En juin 2025, au Salon du Bourget, l'AMIAD et Dassault Aviation signaient un accord de R&D sur l'intégration de l'IA dans le combat aérien. Quelques mois plus tôt, l'agence lançait Pendragon : d'ici 2027, une force de robots coordonnés par une IA collective sera déployée au sein de l'armée de Terre. Derrière ces annonces se profile un enjeu que les armées ne peuvent pas ignorer. Car si la cybersécurité des systèmes d'information militaires fait l'objet d'une doctrine structurée depuis plusieurs années, que se passe-t-il lorsque c'est le système d'IA lui-même, que ce soit son modèle, ses données d'entraînement ou ses inférences en temps réel qui est attaqué ? Cette question invite à repenser la cybersécurité militaire dans sa profondeur. Non plus seulement comme la protection des réseaux et des infrastructures, mais comme la garantie de l'intégrité, de la fiabilité et de la robustesse des systèmes intelligents qui informent, assistent et parfois conditionnent la décision au combat.

Le cyberespace : un nouveau champ de confrontation

En 2010, le Département de la Défense des États-Unis a défini le cyberespace comme « un domaine mondial au sein de l'environnement de l'information, constitué du réseau interdépendant des infrastructures de technologies de l'information et des données résidentes, incluant Internet, les réseaux de télécommunications, les systèmes informatiques, ainsi que les processeurs et contrôleurs intégrés ». Or, seulement trois ans plus tard, en 2013, le Ministère des Armées français expliquait que ce cyberespace « *est désormais un champ de confrontation à part entière* ». Pourquoi une telle évolution ? Parce qu'il s'agit aussi d'un terrain privilégié d'agressions invisibles : les cyberattaques.

Selon l'Agence nationale de la sécurité des systèmes d'information (ANSSI), une cyberattaque est un « *événement visant à compromettre un ou plusieurs systèmes informatiques dans le but de servir des intérêts malveillants* ». Autrement dit, une action conçue pour fragiliser, miner ou détourner un dispositif numérique à des fins hostiles. Et les cibles potentielles abondent. Les infrastructures militaires sont particulièrement vulnérables : ordinateurs, serveurs, périphériques, mais aussi smartphones, tablettes et objets connectés, qu'ils soient reliés à Internet ou isolés du réseau. Les vecteurs d'attaque, eux, sont tout aussi diversifiés. Ils vont des logiciels malveillants aux manipulations humaines appelée ingénierie sociale, en passant par des actions physiques ciblées. Un exemple, facile à comprendre, est celui de brancher une clé USB infectée sur un ordinateur pourtant sécurisé.

Les cyberattaques se déclinent sous de multiples formes. Certaines consistent en attaques par déni de service (DDoS) pour paralyser les systèmes critiques. D'autres passent par la chaîne d'approvisionnement afin d'introduire des backdoors ou des malwares. D'autres encore reposent sur l'interception de communications sensibles, typiques des attaques dites *man-in-the-middle*. Le spectre est large. Certaines attaques visent à falsifier des données afin d'induire en erreur la chaîne décisionnelle, ce qui revient, ni plus ni moins, à influer directement sur le cours des opérations. D'autres, plus offensives, s'emploient à prendre le contrôle de systèmes d'armes, dans le but de les neutraliser ou, pire encore, de les retourner contre leur propriétaire.

On comprend là que le cyberespace ne se contente pas d'accompagner les conflits modernes, il en redessine les contours. Il devient un théâtre d'affrontements à part entière, un espace où l'avantage opérationnel peut se gagner ou se perdre sans bruit, sans éclat, mais avec des effets bien réels sur le terrain. En somme, il s'impose bel et bien comme un nouveau champ de confrontation, aux règles encore incertaines et aux limites souvent mouvantes. Un champ singulier où l'avantage opérationnel peut être conquis sans qu'un seul tir ne soit tiré.

L'IA de défense : entre avancées stratégiques et nouvelles vulnérabilités

Or, ce champ de confrontation ne cesse de s'élargir. Et son extension la plus récente tient en deux lettres : IA. Car l'intelligence artificielle ne fait pas que s'inviter sur le théâtre d'opérations, elle en reconfigure les fondements mêmes.

