas de détection d'un nombre anormalement élevé d'appel aux API. Limiter l'information en réponse aux requêtes entrantes.
Un attaquant accède aux données d'apprentissage et tente une exfiltration	Divulgation d'information	Mise en place d'une politique d'accès aux données basée sur les mécanismes de moindres privilèges. Les données d'apprentissage doivent être chiffrées.
Un attaquant accède aux clés ou aux droits administrateurs	Élévation de privilèges Divulgation d'information	Utilisation d'un système d'authentification multi facteurs pour accéder au service.
Un attaquant exploite une vulnérabilité sur un système IA/ML	Élévation de privilège Dénis de service	Le système IA/ML expose uniquement une API publique. L'application doit être durcie contre les attaques et faire l'objet d'un scan régulier de vulnérabilités.

9.9 Tests de sécurité sur les systèmes IA/ML – Spécificités

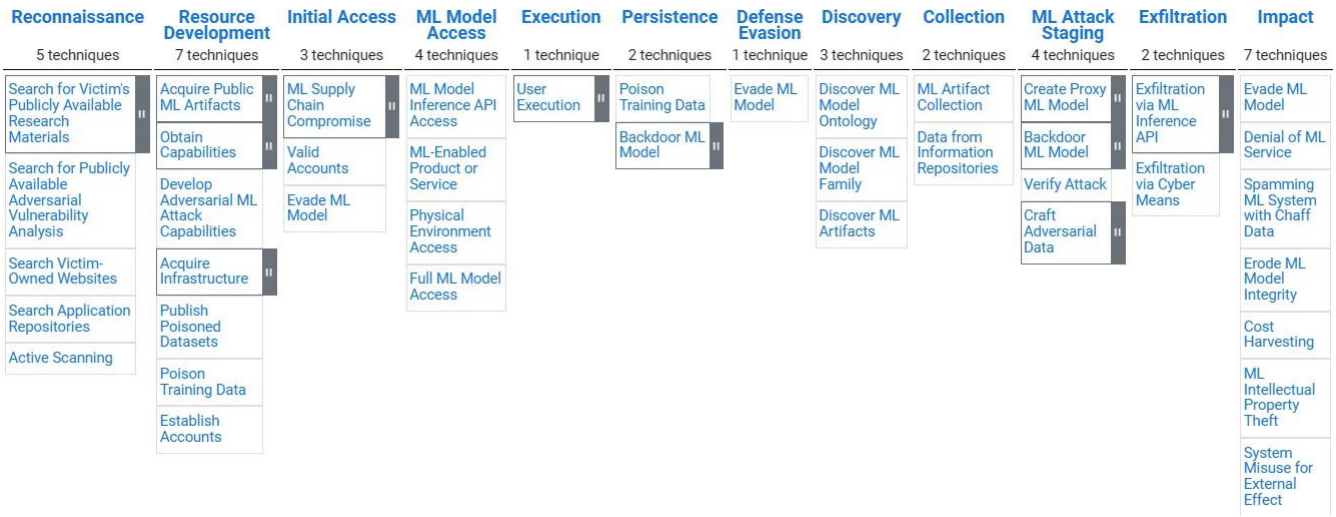
Les tests de pénétration à travers l'identification et l'exploitation de vulnérabilités existantes est un standard dans le domaine de la cybersécurité traditionnelle et sont menés classiquement par les équipes en charge de la sécurité (le plus souvent, ces tests sont externalisés à des Red Teams). Les tests de pénétration sur les systèmes AI/ML constituent cependant un domaine relativement nouveau et en cours de maturation.

Fort heureusement, il est possible de s'appuyer sur un nouveau référentiel spécifique aux systèmes IA/ML récemment développé sur les bases de MITRE ATT&CK. Il s'agit du référentiel MITRE ATLAS (Adversial Threat Landscape for Artificial Intelligence Systems). Le référentiel est disponible sur le site : https://atlas.mitre.org/. MITRE ATLAS est une base de connaissances sur les tactiques, techniques et études de cas adverses pour les systèmes d'apprentissage automatique, basée sur des observations du monde réel, des démonstrations de Red Teams IA/ML, des groupes de sécurité et de la recherche universitaire. Le site documente les attaques réelles qui se sont produites à travers sa section d'étude de cas.

MITRE ATLAS est calqué sur le référentiel MITRE ATT&CK® et ses tactiques et techniques en sont complémentaires. MITRE ATT&CK est lui un référentiel de cyberattaques généralistes basé sur des observations de comportements adverses, classées par tactiques et techniques. Ce référentiel - largement accepté - est une base de connaissances ouverte et accessible qui fournit une représentation complète des comportements d'attaque.

MITRE ATT&CK et MITRE ATLAS aident à mieux classer les attaques, à comprendre le comportement des adversaires et à évaluer les risques d'une organisation. Les équipes de sécurité peuvent également utiliser ces référentiels pour mieux comprendre le mode opératoire des adversaires dans différents scénarii et mettre au point des stratégies informées pour la détection et la prise en charge de ces comportements.

Le schéma ci-dessous illustre l'organisation du référentiel tel qu'il est présenté sur le site https://atlas.mitre.org/. Cette organisation permet de disposer d'une typologie documentée d'attaques pour supporter les tests d'intrusion et évaluer correctement les risques pesant sur les systèmes.



L'application des meilleures pratiques pour les tests de sécurité d'un système IA/ML avant passage en production, devrait reposer sur le référentiel MITRE ATT&CK pour ce qui concerne l'infrastructure et les librairies connexes supportant la solution et faire usage du référentiel MITRE ATLAS pour ce qui concerne le système IA/ML à proprement parlé.

L'organisation des tests devrait suivre le modèle Red Team / Blue Team afin de valider la sécurité du système.

- La Red Team a pour objectif de tester l'efficacité des contrôles de sécurité mis en place et de mener une série d'attaques contre le système afin d'en vérifier la robustesse.

- La Blue Team a la charge de mettre en place les contrôles internes et la sécurisation du système IA/ML pour contrer à la fois les actions de la Red Team et plus généralement des attaquants.

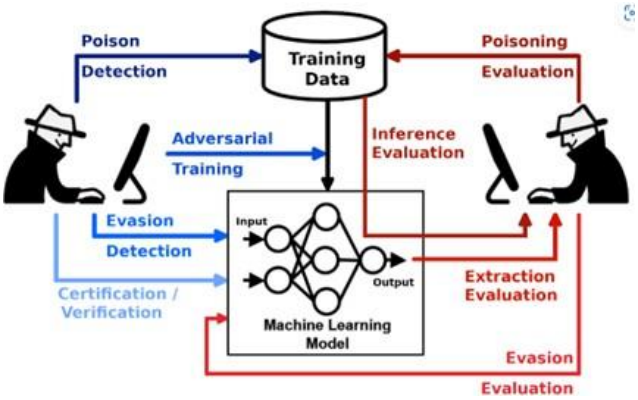
9.10 Automatisation des attaques sur les systèmes IA/ML

Microsoft a mis à disposition de la communauté un outil open source Counterfit pour automatiser les tests de vulnérabilités. La solution embarque des algorithmes d'attaque préchargés pour tester les systèmes d'apprentissage automatique.

Le développement de l'outil par les équipes Microsoft vient du besoin d'évaluer les vulnérabilités de leurs propres systèmes d'IA/ML. Counterfit était à l'origine d'une collection de scripts d'attaque écrits pour cibler des modèles d'IA/ML développés par Microsoft, et a progressivement évolué vers un outil d'automatisation pour attaquer plusieurs systèmes à grande échelle.

À présent, la solution comprend un outil de ligne de commande et une couche d'automatisation générique pour permettre aux développeurs de simuler des cyberattaques contre des systèmes IA/ML (Afifi-Sabet, 2021).

En 2021, la bibliothèque « open source » en Python Adversarial Robustness Toolbox (ART) pour la sécurité de l'apprentissage automatique a été mise à disposition de la communauté sur GITHUB (https://github.com/Trusted-AI/adversarial-robustness-toolbox). Le projet regroupe un total de 82 contributeurs à ce jour.



ART fournit un ensemble d'outils permettant aux développeurs et aux chercheurs d'évaluer, de défendre, de certifier et de vérifier les modèles et applications d'apprentissage automatique contre les menaces telles que l'évasion, l'empoisonnement et l'extraction.

ART prend en charge la plupart des frameworks d'apprentissage automatique (TensorFlow, Keras, PyTorch, MXNet, scikit-learn, XGBoost, LightGBM, CatBoost, GPy, etc.), tous les types de données (images, tableaux, audio, vidéo, etc.) et l'apprentissage automatique (classification, détection d'objets, reconnaissance vocale, génération, certification, etc.).

10 LE MARCHE DE L'IA DANS LA CYBERSECURITE

10.1 Adoption de l'IA dans la cybersécurité par les organisations

La généralisation du télétravail, l'extension des usages des offres de « cloud computing », l'éclatement des données des entreprises, ainsi que la croissance des appareils (IoT) connectés aux plateformes cloud, les interconnexions croissantes avec les fournisseurs, engendrent une multiplication et une diversité croissante des cybermenaces externes ou internes.

Cette situation explique l'augmentation « exponentielle » des cyberattaques contre les entreprises de tous les secteurs d'activité, et contre les agences gouvernementales. Elle souligne le besoin de nouvelles solutions avancées de cybersécurité. Dans ce contexte, l'IA appliquée à la cybersécurité est perçue comme une piste naturelle pour développer et commercialiser de nouvelles solutions de cybersécurité ou bien améliorer les performances des solutions existantes. Les technologies de l'IA s'inscrivent naturellement dans les stratégies de défense en profondeur des systèmes d'information. Plus précisément, L'IA/ML couplée à l'automatisation des tâches est une réponse possible pour traiter cette situation et doit permettre l'accroissement de la productivité des centres opérationnels de cybersécurité.

Les technologies d'IA/ML sont ainsi amenées à transformer la sécurité en automatisant la détection des "patterns" d'attaques, en aidant à la décision via l'analyse automatisée de données. L'automatisation permet l'orchestration des tâches, l'amélioration des temps de réponse aux incidents et la réduction de la charge de travail pour les analystes humains et donc d'exposer des gains de productivité.

De plus en plus d'organisations adoptent ce type de technologies afin d'améliorer leur posture de sécurité mais également pour réduire leurs coûts de fonctionnement. Ainsi, une enquête d'IBM réalisée en 2022 auprès de 1000 entreprises intervenant dans 16 secteurs d'activités et dans les 5 régions du globe montrent que la majorité des entreprises, à l'échelle mondiale et dans tous les secteurs, adoptent ou envisagent d'adopter l'IA et l'automatisation dans leurs fonctions de sécurité. 64 % des répondants ont mis en œuvre l'IA pour les capacités de sécurité dans au moins un des processus du cycle de vie de la sécurité, et 29 % l'envisagent (Muppidi, et al., 2022).

Concernant les apports perçus sur la sécurité opérationnelle des organisations, 67 % des utilisateurs pensent que l'IA/ML présente un apport décisif pour le filtrage et le traitement automatique des alertes de premier niveau, 66 % des utilisateurs pensent que son usage permet l'amélioration de la détection des menaces de type "zero-day", 65 % estiment que cette technologie doit permettre la réduction du bruit et des faux positifs. Enfin, 61 % positionnent les possibilités de corrélation entre le comportement des utilisateurs et les indicateurs de menaces comme un apport possible (Muppidi, et al., 2022).

10.2 L'importance du marché de l'IA dans la cybersécurité

Le marché mondial de l'IA pour la cybersécurité était estimé à 10,73 milliards de dollars en 2020, à 13,29 milliards de dollars en 2021 et à 16,5 milliards de dollars en 2022. Les projections pour 2023 font état d'un marché valorisé à plus de 22 milliards dollars. Le taux de croissance annuel du marché se situe entre 22 et 24 % et devrait continuer à croitre jusqu'en 2030. À cette date, la valeur du marché est projetée à plus de 93 milliards de dollars (Grand View Research, 2022).

Les acteurs majeurs du domaine de la cybersécurité et des secteurs connexes adoptent une stratégie de croissance externe pour intégrer ce marché de l'IA appliquée à la cybersécurité. Nous assistons à des rachats d'entreprises spécialisées en IA ou encore à l'établissement de partenariats stratégiques

par certains acteurs majeurs afin de développer leur portefeuille ou y intégrer les capacités de IA/ML (Ibid.).

Actuellement, les USA dominent complètement le marché puisque les entreprises américaines représentent plus de 37 % du secteur. Cette domination est favorisée par le développement de la 5G et de l'internet des objets ainsi qu'une recherche très dynamique dans le secteur. De façon général, la région APAC (Asie - Pacifique) est en avance sur l'Europe dans le domaine (Ibid.).

10.3 Quelques offres de cybersécurité augmentée avec de l'IA

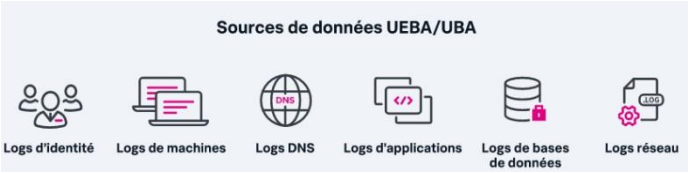
Nous présentons dans cette section une sélection de quelques entreprises de cybersécurité « pure-players » identifiées comme « avant-gardistes » par l'utilisation qu'elles font de l'IA/ML dans leurs produits. Elles proposent de nouvelles approches et stratégies de protection contre les attaques profitant de la maturité des technologies dérivées de l'IA. Ces entreprises interviennent dans différents domaines de la cybersécurité.

Bien évidemment, les entreprises intervenant dans ce domaine ne se limitent pas à cette liste, mais notre objectif n'est pas de faire un annuaire exhaustif de ces entreprises. Nous souhaitons simplement présenter les quelques entreprises reconnues comme parmi les plus innovantes dans ce secteur particulier qu'est la cybersécurité utilisant l'IA (Akash, 2022).

10.3.1 CrowdStrike

Crowdstrike est une entreprise de cybersécurité américaine réputée, fondée en 2011 et basée à Austin au Texas. Elle compte environ 6 000 collaborateurs (Wikipédia/CrowdStrike, 2022) pour un chiffre d'affaires de 1,45 Milliards de dollars en 2021 (Rendementbourse, 2022).

Cette compagnie commercialise notamment le produit « Micro Focus Interset User and Entity Behavioral Analytics (UEBA) », un système pour l'analyse du comportement des entités et utilisateurs.



La solution aide les équipes de sécurité à identifier et prendre en charge les menaces internes qui pourraient autrement échapper à la détection. Il s'agit d'identifier à travers l'analyse des fichiers de logs et du trafic les actions anormales ou suspectes. Interset UEBA s'appuie sur des modèles d'apprentissage automatique non supervisé pour extraire les entités disponibles (utilisateurs, machines, adresses IP, serveurs, imprimantes, etc.) à partir des fichiers journaux et observe les événements pour déterminer les comportements « normaux » / « habituels ». Le système est ainsi calibré - dans un premier temps - en collectant les données du trafic pour définir une « ligne de base ».

À l'issue de cette phase d'apprentissage, les nouveaux événements sont évalués par rapport aux comportements précédemment observés et des scores de risques sont calculés pour indiquer les entités/actions les plus suspectes. Tout écart à la ligne de base est détecté et remonté sur une console (CrowdStrike, 2020).

Dit autrement, l'UBA/UEBA examine les écarts dans le comportement des utilisateurs et des actifs par rapport aux actions passées ou à des groupes de pairs. La solution créée et exploite une référence pour les utilisateurs, les appareils, les applications, les comptes à privilèges et les comptes de service partagés. Le système détecte les déviations standards par rapport à la « norme ». Un score est calculé pour indiquer l'intensité de la menace en question pour faciliter l'interprétation des résultats par les équipes du SOC.

10.3.2 DarkTrace

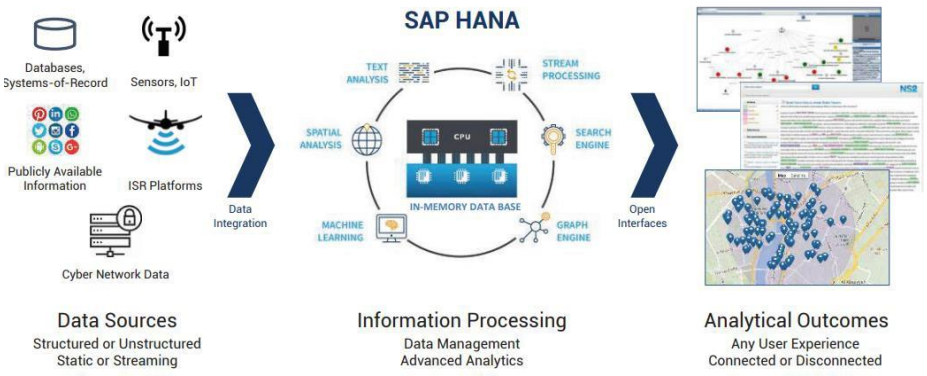
DarkTrace est une société anglo-américaine fondée en 2013 et spécialisée dans la cyberdéfense. Elle est basée à Cambridge (Angleterre) et à San Francisco (USA). La compagnie compte environ 2000 salariés et sert environ 6800 clients à travers le monde, pour un chiffre d'affaires en 2021 de l'ordre de 415 millions de dollars (Mériot, 2022). D'après le journal Challenges, l'entreprise a été initialement fondée par des mathématiciens et d'anciens agents du renseignement issus notamment du MI-6 britannique. Ses fondateurs ont développé une solution logicielle qui permet de sécuriser les organisations contre les cybermenaces grâce à l'IA (Mériot E., 2022). La compagnie met en avant le fait qu'elle dispose d'un centre de recherche comprenant plus de 150 chercheurs, entièrement dédié à l'IA. Par ailleurs, depuis le rachat de Cybersprint par Darktrace, l'entreprise dispose d'un deuxième centre de R&D est opérationnel à La Haye, aux Pays-Bas (Darktrace, s.d.).

