

PÔLE D'EXCELLENCE
CYBER



Livre blanc

IA de confiance dans la défense

Juin 2025





Sommaire

Préface

Bertrand RONDEPIERRE (AMIAD) 4

Introduction

Karl NEUBERGER/ Olivier DENTI 6

PARTIE I : L'IA un enjeu stratégique 8

L'IA dans les cyberattaques : un multiplicateur de risques 10

La souveraineté de l'IA pour la Défense : un impératif stratégique pour
une IA de confiance 12

Coopération et confiance dans les projets complexes 14

Les enjeux et perspectives de l'IA de défense à l'échelle 16

L'enjeu n'est pas l'automatisation de la guerre mais la distanciation des
combattants. 18

L'IA de Confiance dans la Défense : opportunités et Défis. 20

PARTIE II : Vers une opérationnalisation de l'IA sur le champ de bataille 22

Industrialisation de l'IA, un pilier de la confiance 24

Construire la confiance en l'IA : la nécessité de la certification.
Le point de vue d'un CESTI 26

Transformations de la guerre : les enjeux de la confiance dans les
applications militaires de l'IA. 28

Vers la sécurisation et la frugalité des fonctions IA embarquées. 30

Face à une IA source de crises, la confiance comme solution ? 32

Standardisation et industrialisation des tests de cyber résilience des IA. . 34

L'IA de confiance est une illusion : construisons des produits de
confiance 36



Copyright Pôle d'excellence cyber©. Édition de juin 2025.

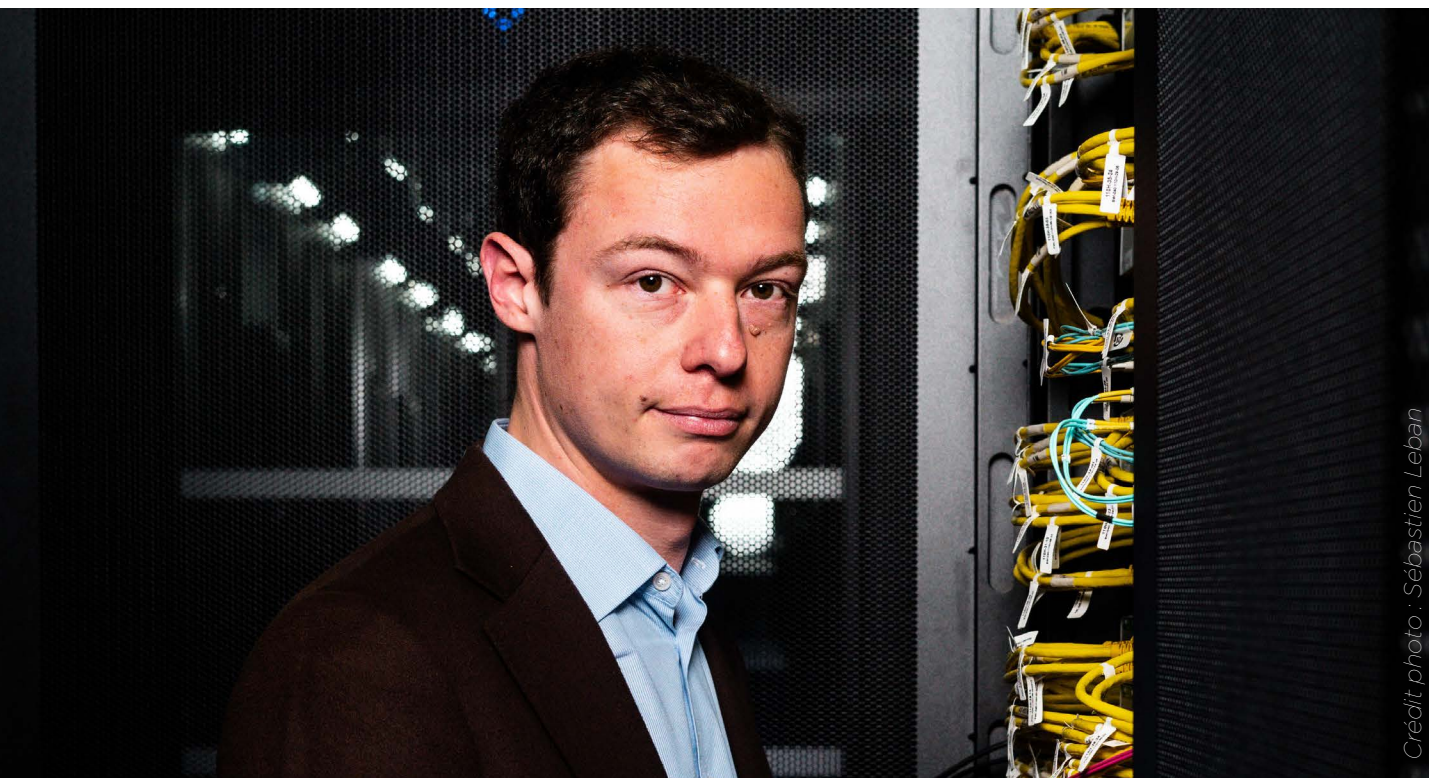
Cette œuvre est mise à disposition sous licence Creative Commons,

Attribution - Pas d'Utilisation Commerciale - Pas de Modification 3.0 France.