Emmanuel Chiva, Délégué général pour l'armement en France de juillet 2022 à novembre 2025, met le doigt sur un enjeu capital : l'IA devrait permettre aux machines non seulement d'analyser des situations d'une grande complexité, mais également de prendre des décisions en temps réel et, ce faisant, de décharger l'humain de tâches jugées soit trop périlleuses, soit à faible valeur ajoutée. Ceci étant dit, pour les armées, l'IA fait feu de tout bois dans plusieurs domaines cruciaux. L'assistance au commandement, d'une part. La surveillance et la reconnaissance, d'autre part. Sans compter le déploiement de systèmes robotiques autonomes, véritable cheval de bataille de demain, ainsi que l'entraînement des forces armées. Bref, le spectre d'applications ne manque pas d'envergure. Qui plus est, la liste s'étoffe au fur et à mesure que la recherche en IA repousse les frontières du possible. Au bout du compte, maîtriser l'IA sur le théâtre d'opérations n'est plus un simple atout dans sa manche. C'est un impératif catégorique pour tout pays qui nourrit l'ambition légitime de garder une longueur d'avance et, par ricochet, de préserver un avantage opérationnel décisif face à l'adversaire.

Cependant, l'intégration de systèmes d'IA introduit également de nouvelles vulnérabilités pour leurs utilisateurs. Dès 2019, la ministre des armées de l'époque, Madame Florence Parly, soulignait les risques potentiels liés à l'IA, en insistant sur les divers dangers à prendre en compte dès la phase de conception : « *La manipulation des données d'apprentissage, les biais cognitifs transmis par l'homme aux algorithmes, les systèmes désorientés et mis en défaut par un simple bout de scotch, les systèmes hackables à distance : les facteurs de risques que nous devons évaluer et maîtriser dès la conception sont extrêmement nombreux* ». Ces avertissements, formulés il y a plusieurs années, n'ont rien perdu de leur acuité. Bien au contraire : ils ont pris une résonance nouvelle à mesure que l'IA militaire s'est déployée.

Car les vulnérabilités pointées par la ministre ne sont pas des accidents de parcours — elles sont structurelles. Et elles font écho, de façon troublante, aux menaces cyber déjà évoquées. L'émergence de l'utilisation de l'IA par les armées élargit, et de façon considérable, ce nouveau champ cyber de confrontation, déjà foisonnant. En effet, une attaque dirigée contre un modèle d'IA de défense pourra, à juste titre, être considérée comme une cyberattaque. Pourquoi ? Parce qu'elle exploite les mêmes fragilités et poursuit des objectifs similaires à ceux des intrusions informatiques plus classiques. Dit autrement, l'ennemi n'a pas besoin de réinventer

ses méthodes : il lui suffit d'adapter ses pratiques aux nouveaux outils. Et l'on comprend vite que l'IA, censée renforcer la résilience opérationnelle, peut tout aussi bien se transformer en talon d'Achille.

Toutes les menaces de cybersécurité sont ici transposables. Vol de données, perturbation des services, manipulation d'informations sensibles : rien n'échappe au spectre des risques. Et les motivations de l'attaquant, loin d'être inédites, se calquent sur celles observées face à une infrastructure informatique traditionnelle. Il cherchera à compromettre la confidentialité, l'intégrité ou la disponibilité des données, mais aussi, ce qui est plus préoccupant encore, le modèle d'IA lui-même. Les finalités de telles offensives sont multiples et parfois particulièrement sournoises. Il peut s'agir de détourner les décisions d'un algorithme afin de semer la confusion au sein de la chaîne de commandement, de subtiliser des informations critiques pour alimenter sa propre puissance, ou encore de rendre un modèle totalement inutilisable, avec pour effet direct la paralysie d'un dispositif de défense. En d'autres termes, ces attaques exploitent avec une redoutable efficacité les failles d'un système, tout comme le ferait une cyberattaque dite classique. C'est tout l'équilibre opérationnel qui peut vaciller, parfois pour un détail apparemment insignifiant.

Panorama des attaques : une taxonomie au service de la compréhension des menaces

Pour anticiper ces menaces, il convient d'abord de les nommer et de les classer avec précision. La Commission nationale de l'informatique et des libertés (CNIL) distingue trois grandes familles d'attaques visant les modèles d'IA : les infections, les manipulations et, enfin, les exfiltrations. Trois mots, presque cliniques, derrière lesquels se cachent des réalités autrement plus préoccupantes. Or, les systèmes d'IA civils et militaires partagent bien souvent une base technologique commune, parfois même strictement identique, ou du moins convergente : cette taxonomie s'applique donc pleinement au champ militaire.