Darktrace a développé une plate-forme de type UBA/UEBA qui utilise des méthodologies d'IA et remplit des bases de règles de statut grâce à un apprentissage non supervisé. La première chose que le système doit faire lorsqu'il est installé sur un réseau est d'établir une ligne de base d'activité « normale ». Les modèles de trafic pour chaque réseau, l'activité de chaque appareil sur le réseau et le comportement de chaque utilisateur sont modélisés pour fournir cet enregistrement de conduite standard. Tout écart est par la suite remonté.

La récente acquisition de la société hollandaise Cybersprint a permis à Darktrace de développer une nouvelle solution de prévention actuellement installée en bêta-test chez certains clients (Darktrace Prevent). Cette solution permet d'identifier les vulnérabilités et de fermer les voies d'accès à risque. Ceci est possible en émulant de manière automatique des attaques pour tester les vulnérabilités et en cartographiant les voies d'attaques les plus pertinentes (Darktrace, s.d.).

10.3.3 SAP NS2

SAP National Security Services (NS2) est une filiale américaine indépendante de SAP qui commercialise des logiciels et services pour la sécurité nationale et des infrastructures critiques. La société a été fondée en 2003 et est basée à Bethesda dans le Maryland. Elle emploie actuellement plus de 850 collaborateurs tous de nationalité américaine, tous localisés aux US et travaillant principalement pour le gouvernement fédéral américain (Chowdhry, 2018). La compagnie expose un chiffre d'affaires de 292 millions de dollars pour l'année 2021 (ZoomInfo, s.d.). Elle possède également une division de capital-risque (NS2 Ventures) qui investit principalement dans des sociétés de cybersécurité et d'analyse de données présentant un intérêt pour la sécurité nationale et/ou pour étendre les capacités de la suite SAP (Golden, s.d.).



SAP NS2 utilise les technologies de l'IA et de l'apprentissage automatique pour adresser des problématiques telles que la cybersécurité et la lutte contre le terrorisme. Son offre se base sur la mise en place d'un puit de données (datalake) couplé à un « broker » de données permettant d'agréger les données hétérogènes issues de sources variées (cf. illustration ci-contre provenant du site de SAP NS2). Le système a été spécifiquement créé au départ pour le Ministère de la Défense Américaine (DoD) et est hautement « scalable ». La technologie se base sur les solutions SAP HANA et SAP Intelligence. L'entreprise est également un fournisseur de solution de Cloud souverain autour des solutions SAP (SaaS). Il s'agit

essentiellement d'un système d'analyse de données via des méthodes statistiques et de machine learning en environnement complexe et sensible appliqué à des problématiques de sécurité nationale (Riccio, 2021).

10.3.4 Vade Secure

Vade Secure est une entreprise française basée dans la région Lilloise, spécialisée dans la conception et l'édition de solutions logicielles de sécurité des courriels. Elle adresse les problématiques de détection de courriels abusifs et/ou malveillants (spams, hameçonnages, spear phishing, malwares). La société a été initialement fondée en 2008 et est présente aux USA depuis 2014. Elle compte actuellement un peu moins de 200 collaborateurs (cf. société.com) pour un chiffre d'affaires d'environ 20 millions d'euros pour l'année 2019 (Buyse N., 2021).

La société Vade Secure a été nommée dans l'indice Next40 des start-ups françaises les plus prometteuses. L'entreprise fait ainsi partie des entreprises technologiques les plus dynamiques en France (Arnoux, 2019).

De façon assez intéressante, depuis que l'entreprise française s'est installée aux USA (notamment à Boston), elle fait face à un certain « harcèlement » juridique de la part de ses concurrents directs (Proofpoint/Cloudmark). L'entreprise a ainsi été condamnée par un jury fédéral américain à verser 13,5 millions de dollars au titre de dommages et intérêts compensatoires pour avoir contrefait des droits d'auteur appartenant à Proofpoint et sa filiale Cloudmark (Buyse, 2021).

Les solutions commercialisées par Vade Secure reposent sur une approche comportementale de la détection des menaces et va au-delà d'une analyse de la signature et de la réputation. Elle est renforcée par des technologies brevetées de machine learning et de deep learning (cf. site vadsecure.com). Les solutions intègrent :

- L'analyse comportementale heuristique : les courriels, pages Web et pièces jointes sont examinés en s'appuyant sur des règles heuristiques créées par les équipes de R&D de l'entreprise. Les règles sont continuellement mises à jour pour s'adapter aux menaces émergentes.

- L'analyse en temps réel des pièces jointes pour détecter celles potentiellement dangereuses.

- L'analyse des URL contenues dans les fichiers hébergés sur des services tiers pour détecter les URL malveillantes masquées dans les pièces jointes partagées.

- L'apprentissage automatique — Analyse les courriels et pièces jointes pour repérer les comportements suspects typiques des malwares et ransomwares. Cette approche comportementale du filtrage permet souvent de détecter un malware sans examiner le fichier à proprement parler.

10.3.5 Cynet

Cynet est une « startup » israélienne fondée en 2015 à Tel Aviv dont le siège se trouve néanmoins à Boston (USA). Elle emploie environ 727 collaborateurs pour un chiffre d'affaires estimé à environ 55 millions de dollars pour l'année 2021 (ZoomInfo, s.d.).

Cynet déploie une plate-forme (Cynet 360) pour la détection et la réponse avancées aux menaces. La solution « Cynet 360 Incidence Response Tool » est une plate-forme de protection contre les violations qui utilise l'apprentissage automatique non supervisé et l'automatisation pour gérer les vulnérabilités, les renseignements sur les menaces, analyser le comportement des utilisateurs et assurer la protection

des terminaux au sein d'un système unifié et centralisée. Le système prend en charge les déploiements SaaS, IaaS, hybrides et sur site (Cynet, 2022).

La solution peut être comprise comme un EDR/XDR (Endpoint Detection and Response) qui collecte, analyse aux niveaux des terminaux les déviances comportementales susceptibles de représenter une menace. Il s'agit en fait d'une nouvelle génération d'anti-malware, ne s'appuyant plus uniquement sur des systèmes de signatures pour effectuer une détection des comportements malveillants. L'EDR ajoute des capacités d'analyse comportementale des processus afin d'en déterminer les déviances.

Si l'EDR détecte au niveau des terminaux, le XDR (Extended Detection Response) est capable de collecter et détecter les activités déviantes et potentiellement malveillantes sur des équipements comme les serveurs, le cloud, les réseaux (TEHTRIS, 2022). La Société propose également un service infogéré 24 heures sur 24 et 7 jours sur 7.

10.3.6 Webroot

Webroot est une entreprise américaine fondée en 1997, basée à Broomfield dans le Colorado (USA). La compagnie emploie environ 600 collaborateurs pour un chiffre d'affaires de l'ordre de 215 millions de dollars pour l'année 2015. Il s'agit essentiellement d'un éditeur de solutions de cybersécurité. Par ailleurs, Webroot a été racheté en 2019, par la société Carbonite pour un montant d'environ 618,5 millions de dollar (Zippia, s.d.).

La société propose notamment une solution de type EDR/XDR en mode cloud, basée sur des algorithmes de type deep learning. L'apprentissage s'effectue sur des quantités massives de données (big data). La solution inclue notamment un antivirus/antimalware, du filtrage web, etc. D'après la documentation de Webroot, la compagnie utiliserait plus de 500 classificateurs fonctionnant en parallèle sur les URL, les IP, les fichiers, avec plusieurs langues, pour reconnaître des modèles, déterminer des réputations, et catégoriser les objets de l'internet (Webroot, 2018).

Pour Webroot, la technologie d'apprentissage automatique est devenue un élément essentiel dans tous les domaines de la cybersécurité. Les renseignements sur les menaces basés sur l'apprentissage automatique peuvent alimenter et améliorer les solutions de sécurité traditionnelle comme les « pares-feux » de nouvelle génération, les SIEM, les IPS (Webroot, 2018).

10.3.7 FireEye

FireEye est une société américaine fondée en 2004 et basée à Milpitas en Californie (US). La compagnie compte environ 3200 salariés pour un chiffre de 831 millions d'euros en 2021 (Zippia, s.d.). FireEyes propose notamment « FireEye Network Security », une solution de protection contre les cybermenaces. La solution permet de détecter et résoudre les incidents de sécurité détectés grâce à l'intégration de preuves concrètes, d'une Cyber Threat Intelligence (CTI) exploitable et des workflows de réponse à incident. La solution repose sur deux piliers : le moteur MVX (Multi-Vector Virtual Execution) et des technologies d'IA et de machine learning (IA/ML).

MVX est un moteur d'analyse dynamique sans signature qui inspecte le trafic réseau pour identifier les attaques conçues pour contourner les systèmes de défense traditionnels basés sur les signatures et les politiques de sécurité. Les multiples moteurs IA/ML et de corrélation constituent un ensemble de moteurs contextuels dynamiques basé sur des règles (Fiche Produit FireEye, 2021).

Son système de prévention des intrusions (IPS), utilise les méthodes classiques de correspondance de signatures pour identifier les attaques courantes.

D'après la compagnie, le moteur Multi-Vector Virtual Execution (MVX) peut détecter parfois certaines attaques zero-day car il ne s'appuie pas sur des signatures. La solution a été entrainée sur des informations en temps réel collectées lors de milliers d'heures de réponse à incident. Les moteurs de la solution bloquent les phases d'infection, de compromission et d'intrusion d'une chaîne d'attaque en identifiant les exploits malveillants, les malwares, les attaques de phishing et les rappels aux serveurs de commande et contrôle (Fiche Produit FireEye, 2021). Le système permet de réduire la charge de travail des équipes du SOC. La solution est composée des briques présentées dans le schéma ci-dessous.

10.3.8 Callsign

Callsign est une société britannique fondée en 2012 basée à Londres (Grande Bretagne) et à Palo Alto en Californie (USA). La société compte environ 321 salariés pour un chiffre d'affaires de l'ordre de 67 millions de dollars en 2021 (ZoomInfo, s.d.). L'entreprise propose une solution d'authentification forte basée sur l'IA/ML pour valider l'identité d'une personne. Bien que la solution puisse être utilisée seule, elle est dans les faits utilisée en tant que système d'identification multi-facteurs (MFA). Pour être plus précis, il s'agit d'un système IDA (Intelligence Driven Detection). Callsign a développé des algorithmes d'apprentissage en profondeur (deep learning) qui évaluent des centaines de points de données pour s'assurer que la personne qui utilise l'appareil est bien celle qu'elle prétend être. En plus des données biométriques standards, le système reconnait par exemple la façon de taper au clavier ou la façon d'utiliser la souris. Le système analyse les milliers de données produites lors de l'interaction avec l'ordinateur pour procéder à l'authentification. La technologie de Callsign est déjà utilisée par de grandes banques telles que la Lloyds Bank ou la Deutsche Bank avec des centaines de milliers d'utilisateurs authentifiés à travers le monde (Nanalyze, 2017).

10.3.9 Blue Hexagon

Blue Hexagon est une startup américaine fondée en 2017 et est basée à Sunnyvale, en Californie. En 2021, elle comptait 29 employés pour un chiffre d'affaires d'environ 3 millions d'euros (Growjo, s.d.). En octobre 2022, Blue Hexagon a été racheté par Qualys, une entreprise de cybersécurité leader son secteur d'activité. Les termes de la transaction n'ont pas été divulgués.

Blue Hexagon fournit une plate-forme automatisée de défense contre les menaces réseau pour prévenir les cybermenaces. La compagnie propose également des solutions de sécurité d'entreprise qui utilisent les méthodes de « deep learning ». Elle propose une solution de détection des logiciels malveillants basée sur l'apprentissage automatique (EDR/XDR).

En tout état de cause Qualys était intéressé par la technologie d'intelligence artificielle/apprentissage automatique (IA/ML) afin d'exploiter ses données (big data) pour développer une plateforme d'analyse prédictive capable de détecter certaines menaces de type zero-day et donc obtenir un véritable avantage par rapport aux entreprises concurrentes.

Ce rachat permet d'intégrer des technologies d'apprentissage (IA/ML) à Qualys Cloud Platform pour faciliter la conversion de pétaoctets de données hautement intégrées en renseignements pratiques pour les clients (Qualys, 2022).

Qualys souhaite s'appuyer sur sa plateforme dans le Cloud et sur ses plus de 10 000 milliards de points de données pour découvrir des caractéristiques comportementales, dont des exploitations actives des vulnérabilités, identifier les menaces réseau et atténuer les risques de manière adaptative (Qualys, 2022).

10.3.10 Cylance

Cylance est une entreprise de cybersécurité américaine fondée en 2012 et basée à Irvine en Californie (USA). Elle a atteint un chiffre d'affaires de 100 millions de dollars en 2017. Actuellement, ses revenus se situent sous la barre des 50 millions de dollars (Business Wire, s.d.).

Cylance a été racheté en 2018 par BlackBerry pour un montant de l'ordre de 1,4 milliards de dollars, elle est actuellement une division de BlackBerry Limited (BlackBerry Limited, 2018).

La société est certainement l'un des fournisseurs pionniers de solutions de sécurité (EDR) utilisant les technologies de l'IA pour la détection de malwares/virus. Le système de Cylance est capable d'identifier l'utilisation malveillante de la mémoire vive (Attaques fileless) et de déclencher une réponse automatique. L'entreprise est également devenue un leader dans le domaine des IPS (BlackBerry Limited, 2018).

Le cas de cette entreprise est très intéressant car il apporte un éclairage particulier sur les limites possibles des techniques d'IA/ML dans le domaine spécifique des EDR. En effet, un contournement a récemment été découvert et publié par des chercheurs de Skylight, une société fondée par des vétérans de la sécurité du gouvernement israélien.

Après une analyse du produit antivirus de Cylance, les chercheurs ont découvert un biais envers un jeu de données particulier (Lahoti, 2019).

De façon assez standard, l'algorithme d'apprentissage automatique de Cylance a été entrainé pour favoriser les fichiers non malveillants, l'amenant à ignorer le code malveillant s'il détectait des chaînes de caractères spécifiques de fichiers bénins jointes à un fichier malveillant. Les chercheurs ont ainsi pu ajouter des chaînes d'un fichier non-malveillant à un fichier malveillant et éviter la détection.

Les chercheurs ont élaboré une méthode « universelle » permettant de contourner le logiciel en ajoutant simplement une liste sélectionnée de chaînes de caractères à tout fichier malveillant. Selon les chercheurs, la méthode a été efficace à 90 %, lorsqu'elle a été testée sur 384 applications malveillantes (Mello Jr., 2019).

Cylance a reconnu que son algorithme basé sur l'IA/ML était « défectueux », mais a déclaré dans un article de blog de l'entreprise qu'il ne s'agissait pas d'un contournement universel. La société a ajouté qu'elle avait corrigé le problème dans son service cloud et qu'elle le ferait sous peu avec son logiciel EDR (Mello Jr., 2019), (Lahoti, 2019).

Cet exemple illustre bien le fait que si l'IA/ML apporte des solutions vis-à-vis des attaques auxquelles sont soumises les entreprises, elle ne saurait être l'unique solution. Ces technologies ouvrent naturellement la voie à de nouvelles attaques. L'IA/ML ne saurait se substituer à une défense en profondeur du SI de l'entreprise. Les EDR basés sur l'IA/ML sont complémentaires des systèmes fondés

sur des bases de signatures. Ces technologies doivent certainement s'inscrire dans le développement et le déploiement de systèmes hybrides.

10.3.11 Et bien d'autres entreprises

Nous avons déjà évoqué le cas de Google qui n'intervient pas directement sur le marché de la cybersécurité, et qui développe et utilise depuis longtemps l'IA/ML pour sécuriser sa plateforme Gmail à l'aide de solutions développées en interne. Ainsi, la solution de filtrage de mail de Google utilise depuis des années le deep learning (Lou, 2019). Google a, pour ce faire, créé et mis à disposition de la communauté, la bibliothèque de deep learning TensorFlow, compatible avec le langage Python. Précisons ici, que la solution de filtrage de mail ne repose - bien évidement - pas uniquement sur une solution d'IA/ML. Cette solution vient en complément de solutions plus traditionnelles.

De même, et depuis quelques années maintenant, les entreprises historiques de la cybersécurité se mettent à intégrer la technologie IA/ML dans leurs offres afin de garder leurs parts de marché et en conquérir d'autres. Sans être exhaustif, nous pouvons citer les acteurs suivants :

- Fortinet
  - En 2020, l'entreprise a mis sur le marché FortiAI, appliance[32] déployée sur site et qui tire parti de réseaux neuronaux profonds (Deep Neural Networks) pour « accélérer » le traitement des menaces et gérer les tâches manuelles et chronophages d'analyse des menaces. L'entreprise décline l'IA/ML sur sa gamme de produit (Sandbox, EDR, Analyse du trafic Web, solution SIEM, etc.). Ceci montre simplement que Fortinet étend les fonctionnalités de son offre avec des capacités d'intelligence artificiel en profitant des techniques d'apprentissage supervisé et non-supervisé (Fortinet, 2020).

- Check Point
  - En 2022, Check Point a par exemple présenté Check Point Quantum Titan, une nouvelle version de la plateforme de cybersécurité Check Point Quantum. Le produit introduit trois nouveaux logiciels qui exploitent l'IA et l'apprentissage en profondeur, pour fournir une solution de prévention pour les attaques du système de noms de domaine (DNS), le phishing, et assurer la sécurité des IoT autonomes (Check Point Press Releases, 2022).

- Sophos
  - Sophos propose à travers sa solution Intercept X, un système EDR qui utilise l'IA (un réseau de neurones d'apprentissage profond). Le système a, en fait, été développé par Invincea, une société acquise par Sophos en 2017. La solution garde une trace du comportement « normal » sur l'appareil protégé et envoie des notifications lorsque quelque chose d'inhabituel se produit. Lorsque des exploits et des virus sont détectés, l'EDR déclenche des workflows et des actions pour les arrêter et les isoler (Gaikwad, 2022).