*Pour voir une copie de cette licence, visitez <http://creativecommons.org/licenses/by-nc-nd/3.0/fr/> ou écrivez
à Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.*

PARTIE III : Défis techniques de l'IA	38
L'IA de confiance pour la défense : de la conception à l'évaluation.	40
Vers des IA vérifiables	42
L'influence de l'IA sur la cybersécurité : défis et opportunités du point de vue de la recherche.	44
Recommandations de sécurité pour la mise-en œuvre de plateformes d'IA générative dédiées à l'ingénierie des systèmes critiques.....	46
IA raisonnée pour les systèmes décisionnels critiques	48
De la sécurité du Machine Learning	50
Intelligence Artificielle responsable ou de confiance pour la Défense : quelle nouvelle ingénierie ?	52
Fiabilité de l'IA en contexte militaire : Trouver l'équilibre entre autonomie et supervision humaine	54
Les principales vulnérabilités des LLMs face aux attaques malveillantes.	56
La cybersécurité de l'IA au cœur du champ de bataille.....	58
PARTIE IV : Difficulté de la régulation de l'IA dans la défense	60
Systèmes d'Armes Létales Autonomes – aspects juridiques.....	62
Trust by Design : comment être éthique et un leader européen de l'IA ..	64
Comment garantir qu'une IA respectera les principes de proportionnalité et de distinction.....	66
Quelle régulation de l'IA de défense ?	68
SALA - SALIA, une distinction « de confiance » ?.....	70
Notes / Références / Annexes	72
Conclusion	78

PRÉFACE



Crédit photo : Sébastien Leban

En ce début 2025, dans un contexte géopolitique marqué par des tensions croissantes redéfinissant l'ordre mondial, l'intelligence artificielle se positionne comme un levier stratégique incontournable ; elle redéfinit profondément le paysage tactique et opérationnel des Armées. Comparable à des révolutions historiques telles que l'introduction de la poudre à canon ou de l'énergie atomique, elle s'impose aujourd'hui comme une technologie clé. Sur les champs de bataille modernes, l'IA est déjà une réalité et couvre un spectre immense, de l'assistance algorithmique aux décisions stratégiques en état-major, jusqu'aux systèmes d'armes autonomes.

Forte de son héritage scientifique, technologique et militaire, la France s'engage résolument dans cette nouvelle ère. Elle aspire à être le leader européen de l'IA dans le domaine de la défense, en s'appuyant sur une expérience militaire de premier rang en Europe ; une expertise métier et des données opérationnelles autonomes ; une ressource nationale de chercheurs et d'ingénieurs IA de niveau mondial, un tissu industriel foisonnant ; des acteurs de l'armement qui délivrent des capteurs et effecteurs d'une performance reconnue ; une solide infrastructure de calcul, y compris dans des environnements classifiés.

Cette ambition reflète une conviction profonde : l'IA constitue un véritable catalyseur de transformation sur le champ de bataille. Elle accélère les dynamiques opérationnelles, elle confère un avantage décisif à ceux qui sauront l'exploiter avec la plus grande efficacité. Dès lors, elle devient une capacité clef à protéger et une cible potentielle à neutraliser chez l'adversaire.

Cependant, cette révolution technologique s'accompagne de défis majeurs. La transformation digitale et capacitaire est essentielle pour tirer pleinement parti des effets de l'intelligence artificielle. Les armées opèrent encore souvent avec des systèmes d'information et des systèmes d'armes développés sur des technologies et architectures axées d'abord sur la performance individuelle, puis, pour certains, sur la connectivité inter-systèmes. Ces approches, bien qu'efficaces, ne répondent plus aux exigences actuelles. Il est désormais impératif d'évoluer vers des écosystèmes véritablement centrés sur le partage et l'exploitation par l'IA des données multi-domaines.

Le second défi concerne l'urgence de la réponse à apporter sur les champs de bataille. L'IA devient consubstantielle aux trois domaines de lutte Cyber. Son usage proliférant impose un rythme accéléré de montée en puissance de l'acteur étatique pour défendre notre espace informationnel et cognitif. L'avènement de l'automatisation par l'IA des drones et robots constituera à n'en pas douter un enjeu militaire dans les années à venir. Nos soldats seront confrontés à des machines « intelligentes » et nous devons les doter des moyens pour y faire face.