Les attaques par infection sont des attaques qui ciblent la phase d'apprentissage des modèles d'IA. En intervenant à ce moment clé du cycle de vie du système, les attaquants peuvent modifier le comportement de l'algorithme de manière significative pour le contrôler de manière dissimulée ou pour détériorer son fonctionnement plus tard. L'attaque peut prendre plusieurs visages : tantôt en insérant des données incorrectes, tantôt en modifiant subrepticement celles qui existent déjà, avec pour conséquence directe d'amener le modèle à apprendre des schémas erronés. Or, dans un tel contexte, l'attaque dite *par infection* agit comme un véritable cheval de Troie : elle pousse le modèle à mal classer les données, à générer des résultats inexacts, biaisés, voire délibérément malveillants. Ainsi, le système devient moins performant, peu fiable, ou encore, ce qui est autrement plus préoccupant, vulnérable à des erreurs critiques. En pratique, imaginons, par exemple, qu'un attaquant injecte dans l'algorithme d'entraînement des données volontairement altérées : l'IA, censée reconnaître des cibles ennemies avec précision, se met alors à confondre l'ami et l'ennemi. Et l'on devine sans peine les conséquences : la sécurité opérationnelle se retrouve mise en péril.

Les attaques par manipulation, plus connues sous le terme anglais adversarial attack, visent à tromper les systèmes d'IA non pas durant l'apprentissage, mais bel et bien lors de leur phase

d'utilisation, une fois celle-ci achevée. Autrement dit, elles surgissent au moment où l'on croit, à tort, que le modèle est stabilisé et fiable. Le principe, en apparence anodin, est redoutable : il s'agit d'introduire de légères perturbations dans les données d'entrée du modèle, perturbations qui donnent naissance à des « *exemples contradictoires* ». À l'œil humain, tout semble normal, rien ne dépasse. Mais pour l'algorithme, ces signaux déformés sont autant de pièges, et les erreurs de prédiction se multiplient. Ici, l'illusion est totale, comparable à un mirage sur la route qui trompe le regard le plus exercé. Ces modifications subtiles, qui exploitent les failles du système comme on enfonce un coin dans une fissure, suffisent à induire des décisions incorrectes : une classification erronée, une évaluation faussée, un jugement biaisé. C'est bien là que réside le danger. Car, en pratique, les conséquences peuvent être lourdes de sens. Imaginons, par exemple, un attaquant introduisant des perturbations quasi invisibles dans les images traitées par les capteurs d'une plateforme militaire. Pour un humain, rien à signaler. Pour l'IA, en revanche, tout bascule : un missile ennemi peut soudain être interprété comme un aéronef civil anodin, ou, à l'inverse, un avion de ligne parfaitement inoffensif être perçu comme une menace à abattre. On le voit, la frontière entre le réel et l'artefact devient dangereusement poreuse.

Les attaques par exfiltration visent, comme leur nom l'indique, à dérober des données critiques issues des systèmes d'IA. Trois grandes catégories se distinguent : l'inférence d'appartenance, l'inversion de modèle et l'extraction de modèle, chacune exploitant à leur manière les failles de l'apprentissage automatique. Premièrement, l'attaque par inférence d'appartenance : elle permet de déterminer si une donnée spécifique a été utilisée lors de l'entraînement. En pratique, l'assaillant joue sur les différences de confiance et de précision que manifeste le modèle selon les données soumises. Si le modèle affiche un niveau de certitude anormalement élevé pour un exemple donné, cela laisse entendre, demi-mot, que cette donnée faisait bel et bien partie de l'ensemble d'apprentissage. Un attaquant pourrait ainsi chercher à vérifier si des informations sensibles, par exemple sur un type particulier de véhicule, d'arme ou de plateforme, figurent dans la mémoire du système. Deuxièmement, l'attaque par inversion de modèle. Ici, le principe est qu'à partir des seules sorties du modèle, l'adversaire tente de reconstituer les données d'entrée qui ont servi à l'entraînement. Autrement dit, il cherche à extraire une représentation moyenne, une sorte d'empreinte statistique, de chaque catégorie apprise. Et de fil en aiguille, ce procédé peut déboucher sur la reconstitution de données sensibles, qu'on croyait pourtant inaccessibles. Enfin, l'extraction de modèle constitue sans doute la forme la plus insidieuse. Il s'agit ni plus ni moins que de voler un modèle entier en traitant le système comme une « *boîte noire* » : nul accès direct au code, nul accès aux données, mais une stratégie patiente de questions-réponses. À force de multiplier les entrées et d'analyser scrupuleusement les sorties, l'attaquant parvient à en élaborer une copie fidèle, presque un clone. Ce faisant, il dérobe non seulement l'architecture globale, mais aussi les paramètres (poids, biais etc.) et même les hyperparamètres, comme le nombre de couches d'un réseau de neurones. Cette situation ne pourrait bien évidemment arriver qu'à la condition qu'un ennemi réudssise à capturer une plateforme équipée d'un modèle d'IA afin de l'interroger à son bon loisir, question après question.