- Symantec
  - Symantec, actuellement connu sous le nom de NortonLifeLock, aide les entreprises à protéger leurs infrastructures contre les menaces de cybersécurité. Symantec utilise l'IA pour élargir ses efforts de détection et de prévention des menaces. Les services de sécurité basés sur l'IA de Symantec incluent la protection des terminaux, la défense des applications de messagerie de même que les infrastructures cloud.

[32] Équipement informatique dédié à une fonctionnalité (ex. sauvegarde, filtrage de flux, système de détection d'intrusion)

- Notons qu'Accenture a racheté en janvier 2020 la division « Cyber Security Services » de Symantec à Broadcom, ce qui en fait l'un des leaders des services infogérés de cybersécurité. Cette division assure la surveillance et l'analyse des menaces à l'échelle mondiale avec une équipe d'environ 300 personnes répartis sur six centres de services à travers la planète. Le service s'appuie sur la solution Symantec Targeted Attack Analytics (EDR &TAA) qui modélise le comportement du réseau et crée une base de performances à l'aide de l'apprentissage automatique. Tout écart par rapport à la « norme » déclenche une alarme. Les fonctionnalités d'IA de TAA reposent sur la plate-forme de cyberdéfense Symantec, qui peut collecter des données de performances à partir de plusieurs points du réseau. TAA est désormais disponible dans la famille de produits Symantec Advanced Threat Protection (Gaikwad, 2022).

- IBM
  - Watson est la plate-forme d'IA, d'apprentissage automatique et d'informatique « cognitive » d'IBM. Elle fournit un large éventail de technologies d'IA pour traiter à la fois des informations structurées et non structurées à partir d'un large éventail de source.
  - La plateforme IBM QRadar Advisor with Watson combine les fonctionnalités d'un SIEM et l'IA (apprentissage automatique) en s'appuyant sur Watson. Le système fournit une aide à l'analyse des alertes. Il permet d'automatiser l'analyse des « root-causes » des alertes et le traitement des menaces répétitives du centre des opérations de sécurité (SOC) et ainsi de réduire les temps de traitement (NOVIPRO, 2019).
  - Par ailleurs, l'application User Behavior Analytics for QRadar aide à déterminer les profils de risques des utilisateurs au sein du réseau et à prendre des mesures lorsque l'application détecte un comportement menaçant. Cette solution ajoute deux fonctions principales à QRadar : le profilage des risques et les identités utilisateurs unifiées (IBM, 2023).

- CISCO
  - En 2013, Cisco a fait l'acquisition de Cognitive Security, une petite société privée de 28 personnes dont le siège est à Prague, en République tchèque. Cognitive Security possédait une expertise dans les techniques d'IA appliquées à la détection des cybermenaces. Leur solution intégrait une gamme de technologies IA/ML pour identifier les menaces via une analyse comportementale des données en temps réel (Cisco, 2013). Depuis, Cisco Cognitive Intelligence offre des fonctionnalités avancées de détection des menaces dans une partie de son portefeuille de sécurité. Via les méthodes d'apprentissage non supervisé, Cognitive Intelligence peut détecter les menaces, telles que les menaces se cachant dans le trafic chiffré, les logiciels malveillants polymorphes, d'après le site de CISCO (Bettex, 2018).

A cette liste, nous pouvons ajouter bien d'autres acteurs tels que : Palo Alto Networks (US), Amazon Web Services (US), Microsoft (US), Trellix (US), ThreatMetrix (US), Securonix (US), Sift Science (US), Acalvio Technologies (US), SparkCognition (US), High-Tech Bridge (Suisse), Deep Instinct (US), SentinelOne (US), Feedzai (US), Vectra Networks (US), Zimperium (US), Argus Cyber Security (Israël), Nozomi Networks (US), BitSight Technologies (USp, Kaspersky Lab (Russie), Bitdefender (Roumanie) ou encore EST (US), Samsung (Corée du Sud), etc., (MarketsandMarkets, 2022).

Cette liste a pour unique mérite de montrer que même s'ils ne sont pas seuls les États-Unis dominent largement le secteur que ce soit au niveau des très grandes entreprises ou au niveau des startup qui constitueront peut-être les géants de demain.

10.3.12 Et les entreprises chinoises ?

De façon paradoxale, dans les deux listes d'entreprises novatrices présentées ci-dessus, nous ne trouvons pas encore d'entreprises chinoises. Ceci s'explique par au moins deux facteurs :

- La Chine conserve un certain retard sur les USA en ce qui concerne les entreprises qui éditent des solutions de cybersécurité.

- Il y a une certaine opacité sur les activités des entreprises chinoises et il est plus difficile de trouver des informations sur elles.

Il existe – bien évidemment - des entreprises chinoises qui sont présentent dans ce secteur d'activité et qui bénéficient de l'appui des autorités du pays. Uniquement à titre d'exemple, nous pouvons citer :

- Antiy Labs est une entreprise de cybersécurité fondée en 2000, basée à Pékin et leader sur le marché chinois. La compagnie est connue pour développer notamment un moteur antivirus pour ordinateurs et téléphones portables. En 2020, la compagnie a fait une levée de fond de 86 millions de dollars (Tracxn, 2023). Elle dispose actuellement de 6 centres de recherche. La société est également présente aux USA et au Canada. Elle fournit tout un portefeuille de protection des terminaux, de surveillance du réseau et de réponses rapides basées sur des technologies propriétaires, parmi lesquelles un moteur de détection des menaces de nouvelle génération, l'analyse homme-machine intégrée, l'IA.

- ThreatBook est un autre exemple, cette entreprise a été créée en 2015 et est basée Pékin. Elle compterait environ 122 collaborateurs pour un revenu d'environ 24 millions de dollars en 2021 (ZoomInfo, s.d.). Elle commercialise une solution de type EDR et une plateforme TDP (Threat Detection Platform) pour la détection de comportements anormaux sur le réseau.

- Tophant est une entreprise fondée en 2014 et basée à Shangaï. L'entreprise compte moins de 100 salariés. Elle commercialise une plateforme de surveillance pour la sécurité réseau qui applique les techniques d'IA pour la détection et la réponse aux comportements anormaux (Crunchbase, s.d.).

Compte tenu des investissements en cours en Chine dans l'IA, nous pouvons anticiper l'apparition de nouvelles entreprises dans le domaine de l'IA pour la cybersécurité, considérant que ces deux domaines sont identifiés comme hautement stratégique pour la sécurité nationale par le gouvernement chinois. Ces entreprises ne manqueront pas de devenir des leaders sur leur marché national et sans doute à l'échelle internationale.

11 ÉVOLUTION ET CONSÉQUENCES DE L'IA AU SEIN DES ORGANISATIONS

En 1943, Thomas Watson alors président directeur général d'IBM et considéré comme une personnalité très visionnaire, prononçait cette célèbre phrase, *« I think there is a world market for maybe five computers »*. Ceci doit nous rappeler qu'en matière de prospective, il convient de rester très modeste et d'être toujours prêt à remettre en cause ses croyances, car en définitive seulement ce qui est démontrable à une valeur scientifique.

Dans ce qui suit, nous devons garder à l'esprit que même une IA de type LLM telle que l'application ChatGPT n'est en dernière instance, qu'une automatisation du processus de compilation, de mémorisation et de synthèse documentaire, couplée à des techniques de traitement du langage naturel, afin de produire un automate conversationnel capable de répondre à des questions d'ordre générale ou plus spécialisées, en prédisant une séquence de mots en sortie, à partir d'une séquence de mots en entrée. Ce type d'IA est par ailleurs continuellement amélioré et mis à jour grâce à l'apprentissage automatique et à l'interaction avec les utilisateurs. Ce qui impressionne finalement, c'est la capacité de ces automates à générer à l'infini des textes structurés et cohérents, des images, des vidéos, de la musique, bref des contenus sur une multitude de sujets.

La puissance du système repose sur un empilement de réseaux de neurones et surtout une masse de données d'apprentissage absolument colossale. Présenté ainsi, cela peut paraître sans conséquence, mais en réalité nous pouvons raisonnablement anticiper un impact bien plus grand que celui qui a été initié au XVIII^{ème} siècle à travers l'automatisation de processus manuels spécifiques pour conduire à l'industrialisation de nos sociétés, puis plus récemment à leur numérisation.

11.1 Prospective des impacts de l'IA sur la cybersécurité

Comme nous l'avons vu tout au long de cette étude, les usages de l'IA/ML dans la cybersécurité sont multiples et remontent déjà à plusieurs dizaines d'années. L'idée d'utiliser cette technologie pour la sécurité des systèmes et des réseaux remonte à 1987, lorsque les chercheurs ont commencé à construire des systèmes de détection d'intrusion (IDS) (Denning, 1987). Si l'IA a une longue histoire avec la cybersécurité, son usage à l'avenir ne peut que se généraliser et s'étendre au niveau de toutes les strates de cette discipline. Clairement, une boucle de rétroaction est enclenchée entre la cybersécurité et l'IA, impliquant à l'avenir des interactions plus fortes et des échanges plus soutenus entre les experts des deux « communautés », qui conduiront au développement de nouveaux produits ou de nouvelles méthodes. Le succès de l'IA réside essentiellement de la convergence et de la massification des données ainsi que de l'accroissement des puissances de calcul. Finalement, la barrière pour appliquer l'IA à la cybersécurité est relativement faible du point de vue de l'accès et l'usage des algorithmes. La barrière réside plus dans l'accès aux données et à l'accès à une puissance de calcul suffisante.

Du point de vue des pratiques, dans les années à venir, la cybersécurité consistera probablement moins à « défendre la forteresse » qu'à progresser vers l'acceptation des cyber-risques permanents, en mettant l'accent sur le renforcement de la résilience et de la capacité de récupération des systèmes. Dans un monde de médias synthétiques de plus en plus sophistiqués et de cyberattaques basées sur l'intelligence artificielle, la cybersécurité portera moins sur la protection de la confidentialité que sur la protection de l'intégrité et de la provenance des informations. Marqueurs de cette tendance, et avec l'avènement de nouvelles technologies d'authentification, les mots de passe pourraient être quasiment obsolètes d'ici 2030. Les progrès de l'intelligence artificielle (IA) et de l'apprentissage automatique (ML) feront que nous vivrons dans un monde dans lequel la distinction entre les humains

et les automates dans le cyberespace sera de plus en plus difficile. Ceci pourrait conduire de nombreuses personnes à réduire leurs activités dans le cyberespace.

Par ailleurs, la sociologie des équipes en charge de la sécurité des systèmes d'information est appelée à évoluer dans les prochaines années en faisant plus de place pour les « data scientists » et les ingénieurs en IA. Les métiers de la cybersécurité devraient donc certainement s'adapter et évoluer. L'IA et l'automatisation devraient aider à atténuer les problèmes causés par la pénurie de compétences observée dans le domaine. Les professionnels de la cybersécurité devront s'adapter et acquérir encore d'avantage d'expertise, en contrepartie leur capacité d'intervention sera augmentée par l'utilisation de l'IA pour la détection et la qualification des menaces mais également pour la génération de rapports pertinents. Bien évidemment, nous pouvons anticiper une augmentation de la cybercriminalité considérant que l'IA va abaisser davantage la barrière d'entrée sur ce « marché ». Le travail des cybercriminels s'en trouvera simplifié. L'anticipation de la croissance de la cybercriminalité dynamise naturellement le marché de la cybersécurité.

Cependant, beaucoup de métiers et notamment ceux de premier niveau seront à termes plus ou moins condamnés. Pour illustrer ce propos, même si l'exemple n'est pas spécifique à la cybersécurité, en France, le groupe bancaire BPCE a mis en place un projet de « Chatbot » en 2018 pour répondre aux demandes de ses clients. Cette intelligence artificielle conversationnelle, baptisée « Emma », a été conçue pour aider les clients à gérer leur compte bancaire et répondre à leurs questions en temps réel. En 2019, le groupe a annoncé que le Chatbot avait traité plus de 20 millions de conversations et qu'il avait répondu à plus de 80 % de demandes des clients de manière autonome. (www.cgt-bpce.fr, s.d.) L'introduction d'Emma a également entraîné la suppression de plusieurs centaines d'emplois dans les centres d'appels de la BPCE, où les employés étaient chargés de répondre aux appels et aux messages des clients. Selon la CGT cette automatisation aurait conduit à la suppression de plus de 500 emplois entre 2018 et 2020.

Pour revenir au domaine de cybersécurité, une étude du fournisseur de sécurité Trend Micro effectuée en 2021 montre que les professionnels de l'informatique pensent que l'IA détruira plus d'emplois qu'elle n'en créera. L'étude a été compilée à partir des entretiens avec 500 directeurs techniques, DSI et responsables informatique. Seuls 9 % des personnes interrogées étaient convaincues que l'intelligence artificielle ne remplacerait certainement pas leur emploi dans les dix prochaines années. 32 % estimaient que les technologies permettront à terme d'automatiser toute la cybersécurité, avec un faible besoin d'intervention humaine. Et environ 19 % des professionnels interrogés estimaient que l'usage de l'IA par les attaquants sera monnaie courante d'ici 2025 (EPSI, 2021).

Dans la suite de ce chapitre, nous tentons d'élargir le champ de notre réflexion au-delà du cas spécifique de cybersécurité. En effet, nous pensons que l'IA impactera toutes les strates des organisations humaines. Il nous est donc apparu indispensable de mener une brève réflexion sur le sujet.

11.2 Incidences de l'IA sur les organisations

L'IA sans que l'on s'en rende compte est finalement déjà une réalité dans les organisations et transforme les pratiques liées au travail. Pour illustrer le propos, nous citerons la présence depuis quelques années déjà, d'agents conversationnels synthétiques (cf. exemple précédant) dans les centres d'appel ou encore le traitement des contrats par les services juridiques.

L'IA permet également une amélioration de l'expérience client dans le domaine notamment de la grande distribution, en permettant une personnalisation améliorée des offres. Ce type d'application est clairement amené à se développer dans les années à venir, transformant de manière profonde cette industrie. Ceci est et sera réalisé en analysant les données des clients de manière massive et systématique et en offrant des recommandations personnalisées, au prix d'un suivi et d'une

surveillance toujours plus grande. Les techniques d'analyse de grandes volumétries de données issues du « Big Data » aident déjà les décideurs dans leurs prises de décisions stratégiques, afin de définir de nouveaux segments de marché ou mieux comprendre leurs cibles.

Dans les années à venir, l'IA aura donc un impact significatif sur les organisations de travail, avec des évolutions majeures notamment en ce qui concerne l'automatisation des tâches et processus qui jusque-là relevaient uniquement de l'intelligence humaine. Dans une vision optimiste, cette tendance devrait permettre de libérer du temps pour les employés afin de se concentrer sur des tâches encore plus créatives et à plus forte valeur ajoutée, améliorant ainsi davantage l'efficacité et la productivité des entreprises.

11.3 Impacts sur l'emploi et le travail

Le développement et l'essor de toute nouvelle technologie induit logiquement une demande de compétences spécifiques, telles que les mathématiques, les statistiques, l'analyse de données et l'IA/ML. Cette demande comme toujours induira une tension sur ces profils et devrait conduire les entreprises à investir dans la formation de leurs collaborateurs pour s'assurer que leurs employés possèdent les compétences requises pour maintenir leur positionnement. De la même façon, les employés devront mener une réflexion quant à leur employabilité dans un environnement en évolution constante, ils n'auront d'autre choix que s'interroger sur leur plus-value au travail.

Cela étant dit et malgré le dogme de « *la destruction créatrice* » promu par Joseph Schumpeter, l'automatisation des tâches de cognition entrainera la perte d'emplois qualifiés. Il appartiendra aux autorités à travers notamment des actions de régulation, d'inciter les entreprises à minimiser leur impact sur le personnel. Dans la théorie de Schumpeter, la destruction créatrice a lieu de manière récurrente, dès lors que l'irruption d'une innovation, réduit les avantages compétitifs de certaines entreprises, et provoque ainsi la recomposition du tissu de production de la valeur avec des destructions d'emplois qui seront remplacés par d'autres.

Bien évidemment, les entreprises qui savent profiter des opportunités créées par ces innovations, sont ainsi capables d'améliorer leur productivité, et se retrouvent dans une position dominante. C'est déjà le cas pour les GAFAMI et autres BHATX qui accompagnent de manière étroite cette révolution.

Clairement, des emplois qualifiés seront détruits dû au fait que la productivité des techniciens et ingénieurs se trouvera forcément augmentée par le développement et la généralisation des usages de l'IA dans tous les secteurs d'activité. Mais les décideurs et les grands acteurs de l'économie – comme attendu – reste sur une vision « à la Schumpeter » pour ce qui concerne le marché du travail.

D'après Sundar Pichai (PDG de Google), *« Les emplois en première ligne sont les travailleurs du savoir, notamment les producteurs de contenus, les comptables, les ingénieurs et les architectes logiciel »*. Venant confirmer ce propos, une étude prospective de Goldman Sachs anticipe que l'IA pourrait affecter, dans les quelques années à venir, environ les deux tiers des emplois actuels (Golman-Sachs, 2023), (Futura-Science, 2023). Ainsi, les seuls modèles génératifs pourraient remplacer totalement jusqu'à un quart des postes actuels. Uniquement, sur le périmètre États-Unis et Europe, environ 300 millions d'emplois à temps plein seraient affectés. Les professions administratives et juridiques seraient impactées avec des suppressions de postes de l'ordre 44 % à 46 %. Les professions liées à l'entretien et au nettoyage devraient être les plus affectées avec jusqu'à 95 % de suppressions de postes, en raison de systèmes automatisés gérés par des IA. Il en est de même pour les services de réparation (85 %), de production (72 %), de transport et de déplacement de matériel (65 %). Les professions liées à l'alimentation seront également touchées à hauteur de 50 %. Les métiers relatifs à l'ingénierie au sens large, devraient voir leurs effectifs réduits de 10 %. Les métiers du commerce et de la finance ne seraient affectés qu'à hauteur de 4 % de suppression de postes (Ibid.).

Comme attendu, les intelligences artificielles devraient également engendrer de nouveaux emplois avec une productivité plus élevée et un coût de main-d'œuvre plus faible qu'aujourd'hui. Toujours d'après les analystes de Goldman-Sachs, la généralisation de l'IA pourrait augmenter la croissance annuelle de la productivité du travail, ce qui générerait un accroissement du produit intérieur brut (PIB) mondial annuel d'environ 7 %. Ce même chiffre de 7 % devrait correspondre au nombre de travailleurs américains qui seront licenciés à cause de la généralisation des usages de l'intelligence artificielle (Ibid.).

L'IA ne fonctionne pas « toute seule », et il est important de noter que les systèmes d'IA sont pensés et conçus pour augmenter et non remplacer les agents humains. Dans ce contexte, il est clair que le remplacement de tous les métiers « intellectuels » n'est pas pour demain. Un consensus semble se dégager dans le milieu des entreprises sur le fait que l'IA génèrera de nouveaux revenus et permettra en même temps d'importants gains de productivité à travers la réduction des structures de coûts. Le risque pour les entreprises est de ne pas prendre la vague de l'IA et de se retrouver dépassées et obsolètes (Vergera, 2023). En réalité, le choix n'existerait plus.

Il convient d'avoir conscience que ces chiffres ne sont issus en définitive que de modèles économiques et se basent sur des hypothèses qui restent à vérifier. L'économie n'a jamais particulièrement brillé par sa capacité prédictive et notamment par sa capacité à prédire les crises. Citer la position de grandes institutions et ou de personnes célèbres ne saurait cacher le fait que ces arguments s'apparentent parfois à l'argumentation d'autorité et illustre la faiblesse des preuves supportant le propos.

Une objection à cette vision finalement assez « optimiste » est, que pour la première fois dans l'Histoire, nous nous retrouvons confronté à une innovation qui automatise les processus qui fondent l'Humanité à savoir la cognition et le travail « intellectuel ». Puisque le travail fonde nos sociétés en permettant la coopération sociale, il est évident que cette technologie touche aux structures profondes de nos civilisations et nous incite à nous questionner sur notre place dans la société.

Sans Doute, pour la première fois de l'Histoire, les générations futures seront confrontées à la possible automatisation d'une grande partie des activités humaines liées au travail et parmi elles les plus gratifiantes.

Sommes-nous à l'aube d'une société post-travail ? Sommes-nous au bord d'une singularité ? Car si nous avons la capacité d'automatiser des chaines complètes de production, une grande partie des processus cognitifs et que nous choisissons de le faire, le travail humain ne pourra plus être à la base de notre société. De façon prosaïque, ce dernier est la source principale des revenus (pour la grande majorité de l'humanité) qui autorise l'accès à la consommation et qui définit en grande partie la position dans la société. Pour Pierre-Yves Gomez (Économiste et Co-initiateur du Courant pour une écologie humaine), *« Le travail fait l'Homme. Il l'humanise. L'Homme est un animal qui travaille »*. D'après ce courant de pensée, il est légitime de se poser la question de savoir si c'est notre humanité même qui pourrait être remise en cause. Notre propos est simplement de souligner que l'avènement de cette technologie doit inciter nos experts en sciences humaines à s'emparer de ce sujet afin de développer de nouvelles idées et d'alerter la société lorsque cela est nécessaire.

11.4 Limiter les impacts sur la cohésion sociale

Conscientes que l'automatisation à outrance des entreprises est un mouvement de fond et qu'elle fait porter un grave risque sur la cohésion sociale de nos sociétés, des personnalités célèbres, telles que Bill Gates, Benoit Hamon ou encore Mady Delvaux, députée européenne luxembourgeoise, ont avancé l'idée d'une taxe sur les robots afin de financer des éléments de protection sociale, comme un revenu universel (Barilari, 2018). Par robots, nous entendons ici toutes les entités cybernétiques incluant les IA. En effet, le travail humain crée de la richesse dont une partie est également captée par l'état à travers l'imposition pour financer nos services publics.

Bill Gates a ainsi déclaré, *« À l'heure actuelle, si un travailleur humain produit, disons, une richesse de 50 000 dollars dans une usine, ce revenu est taxé. Si une machine vient et fait la même chose, on pourrait penser que nous imposerions le robot à un niveau similaire »*. Et, il poursuit en ajoutant *« Une partie [du financement] peut provenir des profits qui sont générés par les gains d'économie de main-d'œuvre. Une partie peut venir directement d'un certain type de taxe pour les robots* » (Barilari, 2018).

Nous devrons selon toute vraisemblance repenser la manière dont des agents cybernétiques sont valorisés et taxés, pour continuer à faire société. Dit autrement, nous devrons repenser les politiques d'imposition des différentes composantes de la société et réfléchir à la taxation des usages de l'IA marchande (Barilari, 2018).

Pour être complet, il existe trois arguments qui tentent de contredire cette thèse. Comme attendu, le premier argument repose sur le mécanisme « schumpétérien » de la destruction créatrice. Les emplois détruits seraient compensés par la création de nouveaux emplois, comme précédemment évoqué. C'est un argument assez faible, car il ne repose que sur des projections. De plus, nous sommes légitimes à poser la question suivante : quel travail reste-t-il dans une civilisation capable d'automatiser toutes les activités humaines, même la recherche ?

Le second argument est lui plus subtile. L'automatisation de tous les secteurs d'activité du marché, génère des gains de productivité et favorise la baisse des prix. Elle libère en conséquence du pouvoir d'achat, ce qui permet un redéploiement de la demande et de l'emploi vers de nouveaux services. Une taxe sur les robots freinerait alors les gains de productivité et le redéploiement des activités vers de nouveaux métiers. Dans le cas de l'IA, ce mécanisme n'est pas du tout évident. L'IA pose non seulement le problème de la disparition de certains emplois « intellectuels », mais elle a pour conséquence indirecte de freiner les augmentations de salaires, ce qui a pour conséquence de réduire « de facto » le pouvoir d'achat.

Le troisième argument, est que la fabrication et la commercialisation des machines ou logiciels s'insère déjà dans une série de règles fiscales. Notamment, la cotisation sur la valeur ajoutée des entreprises (CVAE en France). Cet argument paraît également faible, en effet d'après le site des impôts en France, le taux maximal d'imposition se situe à 0,375% auquel s'ajoute une taxe additionnelle de 6,72%. Il y a de multiples débats sur les modalités d'une taxation des robots et par extension de l'IA et à ce jour aucun consensus ne s'est dégagé. Mais, il y a urgence à mener cette réflexion et surtout à agir.

Afin d'aboutir à une solution viable et d'adopter une démarche pragmatique, la modélisation mathématique pourrait être utilisée afin de simuler les effets de l'IA sur la répartition des richesses, en tenant compte des changements dans les niveaux de revenus, des coûts de main-d'œuvre, des taux d'emploi, des taux d'automatisation et d'autres facteurs pour simuler les effets d'une possible taxation sur les agents cybernétiques. Bref, il apparaît urgent de développer des modèles mathématiques sérieux capables de simuler les flux économiques sur ordinateur. Les modèles ainsi générés pourraient être utilisés et fournir une analyse plus complète de l'impact de l'IA sur l'économie.

11.5 Des impacts écologiques ?

Si le développement de l'IA et plus particulièrement des LLM à grande échelle pose des questions d'ordre économique et philosophique, il pose également des questions d'ordre écologique, de façon moins évidente. En effet, ce type d'application fait appel au calcul massivement parallèle pour les phases d'entrainement du modèle. Le développement de cette technologie est donc adossé à une puissance de calcul colossale de type HPC (High Performance Computing) et nécessite donc l'accès à des centres de calcul ultra-performants. L'entrainement de ce type de modèle de classe industrielle implique l'utilisation de centaines de milliers de cœurs de processeur pendant des mois. De même leur utilisation en phase de production, nécessite une grande puissance de calcul et beaucoup de mémoire pour fonctionner correctement. Il est donc nécessaire de disposer d'une infrastructure informatique

robuste et hautement évolutive. Ceci conduit naturellement à la consommation d'une grande quantité d'énergie, générant ainsi une empreinte carbone importante.

À titre d'exemple, La consommation énergétique de GPT3 n'a pas été rendue publique, mais d'après les estimations les plus récentes, basées sur son nombre de paramètres à optimiser, qui est de 175 milliards pour la version GPT-3, *« la puissance électrique nécessaire pour un seul entraînement de GPT 3 serait supérieure à 4 GWH, ce qui correspond à la puissance électrique d'une tranche complète d'une centrale nucléaire pendant quatre heures. Et, il ne s'agit là que de l'entrainement que d'une seule version »* (Gaultier, 2022). Les impacts en termes de responsabilité sociétale et environnementale sont ici évidents :

- Consommation électrique élevée au niveau des centres de données ;
- Émission de gaz à effet de serre de façon indirecte puisque la majorité de la production énergétique repose sur les technologies du non-renouvelable ;
- Impact sur les ressources naturelles ;
- Production de déchets électroniques en effet les centres de données génèrent des déchets ;
- Impact sur la biodiversité.

11.6 Maîtrise du marché de l'IA

Finalement, en étudiant la littérature, les questionnements liés autour du développement de l'IA et de ses usages concernent essentiellement les modèles de type LLM sans doute parce qu'ils saturent depuis quelques mois le paysage médiatique. Signalons que les autres domaines de l'IA que nous avons également évoqué tout au long de notre travail posent également questions, mais font finalement moins peur. La raison est sans doute à rechercher à plusieurs niveaux. En effet, la barrière d'entrée sur le marché du LLM est très élevée puisqu'elle requière d'importants moyens tant en compétences humaines qu'en ressources financières. Ceci revient à dire qu'à l'exception des États, seules quelques grandes entreprises ou des start-up adossées à des fonds d'investissement prêts à investir massivement peuvent espérer prospérer sur le développement des LLM. Ce raisonnement peut néanmoins être tempéré par l'usage répandu des modèles généralistes pré-entrainés qui permettent de réduire les temps d'apprentissage et ainsi réduire les coûts de développement, du fait que la convergence en partant de ce type de modèle est plus rapide. Il est entendu que ces modèles font déjà l'objet de transactions commerciales. Il est par ailleurs clair que de nouvelles offres vont apparaître pour mettre à disposition des plateformes de développement LLM pour en simplifier le développement.

À l'opposé des LLM, la plupart des autres systèmes d'IA/ML requièrent beaucoup moins de capacité de calcul, la barrière se situe dans ce cas essentiellement au niveau de l'accès aux données en qualité et en quantité suffisante. Enfin ces inquiétudes sont également liées au manque d'explicabilité et donc de transparence des modèles statistiques (LLM ou pas).

11.7 L'accompagnement au changement dans les entreprises

« L'intelligence artificielle est un moyen d'amplifier l'intelligence humaine, de même que les machines sont un moyen d'amplifier la force physique », d'après Yann Le Cun (Chief AI Scientist for Meta). Mais malgré cette vision idyllique et même romantique finalement, l'équilibre homme-machine au sein des organisations qu'elles soient étatiques ou privées reste un enjeu crucial pour le développement durable et la stabilité sociale. Les organisations devront repenser leurs structures et leurs façons de travailler. La distribution du temps pris par les diverses tâches qui composent le travail se trouvera radicalement modifiée. Clairement, il sera possible de faire plus en moins de temps, ceci est la caractéristique même de l'automatisation. Le champ des compétences attendues dans un métier et les critères de performance vont changer. En définitive, pour rester maître de son destin professionnel, *« chacun d'entre nous doit se concentrer sur la manière dont son métier va évoluer. Il n'a jamais été*

aussi urgent d'acquérir de nouvelles compétences, de se former et de repenser son travail », d'après Tesdal Neeley (Professor à la Harvard Business School) (Vergera, 2023).

Il est important de développer des technologies qui respectent les valeurs humaines, de renforcer les compétences humaines pour travailler avec les technologies de l'IA, et d'établir des réglementations éthiques et sociales. La question de l'impact de l'IA sur le travail et l'économie demeure complexe et en constante évolution, il est donc important de continuer à suivre les développements et les recherches dans ce domaine. Il est peu probable qu'il y ait un quelconque moratoire sur le développement de l'IA, malgré toutes les questions légitimes. Dans ce contexte, il apparait urgent que les organisations et en particulier les entreprises se préparent aux mutations, dans le respect des législations qui ne manqueront pas de « fleurir » dans les années à venir.

L'accompagnement au changement est ici un élément clé pour assurer une transition réussie vers une organisation de travail évoluant dans un environnement saturé par l'IA. Comme toujours, la communication, la formation, la collaboration, la sensibilisation à l'éthique et à la réglementation ainsi que la planification à moyen terme sont autant d'éléments à prendre en compte pour assurer la stabilité du tissu social de l'entreprise et sans doute de la Société.

Dans une approche classique de la gestion du changement, les entreprises doivent fournir des informations claires et transparentes sur l'impact de l'IA sur le travail et l'organisation. Les équipes doivent également être formées pour acquérir les compétences nécessaires. De plus, il conviendra de les impliquer dans la mise en œuvre afin de générer un sentiment de propriété et de responsabilité envers le changement. Une approche structurée de la gestion du changement peut ainsi aider à minimiser les impacts négatifs sur le personnel. Il est nécessaire d'identifier les plus touchés par le changement et de mettre en place des mesures pour les soutenir.

11.8 Une meilleure maîtrise du développement des IA

Ces réflexions n'ont certainement pas échappé aux signataires de l'appel au moratoire sur l'intelligence artificielle en 2023 (Musk, 2023). Parmi les signataires nous retrouvons des noms aussi prestigieux qu'Elon Musk, Steve Wozniak, Yuval Noah Hariri et bien d'autres encore. La lettre des auteurs a été résumée à une demande de moratoire sur les LLM (les IA de type large langage modèle) mais en réalité elle va bien au-delà et incite à la réflexion, quel que soit les motivations des uns et des autres. Les auteurs demandent logiquement, la mise en place de systèmes robustes de gouvernance de l'IA incluant une nouvelle autorité réglementaire dédiée à cette technologie, la surveillance des systèmes d'IA, la mise en place de systèmes permettant de distinguer le réel du synthétique, la mise place de systèmes pour le suivi des fuites de modèles. Les signataires mettent l'accent sur la nécessaire création d'un écosystème complet et robuste d'audit et de certification. Ils pointent également la nécessité de faire évoluer le droit pour la clarification de la responsabilité pour les dommages causés par l'IA. Enfin, ils appellent les pouvoirs publics à mettre en place des financements pour la recherche technique sur la sécurité de l'IA et la création d'institutions dotées de ressources suffisantes pour faire face aux perturbations économiques et politiques dramatiques (en particulier pour la démocratie) que l'IA provoquera. Cette lettre contient finalement la feuille de route relative au déploiement des structures de gouvernance mondiale de l'IA, que nos dirigeants devraient implémenter dans les années à venir.

Pour finir, nous terminons ce travail en citant les mots prononcés en 2021 à la BBC par Brad Smith (PDG de Microsoft) au sujet de l'IA : *« Si nous ne promulguons pas les lois pour protéger le public, la technologie continuera d'avancer à toute allure, et il sera très difficile de la rattraper … Je me souviens constamment des leçons de George Orwell dans son livre 1984. L'histoire fondamentale… était celle d'un gouvernement qui pouvait voir tout ce que tout le monde faisait et entendre tout ce que tout le monde disait tout le temps. Eh bien, cela ne s'est pas produit en 1984, mais si nous ne faisons pas attention, cela pourrait arriver en 2024 »* (Louvet, 2021).

12 ANNEXE - Proposition de règlement du Parlement européen et du Conseil

ANNEXES
à la

Proposition de règlement du Parlement européen et du Conseil

ÉTABLISSANT DES RÈGLES HARMONISÉES CONCERNANT L'INTELLIGENCE ARTIFICIELLE (LÉGISLATION SUR L'INTELLIGENCE ARTIFICIELLE) ET MODIFIANT CERTAINS ACTES LÉGISLATIFS DE L'UNION

ANNEXE I
TECHNIQUES ET APPROCHES D'INTELLIGENCE ARTIFICIELLE
visées à l'article 3, point 1

(a) Approches d'apprentissage automatique, y compris d'apprentissage supervisé, non supervisé et par renforcement, utilisant une grande variété de méthodes, y compris l'apprentissage profond.

(b) Approches fondées sur la logique et les connaissances, y compris la représentation des connaissances, la programmation inductive (logique), les bases de connaissances, les moteurs d'inférence et de déduction, le raisonnement (symbolique) et les systèmes experts.

(c) Approches statistiques, estimation bayésienne, méthodes de recherche et d'optimisation.

13 ANNEXE – Typologie des malwares

Il existe différents types de malwares qu'il est possible de classer en différentes familles. Une classification possible des malwares est la suivante :

- **Les « Trojans »** : il s'agit d'exécutables qui apparaissent comme légitimes et sans danger. Ces programmes malveillants se cachent dans des programmes d'apparence inoffensive. Mais une fois que ces derniers sont lancés, ils exécutent des instructions malveillantes sur l'ordinateur hôte et peuvent par exemple ouvrir une porte d'accès dérobée.

- **Les « Botnets »** : ce sont des programmes dont l'objectif principal est de se propager sur autant d'hôtes que possible au sein d'un même réseau, dans le but, par exemple est de détourner leur capacité de calcul au service des attaquants.

- **Les virus** : un virus est un programme qui s'auto-reproduit en infectant d'autres programmes. Très populaires sous MS-DOS et Windows 3.1, ils sont en nette perte de vitesse comparés aux autres formes de malwares.

- **Les vers** : il s'agit de programmes malveillants qui utilisent internet sous toutes ses formes pour se propager. Ils utilisent des supports comme les mails, les sites web, les serveurs FTP, mais aussi le protocole NetBIOS, etc.

- **Les « Downloaders »** : ce sont des programmes qui une fois installés téléchargent des librairies ou du code au contenu malveillant à partir du réseau. Une fois téléchargé, le code malveillant est exécuté sur l'hôte infecté.