Des implications concrètes sur le champ de bataille au conditions d'un déploiement maîtrisé

Les attaques sur l'IA vont constituer une menace considérable pour les véhicules militaires, car au lieu de chercher à les détruire directement, des armées ennemies pourraient exploiter des vulnérabilités dans leurs modèles d'IA ou sur leurs données pour neutraliser leur capacité opérationnelle. Par exemple, le combat aérien est un domaine maîtrisé par peu de pays et où toucher sa cible reste particulièrement complexe. Or, si un avion de chasse devient aveugle à la suite d'une cyberattaque visant ses capteurs dopés à l'IA, le pilote perd sa capacité de décision sur les systèmes d'armes et de pilotage, ce qui met directement sa vie en danger. Comme l'a souligné Will Roper, ancien secrétaire adjoint de l'US Air Force, les avions modernes comportent des millions de lignes de code. Si l'une d'elles est vulnérable ou défectueuse, même un pays sans les moyens de développer un avion de chasse pourrait « mettre hors service cet avion avec juste quelques frappes », simplement en exploitant ces failles via des cyberattaques ciblées. Cela démontre que la menace pour les véhicules utilisant des modèles d'IA ne proviendra pas uniquement des missiles ou des combats directs, mais aussi des attaques informatiques visant les systèmes vitaux afin de les neutraliser.

Face à ces menaces, il apparaît essentiel de réfléchir aux moyens à mettre en place pour anticiper et contrer ces attaques. Car si l'intelligence artificielle devient peu à peu une pièce maîtresse du champ de bataille, elle en devient, ipso facto, une cible de choix. Et pour être déployée sur un champ de bataille, une IA doit impérativement être explicable, fiable et robuste. Ces trois qualités ne sont pas de simples critères techniques : elles fondent la confiance nécessaire qui devient une véritable condition de cybersécurité opérationnelle.

L'opacité des systèmes d'intelligence artificielle, souvent désignée par l'« effet boîte noire », constitue un obstacle majeur à leur emploi militaire. Une IA qui génère des résultats sans que son raisonnement puisse être compris devient immédiatement suspecte aux yeux des opérateurs. Pire encore, elle se transforme en cible de choix, notamment face à des attaques par infection ou par manipulation, connues sous le nom d'*adversarial attack*. Dans le premier cas, l'attaque consiste à empoisonner les données d'apprentissage afin d'altérer, insidieusement et à long terme, le comportement futur du modèle. Dans le second, plus retors, les *adversarial attack* s'attaquent à la phase d'utilisation : de simples perturbations, invisibles à l'œil nu mais soigneusement introduites dans les données d'entrée, suffisent à provoquer des erreurs de classification. Un détail qui paraît anodin pour un humain peut, en réalité, faire basculer tout un système. Assurer l'explicabilité devient donc une véritable question de cybersécurité. C'est elle qui garantit la traçabilité et l'auditabilité des décisions, permettant non seulement de valider la légitimité d'une action militaire, mais aussi d'en identifier les dérives. Pensons par exemple à un tir fratricide ou à des dommages collatéraux : sans explicabilité, l'analyse des causes reste dans le brouillard. Avec elle, au contraire, l'on dispose d'un outil de vérification et d'alerte. En définitive, l'explicabilité agit comme un bouclier numérique. Elle éclaire les zones d'ombre, réduit l'espace exploitable par l'ennemi et, ce faisant, renforce la résilience globale des systèmes d'IA militaires. C'est là un enjeu cardinal, car dans ce champ de confrontation inédit, comprendre le pourquoi des décisions n'est pas un luxe mais bel et bien une nécessité stratégique.