- **Les « RootKits »** : ces malwares sont conçus pour permettre aux pirates d'installer une série d'outils permettant d'accéder à distance aux ordinateurs infectés. Ils compromettent le plus souvent les hôtes au niveau de leur système d'exploitation. Ils prennent souvent la forme de « drivers » ce qui rend la lutte contre ces programmes très compliquée et souvent inefficace.

- **Les « Ransomwares »** : ce sont des programmes malveillants dont la fonction est de chiffrer les données présentes sur le disque dur des ordinateurs infectés. Les victimes sont alors contactées pour payer une rançon contre les clés de déchiffrement des disques durs infectés.

- **Les spywares** : il s'agit de logiciels espions qui permettent de connaître l'activité sur l'ordinateur infecté. On y retrouve deux grandes familles : les Keyloggers (les "Enregistreurs de touches" stockent dans un fichier souvent chiffré toutes les touches frappées sur le clavier) et les Adwares (qui modifient la page de démarrage du navigateur internet, ou installent un plug-in de recherche sur internet).

- **Les Droppers** : ce sont des fichiers exécutables qui paraissent anodins mais qui installent un ou plusieurs malwares.

- **Les Macros** : les malwares sous forme de macros s'insèrent dans des documents de type "suites bureautiques".

- **Les Flooders** : ils permettent d'inonder une cible (un site web par exemple) afin de l'empêcher de fonctionner. Partie intégrante des Denis de Service Distribués (DDoS), ces programmes passent souvent par une phase de contamination (installation sur le plus de machines possibles) avant la phase d'attaque proprement dite.

- **Les « APTs »** : il s'agit là de cyberattaques prolongées et ciblées visant des vulnérabilités spécifiques du parc d'ordinateurs installés, par lesquelles un pirate et/ou une organisation non autorisée accède au réseau d'une entreprise et passe inaperçue pendant une longue période. Une attaque APT vise généralement à surveiller l'activité réseau et à voler des données plutôt qu'à porter atteinte au réseau ou à l'organisation.

- **Les « Zero days »** : il s'agit de malwares qui exploitent des vulnérabilités qui n'ont pas encore été publiée par la communauté des chercheurs et analystes. Ces malwares ne sont par définition pas détectable et possède une très grande valeur pour les pirates ou certaines organisations gouvernementales.

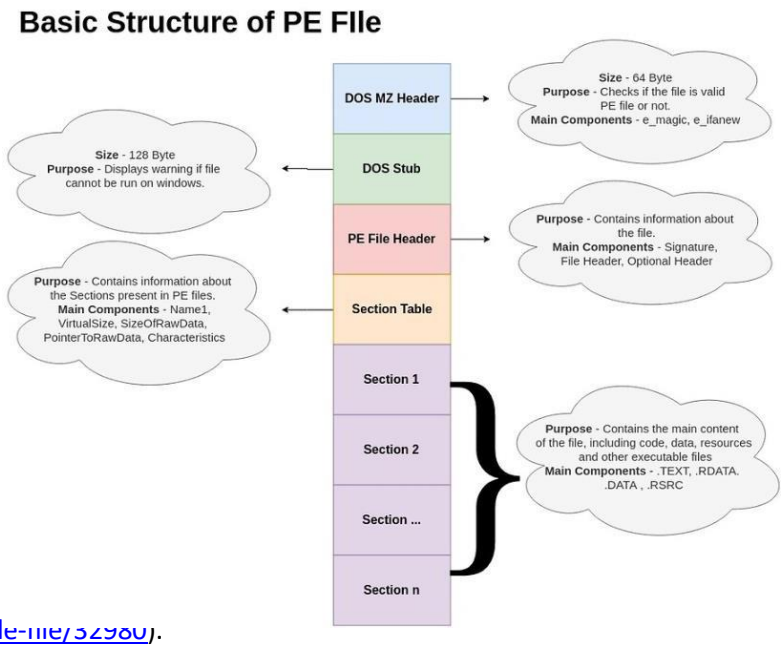
14 ANNEXE – Expérimentation : Apprentissage pour la détection statique de malwares, méthodologie et implémentation

14.1 Utilisation de l'IA/ML pour l'analyse statique de malwares

14.1.1 Analyse du PE file

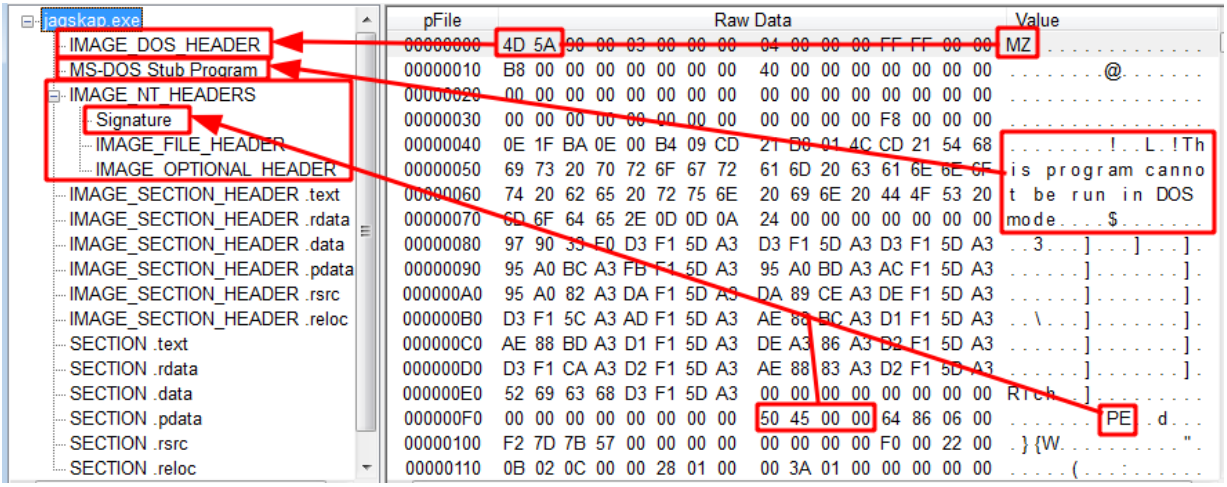
Si l'on se focalise sur les malwares ciblant spécifiquement le système d'exploitation Microsoft qui constitue près de 75 % du parc des ordinateurs installés (cf. site https://fr.statista.com/infographie/20455/parts-de-marche-des-systemes-exploitation-pour-ordinateurs-dans-le-monde/), chaque fichier exécutable (.exe, .dll, .sys) doit satisfaire aux spécifications contenues dans le format PE file **(PE pour Portable Executable)**.

Le loader Windows exécute les directives de chargement présentes dans les différentes sections du PE. À ce titre les sections PE sont une cible privilégiée pour les développeurs de malwares. Dans les faits, la partie PE d'un fichier exécutable peut être vue comme une sorte de « répertoire » contenant une variété d'objets binaires qui sont exécutés et/ou déchiffrés par le système et pouvant potentiellement l'infecter.



Pour simplifier, un « PE file » est constitué d'un « PE file header », d'une section table et d'une section de données. (cf. illustration ci-contre, source : https://malware.news/t/portable-executable-file/32980).

La détection des malwares se basent sur l'analyse de la structure et du contenu du PE. Nous présentons ci-dessous un exemple de PE tel qu'édité par l'outil PEView) en hexadécimal.



La signature hexadécimale « 50 45 00 00 » indique le début fichier PE.

14.1.2 Analyse via la séquence des n-grams

L'analyse statique des malwares (PE) repose classiquement sur une analyse du code à l'aide du calcul de la fréquence des n-grams les plus présents. Un n-grams est essentiellement une sous-séquence contiguë de n éléments d'une séquence donnée d'éléments. Pour dire les choses autrement, les n-grams sont des séquences de mots. Les mono-grams sont des mots uniques, bi-grams des séquences de deux mots, tri-grams de trois mots et ainsi de suite. Un modèle n-gram permet de modéliser chaque succession de mots par une probabilité. Une des premières applications des n-grams est la génération automatique de texte (Natanelic, 2020).

Dans le contexte de la détection et de la classification des logiciels malveillants, le terme n-grams fait référence aux n-grams de code d'octet ou aux n-grammes d'opcode. L'opcode est la partie du langage machine qui spécifie l'opération à effectuer.

La procédure de base consiste à disposer d'un corpus de programmes malveillants/bénins et à extraire les n-grams d'octets les plus courants, c'est-à-dire les séquences de n octets les plus courantes qui apparaissent dans tous les logiciels malveillants ou au moins dans une classe/famille particulière de logiciels malveillants. Il y a donc deux variables dans cette approche, la taille de la fenêtre n et le nombre de ces séquences fréquentes utilisées pour profiler la classe, appelées longueur de profil. Dans les faits, utiliser les 200 « 2-grams » les plus fréquents sont suffisants pour la détection de malwares via des algorithmes d'IA/ML.

14.1.3 Les algorithmes d'apprentissage pour la détection statique

Dans cette expérimentation, nous nous focaliserons sur trois implémentations différentes d'arbres de décision améliorées par les techniques d'échantillonnage :

- Le XGBClassifier :
  - e**X**treme **G**radient **B**oosting est une implémentation open source optimisée de l'algorithme d'arbres de boosting de gradient. Le Boosting de Gradient est un algorithme d'apprentissage supervisé dont le principe est de combiner les résultats d'un ensemble de modèles plus simple et plus faible afin de fournir une meilleure prédiction. Il s'agit dans les faits de générer plusieurs arbres de décision et de les combiner. L'algorithme construit un premier modèle qu'il va bien sur évaluer. À partir de cette première évaluation, chaque « individu » va être alors pondéré en fonction de la performance de la prédiction

- Les RandomForest
  - Il s'agit d'un algorithme qui se base sur l'assemblage d'arbres de décision indépendants. Le RandomForest est donc composé de plusieurs arbres de décision, entrainés de manière indépendante sur des sous-ensembles de l'ensemble d'apprentissage (bagging). Chacun produit donc une estimation, et c'est la combinaison des résultats qui va donner la prédiction finale. Concrètement, l'algorithme procède selon la méthode du « bagging » qui consiste à réaliser un échantillonnage avec remise des données et entraîner l'algorithme de façon séparée sur chacun de ces échantillons et à la fin assembler les résultats des modèles obtenus.

- Le AdaBoost Classifier
  - Il s'agit de l'un des premiers modèles de « boosting » basé sur la génération d'arbres de décision. L'algorithme s'adapte et tente de s'autocorriger à chaque itération du processus de boosting. AdaBoost donne initialement le même poids à chaque ensemble de données. Ensuite, l'algorithme ajuste automatiquement les poids des

points de données après chaque itération. Les poids des éléments mal classés sont augmentés afin de les corriger pour la prochaine itération. Le processus est répété jusqu'à ce que l'erreur résiduelle, ou la différence entre les valeurs réelles et prévues, tombe sous un seuil acceptable.

14.2 Expérimentation pour l'implémentation d'un détecteur statique de malwares

Nous avons utilisé les trois algorithmes précédemment décrits afin de faire une évaluation des méthodes classiques d'IA/ML pour la détection statique de malwares. Nous présentons dans cette annexe les résultats et le code python associé à cette expérience. Notre but est d'évaluer le niveau de difficulté pour arriver à implémenter une solution de type IA/ML. Nous avons pris cet exemple car il est « assez simple », bien documenté et qu'il ne nécessite pas de grande capacité de calcul.

Nous avons utilisé la base de données REWEMA (Retrieval of 32-bit Windows Architecture Executables Applied to Malware Analysis). Cette base de données a été réalisée à des fins de recherche par Sidney M. L. de Lima et ses collaborateurs de l'Université de Pernambuco au Brésil.

Elle est librement disponible sur le site https://github.com/rewema/rewema. Elle présente l'intérêt de contenir 3136 exécutables malveillants et 3136 autres exécutables bénins choisis avec soin afin de couvrir l'ensemble du spectre des malwares. Ainsi, elle contient un panel exhaustif de malwares tels que des « Trojans », des « RootKits », des virus, des spywares, des botnets, des Flooder, des programmes de DoS, etc. Nous utilisons donc une base de données de faible volumétrie à usage académique suffisante pour notre étude. Pour la phase d'apprentissage nous utilisons 70 % des données, les 30 % restants le sont pour l'évaluation des algorithmes (phase de test).

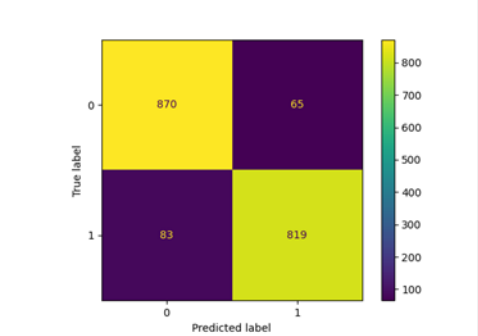
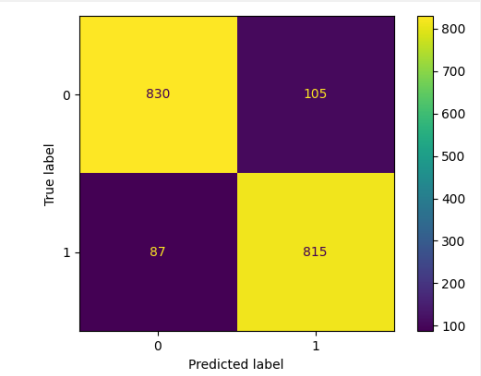
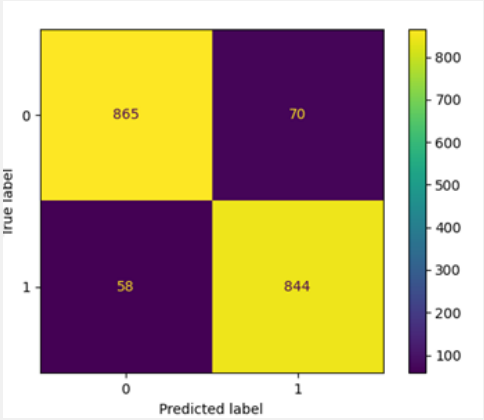
À titre de comparaison, en 2020, l'équipe Sophos AI a annoncé la mise à disposition de SOREL-20M (Sophos-ReversingLabs - 20 millions) un ensemble de données contenant les métadonnées, les étiquettes et les fonctionnalités de 20 millions de fichiers Windows Portable Executable (PE), dont 10 millions d'échantillons de logiciels malveillants désactivés (Harang, 2020). La base de données de malwares est disponible librement en revanche les fichiers « bénins » ne le sont pas. La taille de la base de données est approximativement de 8 TB et est disponible depuis le stockage objet Amazon suivant :

s3://sorel-20m/09-DEC-2020/baselines/checkpoints/FFNN/. Les informations sur cette base sont disponibles sur : https://github.com/sophos/SOREL-20M#a-note-on-dataset-size.

Les tests ont été menés sur une machine virtuelle linux (Ubuntu 22.04) de 8 Gb de RAM et 40 Go de disque dur fonctionnant sur un ordinateur portable.

14.3 Évaluation des performances des classificateurs

Algorithme	Performance	Commentaire
AdaBoostClassifier(n_estimators=250, Random_state=0)	Niveau d'exactitude (accuracy) : 91,94 % Matrice de confusion :	La méthode classe correctement approximativement 92 % des binaires dans un système à 2 états (malwares/bénins) L'ensemble des fichiers bénins de test comprend 935 exemples L'ensemble des fichiers malwares comprend 902 exemples

Algorithme	Performance	Commentaire
		La matrice de confusion montre que sur 902 malwares évalués 819 sont classés correctement et 83 sont classés comme de faux négatifs Sur les 935 fichiers bénins. 870 sont classés correctement et 65 comme faux positifs (malwares)
RandomForestClassifier(n_estimator=100)	Niveau d'exactitude (accuracy) : 89,55 % Matrice de confusion : 	La matrice de confusion montre que sur 902 malwares évalués 815 sont classés correctement et 87 sont classés comme de faux négatifs Sur les 935 fichiers bénins. 830 sont classés correctement et 105 comme faux positifs (malwares)
XGBClassifier (paramètres par défaut)	Niveau d'exactitude (accuracy) : 93,03 % Matrice de confusion : 	La matrice de confusion montre que sur 902 malwares évalués 844 sont classés correctement et 58 sont classés comme de faux négatifs Sur les 935 fichiers bénins. 865 sont classés correctement et 70 comme faux positifs (malwares) C'est l'algorithme le plus performant sur les trois évalués

Nous pouvons constater qu'il est assez « simple » de développer un système basé sur l'IA/ML afin de faire de la détection statique de malware. Il n'y a pas de difficultés particulières pour aboutir à des algorithmes détectant aux alentours de 90-93 % des menaces. Ceci signifie également que certains malwares sont capables d'échapper au système. Bien évidemment, il ne s'agit pas là d'un antimalware exhibant les performances d'un système commercial entrainé sur des centaines de milliers d'exemples. Cette expérience illustre bien que la barrière ne se trouve pas au niveau des algorithmes d'IA/ML mais au niveau des données et de la puissance de calcul pour la phase d'apprentissage.

14.4 Présentation du code utilisé

Le script python ci-dessous permet de construire un modèle de détection statique de malwares via les algorithmes suivants Random Forrest / AdaBoost /XGB classifier. Ce script est une adaptation de la méthode proposée dans le livre d'Emmanuel Tsukerman intitulé « Machine Learning for Cybersecurity – Cookbook » publié aux éditions PACKT en 2019.

```
import pefile
import os
from os import listdir

from sklearn.model_selection import train_test_split

#import collection
import collections
from nltk import ngrams
import numpy as np
import pefile

from sklearn.feature_extraction.text import HashingVectorizer, TfidfTransformer
from sklearn.pipeline import Pipeline

from scipy.sparse import hstack, csr_matrix

from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
from sklearn.svm import SVC

from sklearn.ensemble import AdaBoostClassifier

from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
import matplotlib.pyplot as plt

#
# 0. Defining some convenient functions
#

def read_file(file_path):
    """Reads in the binary sequence of a binary file."""
    with open(file_path, "rb") as binary_file:
        data = binary_file.read()
    return data

def byte_sequence_to_Ngrams(byte_sequence, N):
    """Creates a list of N-grams from a byte sequence."""
    Ngrams = ngrams(byte_sequence, N)
    return list(Ngrams)

def binary_file_to_Ngram_counts(file, N):
    """Takes a binary file and outputs the N-grams counts of its binary sequence"""
```

```
    filebyte_sequence = read_file(file)
    file_Ngrams = byte_sequence_to_Ngrams(filebyte_sequence, N)
    return collections.Counter(file_Ngrams)

def get_Ngram_features_from_sample(sample, K1_most_frequent_Ngrams_list):
    """Takes a sample and produce a feature vector.
    The features are the counts of the K1 N-grams we have selected.
    """
    K1 = len(K1_most_frequent_Ngrams_list)
    feature_vector = K1* [0]
    file_Ngrams = binary_file_to_Ngram_counts(sample, N)
    for i in range(K1):
        feature_vector[i] = file_Ngrams[K1_most_frequent_Ngrams_list[i]]
    return feature_vector


def preprocess_imports(list_of_DLLs):
    """Normalize naming of the imports of a PE file."""
    temp =  [x.decode().split(".")[0].lower() for x in list_of_DLLs]
    return " ".join(temp)

def get_imports(pe):
    """Get a list of the imports of a PE file"""
    list_of_imports = []
    for entry in pe.DIRECTORY_ENTRY_IMPORT:
        list_of_imports.append(entry.dll)
    return preprocess_imports(list_of_imports)

def get_section_names(pe):
    """Gets a list of section names from a PE file"""
    list_of_section_names = []
    for sec in pe.sections:
        normalized_name = sec.Name.decode().replace("x00", "").lower()
        list_of_section_names.append(normalized_name)
    return " ".join(list_of_section_names)

#
# 1. Begin by enumerating our samples and assigning their labels
#

directories_with_labels = [("Benigns/Benigns", 0), ("Malwares/Malwares", 1)]

list_of_samples = []
labels =[]
for data_set_path, label in directories_with_labels:
    samples = [f for f in listdir(data_set_path)]
    for sample in samples:
        file_path = os.path.join(data_set_path, sample)
        list_of_samples.append(file_path)
        labels.append(label)

#
```

```
# 2. Performing a stratified train-test split
#

samples_train, samples_test, labels_train, labels_test = train_test_split(
   list_of_samples, labels, test_size = 0.3, stratify = labels, random_state = 11
   )

#
# 3. We select the 200 most frequent 2-grams as our features
#

N = 2
Ngram_counts_all = collections.Counter([])
for sample in samples_train:
   Ngram_counts_all += binary_file_to_Ngram_counts(sample, N)

K1 = 200
K1_most_frequent_Ngrams = Ngram_counts_all.most_common(K1)
K1_most_frequent_Ngrams_list = [x[0] for x in K1_most_frequent_Ngrams]
print(K1_most_frequent_Ngrams_list)
#
# 4. We extract the N-gram counts, section names, imports and number of sections of
# each sample in our training test and skip over samples whose PE header cannot be parsed[2]

imports_corpus_train = []
num_sections_train = []
section_names_train = []
Ngram_features_list_train = []
y_train = []

for i in range(len(samples_train)):
   sample = samples_train[i]
   try:
      Ngram_features = get_Ngram_features_from_sample(sample, K1_most_frequent_Ngrams_list)
      pe = pefile.PE(sample)
      imports = get_imports(pe)
      n_sections = len(pe.sections)
      sec_names = get_section_names(pe)
      imports_corpus_train.append(imports)
      num_sections_train.append(n_sections)
      section_names_train.append(sec_names)
      Ngram_features_list_train.append(Ngram_features)
      y_train.append(labels_train[i])
   except Exception as e:
      print(sample + ":")
      print(e)

#
# 5.  We use a hashing vectorizer followed by tfidf to convert the imports # and section names,
# both of which a text features into a numerical form

imports_featurizer = Pipeline(
```

```
    [("vect", HashingVectorizer(input="content", ngram_range=(1,2))),
     ("tfidf", TfidfTransformer(use_idf=True))]
    )
section_names_featurizer = Pipeline(
    [("vect", HashingVectorizer(input="content", ngram_range=(1,2))),
     ("tfidf", TfidfTransformer(use_idf=True))]
    )
imports_corpus_train_transformed = imports_featurizer.fit_transform(imports_corpus_train)
section_names_train_transformed = section_names_featurizer.fit_transform(section_names_train)


#
# 6. We combine the vectorized features into a single array
#

x_train = hstack(
    [
        Ngram_features_list_train,
        imports_corpus_train_transformed,
        section_names_train_transformed,
        csr_matrix(num_sections_train).transpose()
    ]
    )

#
# 7. We train a Random Forrest / AdaBoost /XGB classifier on the training
# set and print out its score
#
#clf = SVC(random_state=0)
clf=XGBClassifier()
#clf = RandomForestClassifier(n_estimators=100)
#clf = AdaBoostClassifier(n_estimators=250, random_state=0)
clf = clf.fit(x_train, y_train)

#
# 8. We collect the features of the testing set, just as we did for the
# training set
#
imports_corpus_test = []
num_sections_test = []
section_names_test = []
Ngram_features_list_test = []
y_test = []

for i in range(len(samples_test)):
    sample = samples_test[i]
    try:
        Ngram_features = get_Ngram_features_from_sample(sample, K1_most_frequent_Ngrams_list)
        pe = pefile.PE(sample)
        imports = get_imports(pe)
        n_sections = len(pe.sections)
        sec_names = get_section_names(pe)
```

```
        imports_corpus_test.append(imports)
        num_sections_test.append(n_sections)
        section_names_test.append(sec_names)
        Ngram_features_list_test.append(Ngram_features)
        y_test.append(labels_test[i])
    except Exception as e:
        print(sample + ":")
        print(e)

#
# 9. We apply the previously trained transformer to vectorize the text features and test the classifier
# on the resulting test set


imports_corpus_test_transformed = imports_featurizer.transform(imports_corpus_test)
section_names_test_transformed = section_names_featurizer.fit_transform(section_names_test)

x_test = hstack(
    [Ngram_features_list_test,
    imports_corpus_test_transformed,
    section_names_test_transformed,
     csr_matrix(num_sections_test).transpose()]
    )

#SVC(random_state=0)
print(clf.score(x_test, y_test))
ConfusionMatrixDisplay.from_estimator(clf, x_test, y_test)
plt.show()
```

15 BIBLIOGRAPHIE

A Joint Report by UNICRI and UNCCT, 2021. *Countering Terrorism Online with Artificial Intelligence - An Overview for Law Enforcement and Counter-Terrorism Agencies in South Asia and South-East Asia.* [En ligne]
Disponible sur: https://unicri.it/Publications/Countering-Terrorism-Online-with-Artificial-Intelligence-%20SouthAsia-South-EastAsia
[Accès le 12 12 2022].

Afifi-Sabet, K., 2021. *Microsoft launches open source tool Counterfeit to prevent AI hacking.* [En ligne]
Disponible sur: https://www.itpro.co.uk/technology/artificial-intelligence-ai/359409/microsoft-open-source-counterfit-to-stop-ai-hacks
[Accès le 17 01 2022].

AFP, 2023. *Entre IA "harceleuse" et chatGPT, comment réguler un secteur en pleine mutation.* [En ligne]
Disponible sur: https://www.msn.com/fr-fr/actualite/technologie-et-sciences/entre-ia-harceleuse-et-chatgpt-comment-réguler-un-secteur-en-pleine-mutation/ar-AA17ophU
[Accès le 08 04 2023].

Akash, S., 2022. *Top 10 Cybersecurity Companies Using AI to the Fullest in 2022.* [En ligne]
Disponible sur: https://www.analyticsinsight.net/top-10-cybersecurity-companies-using-ai-to-the-fullest-in-2022/
[Accès le 01 12 2022].

Ambassade de France aux Etats - Unis , 2022. *Introduction d'un projet de régulation des algorithmes d'IA à l'échelle fédérale.* [En ligne]
Disponible sur: https://france-science.com/introduction-dun-projet-de-regulation-des-algorithmes-dia-a-lechelle-federale/
[Accès le 06 05 2023].

ANSSI, s.d. *Glossaire.* [En ligne] Disponible sur: https://www.ssi.gouv.fr/entreprise/glossaire/c/

Anthony, 2023. *Les États-Unis commencent à étudier la possibilité de réglementer les systèmes d'IA comme ChatGPT.* [En ligne]
Disponible sur: https://intelligence-artificielle.developpez.com/actu/343430/Les-Etats-Unis-commencent-a-etudier-la-possibilite-de-reglementer-les-systemes-d-IA-comme-ChatGPT-pour-garantir-que-ces-technologies-soient-legales-efficaces-ethiques-sures-et-dignes-de-confianc
[Accès le 19 04 2023].

Arnoux, C., 2019. *Vade Secure nommé dans l'indice Next40 des start-ups françaises les plus prometteuses.* [En ligne]
Disponible sur: https://www.vadesecure.com/fr/blog/vade-secure-nomme-dans-lindice-next40-des-start-ups-francaises-les-plus-prometteuses
[Accès le 25 12 2022].

Asan, S., 2021. *INTERNET NOUS ENFERME-T-IL DANS UNE BULLE ?.* [En ligne]
Disponible sur: https://cyberjustice.blog/2021/01/13/internet-nous-enferme-t-il-dans-une-bulle/
[Accès le 22 11 2022].

Ashish Vaswani, N. et coll.., 2017. Attention is all you need. *ArXiv,* 6 12, arxiv(arxiv), p. 15.

Barguisseau, L., 2019. *L'intelligence artificielle pour lutter contre les fraudes bancaires.* [En ligne]
Disponible sur: https://www.credigo.fr/actualites/intelligence-artificielle-pour-lutter-contre-fraudes-bancaires.html
[Accès le 11 02 2023].

Barilari, A., 2018. Une taxe sur les robots est-elle un concept d'avenir ?. *Gestion & Finances Publiques*, 01 01, p. 48 à 52.

Barraud, B., 2021 . *Une IA responsable et digne de confiance.* [En ligne]
Disponible sur: https://hal-univ-artois.archives-ouvertes.fr/hal-03659289v1/file/BBARRA~1.PDF
[Accès le 11 04 2023].

Berly, A., Manaouil, C. & Dervaux, A., 2022. *L'intelligence artificielle peut-elle aider à estimer le risque de récidive dans les comportements violents ?.* [En ligne]
Disponible sur: https://hal.science/hal-03491221/document
[Accès le 12 12 2022].

Berthier, T. e., 2022. *"Les usages malveillants de l'intelligence artificielle au service de la cybercriminalité".* [En ligne]
Disponible sur: https://www.areion24.news/2022/04/21/les-usages-malveillants-de-lintelligence-artificielle-au-service-de-la-cybercriminalite/2/
[Accès le 25 04 2023].

Bettex, L., 2018. *L'intelligence artificielle intégrée dans les plates-formes Cisco.* [En ligne]
Disponible sur: https://gblogs.cisco.com/ch-fr/2018/10/30/lintelligence-artificielle-integree-dans-les-plates-formes-cisco/
[Accès le 28 12 2022].

BlackBerry Limited, 2018. *BlackBerry to Acquire Cylance and Add Premier AI and Cybersecurity Capabilities.* [En ligne]
Disponible sur: https://www.blackberry.com/us/en/company/newsroom/press-releases/2018/blackberry-acquisition-press-release
[Accès le 27 12 2022].

Bocoum & Briot Juristes , 2021. *LA CHINE ADOPTE À SON TOUR UNE LOI SUR LA PROTECTION DES DONNÉES.* [En ligne]
Disponible sur: https://www.vill age-justice.com/articles/chine-adopte-son-tour-une-loi-sur-protection-des-donnees,40090.html
[Accès le 27 04 2023].

Bonfils, G., 2023. *Protection des données personnelles : vers un RGPD américain ?.* [En ligne]
Disponible sur: https://incyber.org/protection-donnees-personnelles-vers-rgpd-americain/
[Accès le 06 05 2023].

Bouaziz, D., 2021. *Le risque cyber évalué à 6 000 milliards de dollars en 2021.* [En ligne]
Disponible sur: https://www.ecommercemag.fr/Thematique/paiements-1291/barometre-etude-2187/Breves/etude-risque-cyber-evalue-000-milliards-dollars-2021-360132.htm
[Accès le 22 11 2022].

Bourany, T., 2018. *Les 5V du big data.* [En ligne]
Disponible sur: https://www.cairn.info/revue-regards-croises-sur-l-economie-2018-2-page-27.htm
[Accès le 20 02 2023].

Bourany, T., 2018. *Les 5V du big data.* [En ligne]
Disponible sur: https://www.cairn.info/revue-regards-croises-sur-l-economie-2018-2-page-27.htm
[Accès le 20 02 2023].

Brasseur, C., 2015. *Usages visuels des données & Big data.* [En ligne]
Disponible sur: https://www.cairn.info/revue-i2d-information-donnees-et-documents-2015-2-page-44.htm
[Accès le 20 02 2023].

Brewster, T., 2021. *AI voice cloning is used in a huge heist being investigated by Dubai investigators, amidst warnings about cybercriminal use of the new technology..* [En ligne]
Disponible sur: https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=747334297559
[Accès le 13 02 2023].

Bruno, 2022. *Un projet de loi US sur la confidentialité des données donnerait plus de contrôle sur les informations personnelles collectées.* [En ligne]
Disponible sur: https://www.developpez.com/actu/336011/Un-projet-de-loi-US-sur-la-confidentialite-des-donnees-donnerait-plus-de-controle-sur-les-informations-personnelles-collectees-4-ans-apres-le-RGPD/
[Accès le 13 04 2023].

Business Wire, s.d. *Cylance Customers Propel Company Past $100M Revenue.* [En ligne]
Disponible sur: https://www.businesswire.com/news/home/20180129005200/en/Cylance-Customers-Propel-Company-Past-100M-Revenue
[Accès le 17 12 2022].

Buyse, N., 2021. *Le rêve américain contrarié de la jeune pousse lilloise Vade Secure.* [En ligne]
Disponible sur: https://www.lesechos.fr/tech-medias/hightech/le-reve-americain-contrarie-de-la-jeune-pousse-lilloise-vade-secure-1341199
[Accès le 25 12 2022].

Calvi, T., 2022. *Où en est la réglementation de l'intelligence artificielle aux USA ?.* [En ligne]
Disponible sur: https://www.actuia.com/actualite/ou-en-est-la-reglementation-de-lintelligence-artificielle-aux-usa/
[Accès le 06 05 2023].

Challand, R., 2022. *En Chine, l'intelligence artificielle Tang Yu est devenue PDG.* [En ligne]
Disponible sur: https://www.lesnumeriques.com/vie-du-net/en-chine-l-intelligence-artificielle-tang-yu-est-devenue-pdg-n192725.html
[Accès le 11 12 2022].

Challand, R., 2023. *En pleine euphorie ChatGPT, la Chine veut encadrer l'intelligence artificielle.* [En ligne]
Disponible sur: https://www.lesnumeriques.com/intelligence-artificielle/en-pleine-euphorie-chatgpt-la-chine-veut-encadrer-l-intelligence-artificielle-n208825.html
[Accès le 03 05 2023].

Check Point Press Releases, 2022. *Check Point Software Brings Faster, AI-Enabled Network Security and Advanced Threat Prevention for On-Premise, Cloud and IoT.* [En ligne]

Disponible sur: https://www.checkpoint.com/press-releases/check-point-software-brings-faster-ai-enabled-network-security-and-advanced-threat-prevention-for-on-premise-cloud-and-iot/
[Accès le 27 12 2022].

Cherif, A., 2023. *ETATS-UNIS: UNE PLAINTE DÉPOSÉE CONTRE CHATGPT AUPRÈS DE L'AUTORITÉ DE LA CONCURRENCE.* [En ligne]
Disponible sur: https://www.bfmtv.com/tech/etats-unis-une-plainte-deposee-contre-chat-gpt-aupres-de-l-autorite-de-la-concurrence_AN-202303300530.html#:~:text=Suivant-,Etats%2DUnis%3A%20une%20plainte%20d%C3%A9pos%C3%A9e%20contre%20ChatGPT%20aupr%C3%A8s%20de,l'autorit%C3%A
[Accès le 09 04 2023].

Chowdhry, A., 2018. *How SAP NS2 Provides The U.S. Government With Technology Solutions.* [En ligne]
Disponible sur: https://www.forbes.com/sites/amitchowdhry/2018/03/13/how-sap-ns2-provides-the-u-s-government-with-technology-solutions/?sh=7a77cf9373e6
[Accès le 24 12 2022].

Cisco, 2013. *Cisco Completes its Acquisition of Cognitive Securit.* [En ligne]
Disponible sur: https://www.cisco.com/c/en/us/about/corporate-strategy-office/acquisitions/cognitivesecurity.html
[Accès le 28 12 2022].

CNIL, 2022. *IA : comment être en conformité avec le RGPD ?.* [En ligne]
Disponible sur: https://www.entreprises.gouv.fr/fr/numerique/enjeux/promouvoir-modele-d-ia-ethique
[Accès le 11 03 2023].

CNIL, 2023. *La France ratifie la Convention 108+ du Conseil de l'Europe.* [En ligne]
Disponible sur: https://www.cnil.fr/fr/la-france-ratifie-la-convention-108-du-conseil-de-leurope
[Accès le 11 04 2023].

CNIL, L. d. d. N. d. l., 2022. *Sécurité des systèmes d'IA.* [En ligne]
Disponible sur: https://linc.cnil.fr/sites/default/files/atoms/files/linc_cnil_dossier-securite-systemes-ia.pdf
[Accès le 10 04 2023].

CNIL, S. d. l., 2020. *https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles.* [En ligne]
Disponible sur: https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles
[Accès le 10 04 2020].

Coëffé, T., 2021. *La Chine adopte la loi PIPL, l'équivalent du RGPD pour protéger les données personnelles.* [En ligne]
Disponible sur: https://www.blogdumoderateur.com/chine-pipl-rgpd/
[Accès le 03 05 2023].