La confiance en l'IA suppose également qu'elle soit fiable et robuste. La fiabilité suppose que les performances d'un système demeurent constantes, quelles que soient les conditions de déploiement. Dit autrement, un modèle fiable doit produire les mêmes résultats dans un laboratoire parfaitement contrôlé que sur un théâtre d'opérations où règnent incertitude, stress et imprévus. La robustesse, quant à elle, renvoie à une autre exigence, tout aussi cruciale : la capacité de résister aux perturbations, de fonctionner dans des environnements dégradés, ou encore de supporter des attaques destinées à détourner ses résultats. Autrement dit, il ne suffit pas qu'un modèle soit précis en temps normal, il faut aussi qu'il tienne la route lorsque les vents contraires se lèvent.Comme le souligne l'Amiral Pierre Vandier, la robustesse doit « garantir aux décideurs politiques que quand il va sur la base d'un traitement d'IA de nos données prendre une décision, alors cette décision sera maîtrisée ». Cette notion est donc intimement liée au développement de la cybersécurité des systèmes d'IA de défense.

Sécuriser l'IA de défense : des méthodes traditionnelles aux approches innovantes

Un système d'IA repose essentiellement sur deux éléments clés : des données et un modèle. Assurer la cybersécurité d'une IA de défense revient donc à protéger ces deux aspects tout au long de leur cycle de vie. Les pratiques traditionnelles issues du NIST Cybersecurity Framework constituent un socle de départ, avant que de nouvelles méthodes dédiées viennent compléter l'arsenal.

Ce cadre méthodologique s'articule autour de cinq fonctions clés : identifier, protéger, détecter, répondre et restaurer. Appliqué aux IA de défense, chacune de ces fonctions recouvre des enjeux spécifiques détaillés ci-après.

L'identification consisterait à recenser les actifs critiques et analyser leurs vulnérabilités. Cette étape impliquerait de cartographier les données d'entraînement sensibles issues de capteurs militaires, ainsi que les modèles d'IA déployés sur le terrain. Ensuite, la fonction de protection viserait à assurer la sécurité et la résilience des IA. Elle reposerait sur des mesures classiques de cybersécurité : authentification, chiffrement et gestion des accès pour protéger les données d'entraînement contre l'infection par exemple, mais aussi sur des pratiques de secure coding, la protection des endpoints et la sécurisation du réseau afin de protéger le développement des modèles. Vient ensuite la détection, qui viserait à identifier les incidents de cybersécurité tels qu'une intrusion dans les données d'apprentissage ou une attaque sur un modèle d'IA. Cette fonction serait assurée par une supervision en continu via des systèmes de détection d'intrusion (IDS) et/ ou des solutions de gestion des événements (SIEM). Elle serait complétée par des tests spécifiques menés par des *red teams*, qui vont par exemple simuler des *adversarial attack*, afin de repérer toute tentative de manipulation avant qu'elle ne compromette une mission. En cas d'incident, la réponse pourrait venir des SOC spécialisés dans la surveillance d'IA de défense qui sauraient distinguer une anomalie technique d'une attaque ciblée et de déployer des contre-mesures adaptées. Enfin, la restauration garantirait la continuité des opérations grâce aux plans de reprise et aux sauvegardes, mais aussi en prévoyant la reconstitution de modèles compromis ou leur réentraînement sur des données fiables, condition indispensable pour maintenir la confiance et l'efficacité des systèmes d'IA en

contexte militaire.En définitive, le *NIST Cybersecurity Framework* offrirait un socle précieux pour sécuriser les IA de défense, mais il resterait insuffisant face à la nature dynamique et évolutive des modèles. Dans un contexte militaire, où l'adversaire chercherait activement à exploiter ces vulnérabilités, il conviendrait donc de compléter ce cadre par de nouveaux mécanismes dédiés à plus cybersécuriser l'IA.

La sécurisation des données, essentielles à l'apprentissage, représente un enjeu stratégique majeur, à la fois lors de la phase de développement et de déploiement des modèles d'IA de défense.

Lors de la phase de développement, deux méthodes semblent émerger comme particulièrement prometteuses. La première est le chiffrement homomorphique qui permet de traiter et d'analyser les données sans jamais les déchiffrer, assurant ainsi leur confidentialité même lors des opérations de calcul. La seconde est le *Federated Learning* (« apprentissage fédéré » en français) qui implique la décentralisation de l'entraînement des modèles en permettant à plusieurs appareils d'apprendre localement à partir de leurs propres données, tout en partageant uniquement les résultats d'apprentissage avec le système central. Ainsi, les données brutes ne sont jamais centralisées, ce qui limite considérablement les risques d'attaque ou d'exfiltration de données pendant le développement.

Lors de la phase de déploiement d'un modèle, plusieurs approches complémentaires pourraient être particulièrement pertinentes pour renforcer la sécurité des données qu'il utilise. La première est le *Poison Control*, une méthode de détection des attaques par infection. Elle repose sur une surveillance continue des ensembles de données, de manière à identifier toute anomalie laissant soupçonner l'injection de données falsifiées. Dans un contexte militaire, une telle méthode permettrait, par exemple, de repérer si un adversaire tente de glisser de fausses images satellites dans un système d'analyse géospatiale. La deuxième approche, appelée *Input Control*, concentre son action sur le filtrage des entrées utilisateurs afin de contrer d'éventuelles attaques. Elle repose sur la validation des formats, l'analyse de la cohérence sémantique et, lorsque nécessaire, l'application de règles statistiques plus strictes. Enfin, la troisième approche, dite *Transform Inputs* (« *transformation des entrées* » en français), consiste à modifier légèrement les données avant qu'elles ne soient traitées par le modèle, par exemple en introduisant du bruit aléatoire ou en reformulant les prompts. Cette méthode compliquerait alors la tâche des attaquants qui chercheraient à interroger un modèle d'IA de défense pour en extraire des informations sensibles (qu'il s'agisse de données d'entraînement, de représentations internes ou même d'une copie du modèle lui-même). Dans un système de détection radar, elle pourrait par exemple brouiller partiellement les signaux entrants afin de réduire les risques de manipulation adverse. Au bout du compte, ces trois méthodes poursuivent une même finalité. Limiter l'impact des entrées malveillantes, sans sacrifier la robustesse opérationnelle du modèle. Autrement dit, conjuguer protection et performance, car sur le champ de bataille numérique comme ailleurs, il ne suffit pas de résister : il faut aussi continuer d'agir efficacement.

Après avoir exploré les méthodes innovantes de protection des données, il faut désormais s'intéresser aux modèles eux-mêmes. Leur conception et leur déploiement doivent intégrer de nouvelles approches de cybersécurité dès l'origine, selon le principe du *security by design*,

c'est-à-dire une démarche qui intègre la sécurité comme exigence fondamentale dès la phase de conception plutôt que comme un ajout a posteriori. L'enjeu est d'anticiper les dérives ou attaques, même inconnues au moment du développement, et d'intégrer des contre-mesures afin de garantir la robustesse des IA de défense dans des environnements contestés.

Lors de la phase de développement du modèle, plusieurs méthodes innovantes devraient pouvoir renforcer la résilience des modèles d'IA face aux menaces. L'une de ces techniques est l'apprentissage par exemples contradictoires (dit *adversarial learning*) qui consiste à exposer le modèle à des données volontairement altérées afin de l'habituer à reconnaître et ajouter des exemples malveillants à sa base de données afin d'accroître sa résistance aux *adversarial attack*. Le *Randomized Smoothing* offrirait une autre voie de protection en entraînant une IA à maintenir des prédictions stables, même lorsque les données d'entrée de son algorithme de classification sont perturbées par du bruit, limitant ainsi l'efficacité des cyberattaques. Enfin, les jumeaux numériques devraient représenter une avancée stratégique dans la cybersécurité des IA de défense en permettant de simuler des environnements réalistes et évolutifs pour tester la robustesse des modèles avant leur déploiement, anticiper les vulnérabilités et valider les contre-mesures dans un cadre virtuel sécurisé.