Cohen, Avocate, D., 2023. *CHATGPT ET RGPD : LA PROTECTION DES DONNÉES PERSONNELLES.* [En ligne]
Disponible sur: https://www.village-justice.com/articles/chatgpt-rgpd-protection-des-donnees-personnelles,45280.html
[Accès le 20 04 2023].

Commission européenne , 2021. *Proposition RÈGLEMENT DU PARLEMENT EUROPÉEN ET DU CONSEIL, ÉTABLISSANT DES RÈGLES HARMONISÉES CONCERNANT L'INTELLIGENCE ARTIFICIELLE (LÉGISLATION SUR L'INTELLIGENCE ARTIFICIELLE) ET MODIFIANT CERTAINS ACTES LÉGISLATIFS DE L'UNION.* [En ligne]
Disponible sur: https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A52021PC0206
[Accès le 20 02 2023].

Conseil de l'europe , 2023. *Le « Privacy Symposium » examine le rôle mondial de la convention pour la protection des données ; l'Argentine ratifie le protocole d'amendement.* [En ligne]
Disponible sur: https://www.coe.int/fr/web/portal/-/le-%C2%AB-privacy-symposium-%C2%BB-examine-le-r%C3%B4le-mondial-de-la-convention-pour-la-protection-des-donn%C3%A9es-l-argentine-ratifie-le-protocole-d-amendement#:~:text=L'%C3%A9dition%202023%20du%20colloque,la%20%C2%A
[Accès le 03 05 2023].

Conseil de l'UE , 2022. *Législation sur l'intelligence artificielle: le Conseil appelle à promouvoir une IA sûre et respectueuse des droits fondamentaux.* [En ligne]
Disponible sur: https://www.consilium.europa.eu/fr/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights/
[Accès le 20 02 2023].

Conseil d'Etat , 2014. *Le numérique et les droits fondamentaux.* [En ligne]
Disponible sur: https://www.conseil-etat.fr/publications-colloques/etudes/le-numerique-et-les-droits-fondamentaux
[Accès le 20 04 2023].

Coret, S., 2022. *Les dépenses en intelligence artificielle aux États-Unis atteindront 120 milliards de dollars d'ici 2025.* [En ligne]
Disponible sur: https://intelligence-artificielle.developpez.com/actu/332056/Les-depenses-en-intelligence-artificielle-aux-Etats-Unis-atteindront-120-milliards-de-dollars-d-ici-2025-avec-un-taux-de-croissance-annuel-compose-TCAC-de-26-pourcent-sur-la-periode-de-prevision
[Accès le 21 11 2022].

Council of europe , 2018. *Intelligence artificielle et protection des données : enjeux et solutions possibles.* [En ligne]
Disponible sur: https://rm.coe.int/rapport-sur-l-intelligence-artificielle/16809020ef
[Accès le 20 02 2023].

Creemers, R. & Webster, G., 2021. *Translation: Personal Information Protection Law of the People's Republic of China – Effective Nov. 1, 2021.* [En ligne]
Disponible sur: https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/
[Accès le 12 04 2023].

Crichton (Dalloz), 2023. *Projet de règlement sur l'IA (II) : une approche fondée sur les risques.* [En ligne]
Disponible sur: https://www.dalloz-actualite.fr/flash/projet-de-reglement-sur-l-ia-ii-une-approche-fondee-sur-risques
[Accès le 25 04 2023].

CrowdStrike, 2020. *MICRO FOCUS INTERSET UEBA: UNKNOWN THREAT DETECTION - Combining endpoint data with user and entity behavioral analytics (UEBA) to swiftly reveal hidden threats.* [En

ligne]
Disponible sur: https://www.crowdstrike.com/wp-content/uploads/2020/08/micro-focus-inteset-datasheet.pdf
[Accès le 23 12 2022].

Crunchbase, s.d. *Tophant.* [En ligne]
Disponible sur: https://www.crunchbase.com/organization/tophant
[Accès le 29 12 2022].

Cynet, 2022. *Comprehensive Cybersecurity Made Easy.* [En ligne]
Disponible sur: https://www.cynet.com

Darktrace, s.d. *Darktrace Cyber Centre de recherche en IA.* [En ligne]
Disponible sur: https://fr.darktrace.com/recherche
[Accès le 23 12 2022].

DCSSI, 2004. *La défense en profondeur appliquée aux systèmes d'information.* [En ligne]
Disponible sur: https://www.ssi.gouv.fr/uploads/IMG/pdf/mementodep-v1-1.pdf
[Accès le 21 11 2022].

De Montes, D., 2018. *Comment gérer les Faux-Positifs dans un SOC.* [En ligne]
Disponible sur: https://www.idna.fr/2018/11/06/comment-gerer-les-faux-positifs-dans-un-soc/
[Accès le 21 11 2022].

Deleporte Avocat, B., 2021. *La Chine adopte sa loi sur la protection des données personnelles.* [En ligne]
Disponible sur:
https://www.lagbd.org/La_Chine_adopte_sa_loi_sur_la_protection_des_donnees_personnelles
[Accès le 03 05 2023].

Demichelis, R., 2019. *Les 10 recommandations de l'OCDE pour l'intelligence artificielle.* [En ligne]
Disponible sur: https://www.lesechos.fr/tech-medias/intelligence-artificielle/les-10-recommandations-de-locde-pour-lintelligence-artificielle-1023062
[Accès le 10 04 2023].

Denning, D. E., 1987. An Intrusion-Detection Model.. *IEEE Transactions on Software Engineering,* 13(2), pp. 222-232.

Dickson, B., 2020. *How machine learning removes spam from your inbox.* [En ligne]
Disponible sur: https://bdtechtalks.com/2020/11/30/machine-learning-spam-detection/#:~:text=Spam%20detection%20is%20a%20supervised,data%20sets%20of%20labeled%20emails
[Accès le 24 01 2023].

Dimeglio, Avocat , A., 2023. *[VIDÉO] L'INTERDICTION DE CHATGPT EN ITALIE..* [En ligne]
Disponible sur: https://www.village-justice.com/articles/interdiction-chatgpt-italie,45811.html
[Accès le 10 04 2023].

Donnat., F., 2019. *L'intelligence artificielle, un danger pour la vie privée ?* [En ligne]
Disponible sur : https://www.cairn.info/revue-pouvoirs-2019-3-page-95.htm?ref=doi
[Accès le 11/04/2023]

Dorigny, M., 2022. *Qu'est ce que la défense en profondeur ?.* [En ligne]
Disponible sur: https://www.it-connect.fr/cybersecurite-defense-en-profondeur/
[Accès le 21 11 2022].

EPSI, 2021. *Cybersécurité : l'IA va-t-elle prendre la place de l'humain à l'horizon 2030 ?.* [En ligne]
Disponible sur: https://www.epsi.fr/ia-remplacer-humain-2030
[Accès le 03 05 2023].

Europol, 2023. *Un rapport d'Europol explore le côté obscur de ChatGPT et autres LLM.* [En ligne]
Disponible sur: https://www.lexisveille.fr/un-rapport-deuropol-explore-le-cote-obscur-de-chatgpt-et-autres-llm
[Accès le 20 04 2023].

Fabrion, M., 2022. *Pékin oblige les géants chinois de la tech à lever le voile sur leurs algorithmes.* [En ligne]
Disponible sur: https://www.lesnumeriques.com/pro/pekin-oblige-les-geants-chinois-de-la-tech-a-lever-le-voile-sur-leurs-algorithmes-n189523.html
[Accès le 05 05 2023].

Fabrion, M., 2023. *IA : Elon Musk lance sa start-up pour créer un rival de ChatGPT.* [En ligne]
Disponible sur: https://www.lesnumeriques.com/intelligence-artificielle/ia-elon-musk-lance-sa-start-up-pour-creer-un-rival-de-chatgpt-n208942.html
[Accès le 2 05 2023].

Fasinou, B., 2021. *La Chine propose un contrôle strict des algorithmes en vue de réglementer davantage les services Internet.* [En ligne]
Disponible sur: https://www.developpez.com/actu/317967/La-Chine-propose-un-controle-strict-des-algorithmes-en-vue-de-reglementer-davantage-les-services-Internet-le-pays-veut-interdire-les-algorithmes-qui-creent-l
[Accès le 03 05 2023].

Fernandez-Toro, A., 2018. *Management de la sécurité de l'information.* s.l.:Eyrolles.
Fiche Produit FireEye, 2021. *FireEye Network Security.* [En ligne]
Disponible sur: https://www.fireeye.fr/content/dam/fireeye-www/regional/fr_FR/products/pdfs/fireeye-network-threat-prevention-platform.pdf
[Accès le 26 12 2022].

Figaro avec AFP , 2023. *Intelligence artificielle : la Maison Blanche rappelle les entreprises à leur devoir "moral".* [En ligne]
Disponible sur: https://www.lefigaro.fr/secteur/high-tech/intelligence-artificielle-la-maison-blanche-rappelle-les-entreprises-a-leur-devoir-moral-20230504
[Accès le 04 05 2023].

Fortinet, 2020. *Fortinet présente son appliance pour une détection ultrarapide des menaces basée sur l'auto-apprentissage et l'intelligence artificielle.* [En ligne]
Disponible sur: https://www.fortinet.com/fr/corporate/about-us/newsroom/press-releases/2020/introduces-self-learning-artificial-intelligence-appliance-for-sub-2nd-threat-detection
[Accès le 27 12 2022].

France 5, C dans l'air , 2023. *Intelligence artificielle : ça va trop vite ?.* [En ligne]
Disponible sur: https://www.france.tv/france-5/c-dans-l-air/4740136-emission-du-mardi-4-avril-2023.html

[Accès le 08 04 2023].

Futura-Science, S., 2023. *Les emplois qui seront les plus touchés ou remplacés par les IA générative comme ChatGPT, selon Goldman-Sachs.* [En ligne]
Disponible sur: https://www.futura-sciences.com/tech/actualites/intelligence-artificielle-emplois-seront-plus-touches-remplaces-ia-generative-comme-chatgpt-selon-goldman-sachs-104373/
[Accès le 2023 04 30].

Gaikwad, S., 2022. *Top 15 AI-enabled cybersecurity companies in 2022.* [En ligne]
Disponible sur: https://tealfeed.com/top-15-ai-enabled-cybersecurity-companies-u6fht
[Accès le 28 12 2022].

Ganascia, J.-G., 2017. *Le mythe de la Singularité : faut-il craindre l'intelligence artificielle ?.* s.l.:SEUIL.

Ganascia, J.-G., 2022. *Applications de l'intelligence artificielle.* [En ligne]
Disponible sur: https://www.universalis.fr/encyclopedie/intelligence-artificielle-ia/3-applications-de-l-intelligence-artificielle/
[Accès le 22 11 2022].

Gaultier, D., 2022. *ChatGPT d'OpenAI va-t-il bouleverser le machine learning d'entreprise ?.* [En ligne]
Disponible sur: https://www.journaldunet.com/solutions/dsi/1517829-la-nouvelle-generation-d-ia-proposee-par-openai-va-t-elle-bouleverser-le-monde-de-l-intelligence-artificielle-en-entreprise/
[Accès le 30 04 2023].

GEO, S., 2022. [En ligne]
Disponible sur: https://www.geo.fr/histoire/il-y-a-25-ans-lordinateur-deep-blue-domptait-le-roi-des-echecs-209739
[Accès le 22 04 2023].

Gicat, 2015. *CYBERSÉCURITÉ - (PROTÉGER - RÉAGIR - ANALYSER).* [En ligne]
Disponible sur: https://www.hexatrust.com/wp-content/uploads/2015/04/CYBERSECURITE_FR_compressed.pdf

Giot, R., El-Abed, M. & Rosenberger, C., 2014. *Keystroke Dynamics Authentication.* [En ligne]
Disponible sur: https://hal.science/hal-00990373/document

Golden, s.d. *SAP National Security Services.* [En ligne]
Disponible sur: https://golden.com/wiki/SAP_National_Security_Services-VWXWXYD
[Accès le 24 12 2022].

Golman-Sachs, S., 2023. *Generative AI could raise global GDP by 7%.* [En ligne]
Disponible sur: https://www.goldmansachs.com/insights/pages/generative-ai-could-raise-global-gdp-by-7-percent.html
[Accès le 30 04 2023].

Goodfellow, I. J. et al., 2014. *Generative Adversarial Networks.* [En ligne]
Disponible sur: https://arxiv.org/abs/1406.2661
[Accès le 12 02 2023].

Grand View Research, 2022. *Artificial Intelligence In Cybersecurity Market Size, Share & Trends Analysis Report By Type (Cloud Security, Network Security), By Offering, By Technology, By Application, By Vertical, By Region, And Segment Forecasts, 2022 - 2030.* [En ligne]

Disponible sur: https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-cybersecurity-market-report
[Accès le 28 12 2022].

Grandmontagne, Y., 2018. *Détection des fraudes : la convergence du HPC, du Big Data et l'IA.* [En ligne]
Disponible sur: https://itsocial.fr/enjeux-it/enjeux-securite/cybersecurite/detection-fraudes-convergence-hpc-big-data-lia/
[Accès le 11 02 2023].

Groupe d'experts de haut niveau sur l'IA (GEHN IA), 2019. *Lignes directrices en matière d'éthique pour ujne IA digne de confiance P.17.* [En ligne]
Disponible sur: https://op.europa.eu/fr/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1
[Accès le 07 04 2023].

Growjo, s.d. *Blue Hexagon Revenue and Competitors.* [En ligne]
Disponible sur: https://growjo.com/company/Blue_Hexagon
[Accès le 27 12 2022].

Guembe, B. et al., 2022. *The Emerging Threat of Ai-driven Cyber Attacks: A Review.* [En ligne]
Disponible sur: https://www.tandfonline.com/doi/full/10.1080/08839514.2022.2037254
[Accès le 11 12 2022].

Harang, R., 2020. *Sophos-ReversingLabs (SOREL) 20 Million sample malware dataset.* [En ligne]
Disponible sur: https://ai.sophos.com/2020/12/14/sophos-reversinglabs-sorel-20-million-sample-malware-dataset/
[Accès le 28 01 2023].

Haton, J.-P., 2000. *L'intelligence artificielle.* [En ligne]
Disponible sur: https://vod.canal-u.tv/vod/media/canalu/documents/utls/190900.pdf
[Accès le 12 12 2022].

IBM, 2023. *QRadar User Behavior Analytics.* [En ligne]
Disponible sur: https://www.ibm.com/docs/en/qradar-common?topic=app-qradar-user-behavior-analytics
[Accès le 28 12 2022].

Ijlal, T., 2022. *Artificial Intelligence (AI) Governance and Cyber-Security: A beginner's handbook on securing and governing AI systems.* s.l.:Independently published.

INTERPOL, 2022. *L'avenir de l'action policière.* [En ligne]
Disponible sur: https://www.interpol.int/fr/Notre-action/Innovation/L-avenir-de-l-action-policiere
[Accès le 12 12 2022].

ITU, 2021. *United Nations Activities on Artificial Intelligence (AI).* [En ligne]
Disponible sur: https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2021-PDF-E.pdf

Jääskelä, J., 2020. *Anomaly-Based Insider Threat Detection with Expert Feedback and Descriptions.* [En ligne]
Disponible sur: http://jultika.oulu.fi/files/nbnfioulu-202003171272.pdf
[Accès le 01 02 2023].

Jean, A., 2020. *Une brève introduction à l'intelligence artificielle.* [En ligne]
Disponible sur: https://hal.science/hal-02991385
[Accès le 12 12 2022].

Journodev.tech, 2022. *Apprentissage automatique : les challenges de la qualité des données dans la perspective d'une adéquation aux usages.* [En ligne]
Disponible sur: https://journodev.tech/apprentissage-automatique-les-challenges-de-la-qualite-des-donnees-dans-la-perspective-dune-adequation-aux-usages/
[Accès le 20 02 2023].

Kallenborn, G., 2020. *Ces chercheurs transforment les malwares en images pour mieux les détecter grâce à une IA.* [En ligne]
Disponible sur: https://www.01net.com/actualites/ces-chercheurs-transforment-les-malwares-en-images-pour-mieux-les-detecter-grace-a-une-ia-1912576.html
[Accès le 29 01 2023].

Kania, E. B., 2019. *Chinese Military Innovation in Artificial Intelligence.* [En ligne]
Disponible sur:
https://www.uscc.gov/sites/default/files/June%207%20Hearing_Panel%201_Elsa%20Kania_Chinese%20Military%20Innovation%20in%20Artificial%20Intelligence_0.pdf
[Accès le 22 11 2022].

Kassem, A. K., 2022. *Intelligent system using machine learning techniques for security assessment and cyber intrusion detection.* [En ligne]
Disponible sur: https://theses.hal.science/tel-03522384/document
[Accès le 22 11 2022].

Koller, R., 2022. *La Chine règlemente les recommandations algorithmiques pour protéger les utilisateurs (et l'Etat).* [En ligne]
Disponible sur: https://www.ictjournal.ch/news/2022-01-12/la-chine-reglemente-les-recommandations-algorithmiques-pour-proteger-les
[Accès le 03 05 2023].

Kriegler, S. H., 2020. *Artificial Intelligence Guided Battle Management: Enabling Convergence in Multi-Domain Operations.* [En ligne]
Disponible sur: https://apps.dtic.mil/sti/pdfs/AD1159377.pdf
[Accès le 21 11 2022].

La finance pour tous, 2022. *Baisse de la fraude des paiements sur Internet en France en 2021.* [En ligne]
Disponible sur: https://www.lafinancepourtous.com/2022/07/28/baisse-de-la-fraude-des-paiements-sur-internet-en-france-en-2021/
[Accès le 10 02 2023].

Lahoti, S., 2019. *A universal bypass tricks Cylance AI antivirus into accepting all top 10 Malware revealing a new attack surface for machine learning based security.* [En ligne]
Disponible sur: https://hub.packtpub.com/a-universal-bypass-tricks-cylance-ai-antivirus-into-accepting-all-top-10-malware-revealing-a-new-attack-surface-for-machine-learning-based-security/
[Accès le 27 12 2022].