Une fois déployés, les modèles d'IA devront aussi être protégés par de nouvelles approches de cybersécurité spécialement conçues pour contrer les menaces opérationnelles. Les « auto-encodeurs » semblent offrir une première ligne de défense. Placés en amont du modèle, ils transforment les entrées malveillantes afin de limiter les *adversarial attack*, tandis qu'en aval, ils réduisent la quantité d'informations sensibles divulguées, compliquant ainsi les tentatives d'extraction de données. Les *Generative Adversarial Networks* constituent une autre solution prometteuse. En générant des données fictives proches des originales mais expurgées de tout élément sensible, et en filtrant avec soin les entrées suspectes, ces techniques permettent à la fois de limiter les fuites d'information et de déjouer des attaques par infection ou manipulation. Autrement dit, elles construisent une sorte de pare-feu intelligent, capable de brouiller les pistes pour l'adversaire. Enfin, une piste supplémentaire mérite l'attention : la supervision des IA de défense par d'autres IA. Cette approche, qui pourrait sembler paradoxale à première vue, ouvre pourtant la voie à un renforcement considérable de la sécurité. Grâce à une veille continue, ces systèmes détecteraient les moindres déviations statistiques dans les prédictions, comme on repère une fausse note dans une partition bien rodée. Résultat : l'identification précoce des comportements anormaux, le déclenchement de contre-mesures adaptées et, surtout, l'enrichissement des retours d'expérience, essentiels pour affiner les modèles. En somme, l'idée n'est plus seulement de protéger l'IA, mais de lui donner la capacité de se surveiller elle-même, d'apprendre de ses failles et de transformer chaque tentative d'attaque en source de résilience accrue.

Intégrer l'IA dans nos systèmes d'armes et acculturer les militaires à son utilisation est aujourd'hui un impératif stratégique pour tout pays souhaitant conserver un avantage sur ses adversaires. Toutefois, utiliser cette innovation ne se fait pas sans faire peser de nouvelles menaces sur nos armées. L'IA est vulnérable à des attaques numériques, par infection, manipulation ou exfiltration, qui peuvent avoir des conséquences dévastatrices sur le terrain et mettre en danger la vie de nos forces armées ou même de civils. Dans ce contexte, la cybersécurité des modèles et des données d'IA doit évidemment passer par l'application de

mesures traditionnelles mais elle doit aussi s'appuyer sur des approches innovantes pour accroître leur robustesse.

Valentin AUBERT

Président de la Commission de l'Innovation Duale de l'INAS

Pour aller plus loin

- Braiek, H. B., & Khomh, F. (2024). Machine learning robustness: A primer. arXiv. http://arxiv.org/abs/2404.00897
- Billois, G., Bossuet, R., & Pierre-Louis, C. (2024, mars 13). Sécuriser l'IA : les nouveaux enjeux de cybersécurité. RiskInsight. https://www.riskinsight-wavestone.com/2024/03/securiser-lia-les-nouveaux-enjeux-de-cybersecurite/
- CNIL. (n.d.). Attaque par exfiltration de modèle (model evasion attack). CNIL. https://www.cnil.fr/fr/definition/attaque-par-exfiltration-de-modele-model-evasion-attack
- Commission chargée de l'élaboration du Livre blanc sur la défense et la sécurité nationale. (2013). Livre blanc : Défense et sécurité nationale. Présidence de la République, Direction de l'information légale et administrative. https://www.vie-publique.fr/rapport/34001-livre-blanc-2013
- Department of Defense (DoD). (2010). Joint publication 1-02: Department of Defense dictionary of military and associated terms. https://fas.org/irp/doddir/dod/jp1_02.pdf
- LCP. (2024, juin 16). Le journal de la Défense – Intelligence artificielle : les armées accélèrent. LCP. https://lcp.fr/programmes/le-journal-de-la-defense/intelligence-artificielle-les-armees-accelerent-285421
- Meunier, L. (2022). Adversarial attacks: A theoretical journey [Doctoral dissertation, Université Paris Sciences et Lettres]. HAL. https://theses.hal.science/tel-04056444
- National Institute of Standards and Technology (NIST). (2024). The NIST Cybersecurity Framework (CSF) 2.0 (NIST CSWP 29). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.CSWP.29
- Vallet, F. (2022). Petite taxonomie des attaques des systèmes d'IA. Laboratoire d'Innovation Numérique de la CNIL.