Larue, J., 2022. *Attaque vs défense : qui bénéficie le plus de l'IA dans le cyberespace ?.* [En ligne]
Disponible sur: https://portail-ie.fr/analysis/4080/attaque-vs-defense-qui-beneficie-le-plus-de-lia-dans-le-cyberespace
[Accès le 22 11 2022].

Laurain, A., 2022. *Intelligence artificielle et big data en médecine : une nouvelle corde à notre arc.* [En ligne]
Disponible sur: https://www.cairn.info/revue-hegel-2022-1-page-1.htm
[Accès le 20 02 2023].

Légifrance, 2016. *LOI n° 2016-1321 du 7 octobre 2016 pour une République numérique.* [En ligne]
Disponible sur: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746
[Accès le 25 02 2023].

Légifrance, 2017. *Décret n° 2017-330 du 14 mars 2017 relatif aux droits des personnes faisant l'objet de décisions individuelles prises sur le fondement d'un traitement algorithmique.* [En ligne]
Disponible sur: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000034194929
[Accès le 25 02 2023].

Levy, L., 2021. *Chapitre 10. Data + iA = éthique : équation impossible ?.* [En ligne]
Disponible sur: https://www.cairn.info/sortez-vos-donnees-du-frigo--9782100817054-page-135.htm#s2n5
[Accès le 22 03 2023].

Libeau, D., 2023. *"J'ai déposé une plainte à la CNIL contre ChatGPT".* [En ligne]
[Disponible sur :https://blog.davidlibeau.fr/jai-depose-une-plainte-a-la-cnil-contre-chatgpt/
[Accès le 08/04/2023].

Li, Z. et al., 2018. *VulDeePecker: A Deep Learning-Based System for Vulnerability Detection.* [En ligne]
Disponible sur: https://arxiv.org/pdf/1801.01681.pdf
[Accès le 13 02 2023].

Loiseau, J.-C. B., 2019. *Rosenblatt's perceptron, the first modern neural network.* [En ligne]
Disponible sur: https://towardsdatascience.com/rosenblatts-perceptron-the-very-first-neural-network-37a3ec09038a
[Accès le 21 11 2022].

Lou, R., 2019. *Gmail utilise l'IA pour bloquer 100 millions de spams supplémentaires par jour.* [En ligne]
Disponible sur: https://www.journaldugeek.com/2019/02/07/gmail-utilise-lia-bloquer-100-millions-de-spams-supplementaires-jour/
[Accès le 29 12 2022].

Louvet, B., 2021. *2024 pourrait ressembler au « 1984 » de Georges Orwell, prévient le président de Microsoft.* [En ligne]
Disponible sur: https://sciencepost.fr/2024-pourrait-ressembler-au-1984-de-georges-orwell-previent-le-president-de-microsoft/
[Accès le 06 05 2023].

Malki, F., 2023. *La Chine souhaite encadrer l'intelligence artificielle générative.* [En ligne]
Disponible sur: https://www.lasemainedecastres.fr/la-chine-souhaite-encadrer-lintelligence-artificielle-generative/#mettre-en-place-une-inspection-de-securite-pour-les-outils-dia

[Accès le 03 05 2023].

Marin, J., 2023. *Les eurodéputés se rapprochent d'un accord sur la réglementation des IA génératives.* [En ligne]
Disponible sur: https://www.usine-digitale.fr/article/les-eurodeputes-se-rapprochent-d-un-accord-sur-la-reglementation-des-ia-generatives.N2121476
[Accès le 08 05 2023].

MarketsandMarkets, 2022. *Artificial Intelligence in Cybersecurity Market by Offering (Hardware, Software, and Service), Deployment Type, Security Type, Technology (ML, NLP, and Context-Aware), Application (IAM, DLP, and UTM), End User and Geography - Global Forecast to 2028.* [En ligne]
Disponible sur: https://www.marketsandmarkets.com/Market-Reports/artificial-intelligence-security-market-220634996.html
[Accès le 28 12 2022].
McCulloch, W. & Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*.

Mello Jr., J. P., 2019. *When machine learning is hacked: 4 lessons from Cylance.* [En ligne]
Disponible sur: https://techbeacon.com/security/when-machine-learning-hacked-4-lessons-cylance
[Accès le 27 12 2022].

Mériot, E., 2022. *DARKTRACE : LA PREMIÈRE PLATEFORME DE CYBERDÉFENSE AUTONOME AU MONDE.* [En ligne]
Disponible sur: https://www.challenges.fr/entreprise/darktrace-la-premiere-plateforme-de-cyberdefense-autonome-au-monde_818026
[Accès le 23 12 2022].

Messina, A., 2023. *Vers un encadrement de l'intelligence artificielle générative en Chine.* [En ligne]
Disponible sur: https://siecledigital.fr/2023/04/12/vers-un-encadrement-de-lintelligence-artificielle-generative-en-chine/
[Accès le 12 04 2023].

Miailhe, N., 2018. *Géopolitique de l'Intelligence artificielle : le retour des empires ?.* [En ligne]
Disponible sur: https://www.cairn.info/revue-politique-etrangere-2018-3-page-105.htm
[Accès le 22 11 2022].

Mohammad, R. M., Thabtah, F. & McCluskey, L., 2015. *Phishing Websites Features.* [En ligne]
Disponible sur: https://archive.ics.uci.edu/ml/machine-learning-databases/00327/
[Accès le 25 01 2023].

Msika, S., 2020. *Renforcement de systèmes de détection d'intrusions par des attaques GAN et métaheuristiques.* [En ligne]
Disponible sur: https://publications.polymtl.ca/4192/1/2020_SimonMsika.pdf
[Accès le 22 11 2022].

Muppidi, S., Fisher, L. & Parham, G., 2022. *AI and automation for cybersecurity.* [En ligne]
Disponible sur: https://www.ibm.com/downloads/cas/9NGZA7GK
[Accès le 30 12 2022].

Musk, E., 2023. *Pétition pour geler la recherche sur les LLMs d'Elon Musk, Yoshua Bengio, Victoria Krakovna, Max Tegmark. ...*[En ligne]

Disponible sur: https://24pm.com/intelligence-artificielle/ia-generative/966-petition-pour-geler-la-recherche-sur-les-llms-d-elon-musk-yoshua-bengio-victoria-krakovna-max-tegmark
[Accès le 30 04 2023].

Nanalyze, 2017. *Callsign – Artificial Intelligence for Authentication.* [En ligne]
Disponible sur: https://www.nanalyze.com/2017/07/callsign-ai-authentication/
[Accès le 23 12 2022].

Natanelic, B., 2020. *NLP : Génération de texte par n-grams.* [En ligne]
Disponible sur: https://beranger.medium.com/nlp-génération-de-texte-par-n-grams-3894187f6cd4
[Accès le 28 01 2023].

Nataraj, L., Karthikeyan, S., Jacob, G. & Manjunath, B. S., 2011. *Malware images: visualization and automatic classification.* [En ligne]
Disponible sur: https://dl.acm.org/doi/10.1145/2016904.2016908

Noël, J.-C., 2018. *Intelligence artificielle : vers une nouvelle révolution militaire ?.* [En ligne]
Disponible sur: https://www.ifri.org/sites/default/files/atoms/files/fs84_noel.pdf
[Accès le 22 11 2022].

NOVIPRO, 2019. *IBM QRADAR ADVISOR WITH WATSON: UN SIEM EFFICACE, RENFORCÉ PAR L'IA.* [En ligne]
Disponible sur: https://hub.novipro.com/fr/ibm-qradar-advisor-with-watson-un-siem-efficace-renforcé-par-lia
[Accès le 28 12 2022].

OCDE, I. j. d., 2019. *Recommandation du Conseil sur l'intelligence artificielle.* [En ligne]
Disponible sur: https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0449
[Accès le 10 04 2023].

OTAN, 2022. *Démonstration d'une nouvelle technologie pour la lutte contre le terrorisme dans des lieux très fréquentés.* [En ligne]
Disponible sur: https://www.nato.int/cps/fr/natohq/news_195801.htm?selectedLocale=fr
[Accès le 12 12 2022].

Pariser, E., 2012. *The Filter Bubble: What The Internet Is Hiding From You.* s.l.:Penguin.

Petel, A., 2023. Quelle réglementation européenne sur l'intelligence artificielle ?. *I2D - Information, données & documents 2022/1 (n° 1), pages 22 à 28*.

Picciau, K., 2022. *Qu'est-ce qu'une cyberattaque ?.* [En ligne]
Disponible sur: https://www.stoik.io/cyberattaque
[Accès le 17 01 2022].

Qualys, 2022. *Qualys rachète la plateforme d'IA/ML de Blue Hexagon.* [En ligne]
Disponible sur: https://www.qualys.com/company/newsroom/news-releases/france/qualys-rachete-la-plateforme-dia-ml-de-blue-hexagon/
[Accès le 27 12 2022].

Rendementbourse, 2022. *CrowdStrike Holdings, Inc.*. [En ligne]
Disponible sur: https://rendementbourse.com/crwd-crowdstrike-holdings-inc/finances
[Accès le 23 12 2022].

Rey, N., 2022. *La Maison Blanche dévoile la "déclaration des droits" de l'IA.* [En ligne]
Disponible sur: https://intelligence-artificielle.developpez.com/actu/337334/La-Maison-Blanche-devoile-la-declaration-des-droits-de-l-IA-des-directives-nationales-non-contraignantes-qui-pourraient-eclairer-les-decisions-politiques-et-commerciales-futures/
[Accès le 06 05 2023].

Riccio, N., 2021. *Collaborative A/AI Ops: Delivering Security, Efficiency, Speed, and Scale with Marina.* [En ligne]
Disponible sur: https://www.sapns2.com/collaborative-a-ai-ops-delivering-security-efficiency-speed-and-scale-with-marina/
[Accès le 24 12 2022].

Rotaru, V. et al., 2022. *Event-level prediction of urban crime reveals a signature of enforcement bias in US cities.* [En ligne]
Disponible sur: https://www.nature.com/articles/s41562-022-01372-0

Russell, S., 2016. *Q & A: The future of artificial intelligence.* [En ligne]
Disponible sur: https://people.eecs.berkeley.edu/~russell/research/future/q-and-a.html
[Accès le 22 11 2022].

Samson Jr , R., 2022. *Top 10 Intrusion Detection And Prevention Systems.* [En ligne]
Disponible sur: https://www.clearnetwork.com/top-intrusion-detection-and-prevention-systems/
[Accès le 30 01 2023].

Schaeffer, F., 2020. *La Chine prête à tout pour être le leader mondial de l'IA.* [En ligne]
Disponible sur: https://www.lesechos.fr/tech-medias/intelligence-artificielle/la-chine-prete-a-tout-pour-etre-le-leader-mondial-de-lia-1173173
[Accès le 22 11 2022].

strategy, E. d., 2023. *strategy, EU digital.* [En ligne]
Disponible sur: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence
[Accès le 19 03 2023].

Tabouy, L., 2022. *Introduction. Quel monde voulons-nous ? Jusqu'où pouvons-nous et devons-nous aller ?.* [En ligne]
Disponible sur: https://www.cairn.info/revue-realites-industrielles-2022-3-page-6.htm
[Accès le 22 11 2022].

Tarpin, L., 2023. *LA FRANCE RATIFIE LA CONVENTION 108 + DU CONSEIL DE L'EUROPE.* [En ligne]
Disponible sur: https://cyberjustice.blog/2023/04/24/la-france-ratifie-la-convention-108-du-conseil-de-leurope/
[Accès le 03 05 2023].

Teboul, B., 2022. *Le tournant cognitif de la cybersécurité: changement de paradigme et prolégomènes à la cybersécurité cognitive..* [En ligne]
Disponible sur: https://hal.science/hal-03639141/document
[Accès le 22 11 2022].

Technologies, S., 2023. *ChatGPT Malware: A New Threat in Cybersecurity.* [En ligne]
Disponible sur: https://www.sangfor.com/blog/cybersecurity/chatgpt-malware-a-new-threat-in-cybersecurity
[Accès le 04 03 2023].

TEHTRIS, 2022. *XDR vs EDR, quelles différences et quels avantages ?.* [En ligne]
Disponible sur: https://tehtris.com/fr/blog/xdr-vs-edr-quelles-differences-quels-avantages
[Accès le 26 12 2022].

Thales, 2021. *Reconnaissance faciale : 7 tendances à suivre pour 2021.* [En ligne]
Disponible sur:
https://www.thalesgroup.com/fr/europe/france/dis/gouvernement/biometrie/reconnaissance-faciale
[Accès le 06 02 2023].

Thibodeau , P., 2016. *La Maison Blanche met en garde contre les dérives de l'intelligence artificielle.* [En ligne]
Disponible sur: https://www.lemondeinformatique.fr/actualites/lire-la-maison-blanche-met-en-garde-contre-les-derives-de-l-intelligence-artificielle-64739.html
[Accès le 20 04 2023].

Thibout, C., 2018. *La France est-elle armée dans la course à l'intelligence artificielle ?.* [En ligne]
Disponible sur: https://www.iris-france.org/109396-la-france-est-elle-armee-dans-la-course-a-lintelligence-artificielle/
[Accès le 22 11 2022].

Tracxn, 2023. *Antiy Labs.* [En ligne]
Disponible sur: https://tracxn.com/d/companies/antiy-labs/__XPWLfTz9OGfO2-pKjzDmYpqpx9eYvc2j7hDfruN8zrY/funding-and-investors
[Accès le 29 12 2022].

Trueman, C., 2023. *Premiers pas des États-Unis et de la Chine pour réguler l'IA générative.* [En ligne]
Disponible sur: https://www.lemondeinformatique.fr/actualites/lire-premiers-pas-des-etats-unis-et-de-la-chine-pour-reguler-l-ia-generative-90140.html
[Accès le 03 05 2023].

Van Veen, F., 2016. *Neural Network Zoo.* [En ligne]
Disponible sur: https://www.asimovinstitute.org/neural-network-zoo/
[Accès le 23 04 2023].

Vang, J., 2020. *Big Data and AI: Why We No Longer Have Free or Fair Elections.* [En ligne]
Disponible sur: https://www.glimpsefromtheglobe.com/topics/politics-and-governance/big-data-and-ai-why-we-no-longer-have-free-or-fair-elections/
[Accès le 22 11 2022].

Vergera, I., 2023. *L'intelligence artificielle va-t-elle supprimer 300 millions d'emplois?*. *Le Figaro*, 02 05, p. 12.

Vilain, Z., 2023. *Nous avons porté plainte devant la CNIL hier contre #OpenAI #ChatGPT suite à une demande d'accès à mes données personnelles restée sans réponse.* [En ligne]
Disponible sur : https://twitter.com/Zoe_Vilain/status/1643554710316564480?ref_src=twsrc%5Etfw
[Accès le 08/04/2023]

Villani, C., 2018. *Donner un sens à l'intelligence artificielle pour une stratégie nationale et européenne.* [En ligne]
Disponible sur: https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf
[Accès le 22 11 2022].

Vrankulj, A., 2013. *Coursera looks to verify online student identity with photo, keystroke dynamics.* [En ligne]
Disponible sur: https://www.biometricupdate.com/201301/coursera-looks-to-verify-online-student-identity-with-photo-keystroke-dynamics
[Accès le 06 02 2023].

W., M. C. W. e. P., s.d.
Webroot, 2018. *The Webroot Approach to Machine Learning.* [En ligne]
Disponible sur: https://www-cdn.webroot.com/1215/2510/8234/Machine-Learning-Webroot-Approach-WP_US.pdf
[Accès le 26 12 2022].

Wolfe, C. R., 2022. *Language Model Scaling Laws and GPT-3.* [En ligne]
Disponible sur: https://cameronrwolfe.substack.com/p/language-model-scaling-laws-and-gpt
[Accès le 02 04 2023].

Yonnet, P., 2022. *L'IA joue un rôle dans les SERPs de Google….* [En ligne]
Disponible sur: https://www.neper.fr/2022/02/13/lia-joue-un-role-dans-les-serps-de-google/
[Accès le 11 12 2022].

ZED.NET, 2023. *ChatGPT : les premières plaintes françaises enregistrées par la CNIL.* [En ligne]
Disponible sur: https://www.zdnet.fr/actualites/chatgpt-les-premieres-plaintes-francaises-enregistrees-par-la-cnil-39956702.htm
[Accès le 08 04 2023].

Zhu, J., 2022. *The Personal Information Protection Law: China's Version of the GDPR?.* [En ligne]
Disponible sur: https://www.jtl.columbia.edu/bulletin-blog/the-personal-information-protection-law-chinas-version-of-the-gdpr
[Accès le 03 05 2023].

Zippia, s.d. *FireEye.* [En ligne]
Disponible sur: https://www.zippia.com/fireeye-careers-4343/
[Accès le 26 12 2022].

Zippia, s.d. *Webroot.* [En ligne]
Disponible sur: https://www.zippia.com/webroot-careers-44206/
[Accès le 26 12 2022].

Zolynski, C., 2015. *Big data : pour une éthique des données.* [En ligne]
Disponible sur: https://www.cairn.info/revue-i2d-information-donnees-et-documents-2015-2-page-25.htm#no1
[Accès le 20 02 2023].

ZoomInfo, s.d. *Callsign.* [En ligne]
Disponible sur: https://www.zoominfo.com/c/callsign-inc/356463883
[Accès le 26 12 2022].

ZoomInfo, s.d. *Cynet.* [En ligne]
Disponible sur: https://www.zoominfo.com/c/cynet/368635112
[Accès le 26 12 2022].

ZoomInfo, s.d. *SAP NS2.* [En ligne]
Disponible sur: https://www.zoominfo.com/c/sap-ns2/347547546
[Accès le 24 12 2022].

ZoomInfo, s.d. *ThreatBook.* [En ligne]
Disponible sur: https://www.zoominfo.com/c/threatbook/425712803
[Accès le 29 12 2022].