Un troisième défi réside dans la capacité des organisations à intégrer les modalités de conception de l'IA fondées sur l'expérience utilisateur. Cela implique de lier étroitement la préparation de l'avenir à la conduite des activités présentes, pour garantir que les solutions produites soient adaptées aux impératifs opérationnels et ce, dans le strict respect du droit des conflits armés.

Enfin, le dernier défi, porte sur l'alignement des usages de l'IA avec nos valeurs fondamentales, afin d'assurer son acceptabilité et de renforcer la légitimité de son emploi, tant au sein des forces armées que dans l'opinion publique. L'IA doit-elle se substituer au combattant ou simplement l'assister ? Ces interrogations appellent une réflexion collective impliquant industriels, chercheurs, universitaires et acteurs publics. Nous avons pensé, dès les années 60 avec l'atome, la question de la maîtrise souveraine des technologies sensibles. Aujourd'hui, nous devons concevoir nos IA en adéquation avec les règles et les responsabilités liées à l'emploi de la force. Cela passe par une définition claire des cadres et des domaines d'emploi, ainsi que par leur appropriation à tous les niveaux. L'enjeu est de garantir une maîtrise souveraine et responsable de cette technologie tout en répondant aux exigences opérationnelles.

La France ne peut pas avancer seule dans cette transformation technologique. Elle doit inscrire son action dans une dynamique de coopération internationale. Dans un contexte marqué par des tensions internationales croissantes, établir des partenariats solides et durables apparaît comme une nécessité incontournable.

Cet ouvrage, fruit d'un travail collectif, se propose d'explorer ces enjeux à travers quatre axes fondamentaux : stratégique ; opérationnel ; technique ; juridique et éthique. Il constitue un appel à conjuguer innovation et responsabilité pour bâtir une défense française moderne et souveraine. Les réflexions qui y sont développées traduisent une ambition claire : faire de l'intelligence artificielle un levier stratégique au service de notre sécurité collective tout en restant fidèle aux principes qui nous rassemblent.

L'avenir de notre défense repose sur notre capacité à maîtriser et adopter ces technologies sans céder ni à l'angélisme ni au fatalisme. La France continuera à défendre sans concession une approche responsable du développement technologique dans le domaine militaire. Le comité d'éthique de la défense a rendu public son rapport en début d'année 2025. Parmi ses recommandations, je retiens particulièrement que les systèmes à base d'IA obéissent aux mêmes contrôles de licéité que tout système d'armes. L'IA de défense exige de tous, du concepteur, du responsable de la donnée, à l'utilisateur et au chef militaire, une conscience aiguisée de sa responsabilité et une appréhension totale des conditions d'emploi du système. Dans le domaine de la défense, l'IA responsable ou l'IA de confiance sont des truismes en tant que finalité mais qui doivent cependant être documentés et déclinés dans les faits. Ce livre blanc est une étape essentielle dans cette démarche ambitieuse et lucide qui caractérise notre doctrine de défense d'aujourd'hui comme celle de demain.

Bertrand RONDEPIERRE

Directeur de l'Agence Ministérielle pour l'Intelligence Artificielle de Défense (AMIAD)

INTRODUCTION



Karl NEUBERGER

VP Capgemini Invent
France



Olivier DENTI

Directeur Data / IA
Capgemini Invent
France



Ce livre blanc, fruit d'un travail collectif porté par le pôle d'excellence cyber (PEC), se penche sur les multiples enjeux liés à l'intelligence artificielle dans la défense. Il met en perspective les défis techniques, éthiques, juridiques et opérationnels qui accompagnent cette transformation majeure.

À l'heure où les innovations se succèdent à un rythme inédit, il devient essentiel de cerner à la fois le potentiel de l'IA et les responsabilités qu'elle implique. Cet ouvrage propose un parcours à travers 28 articles, chacun offrant un éclairage singulier sur la question de la confiance dans l'IA appliquée à la défense. Des enjeux de qualification et de sécurité aux dilemmes soulevés par les systèmes autonomes, les contributions croisent regards d'experts et retours d'expérience, élargissant ainsi notre compréhension des risques et des opportunités.

Au fil des pages, se dessine une vision concrète de l'IA de confiance, fondée sur la diversité des compétences et la complémentarité des expertises. L'approche pluridisciplinaire qui s'impose vise à garantir des systèmes performants, conformes aux attentes opérationnelles, mais aussi respectueux de nos valeurs et des cadres juridiques internationaux. Cette exigence d'intégration favorise un usage responsable et conforme aux principes démocratiques et au droit international humanitaire.

Pourtant, la spécificité de l'IA réside dans son évolution continue. Systèmes adaptatifs, apprentissage permanent : cette dynamique impose un contrôle constant et transparent, de la conception à l'exploitation. Maintenir la fiabilité et la sécurité des systèmes, même lorsqu'ils évoluent, devient une priorité. La transparence à chaque étape conditionne la confiance des utilisateurs, qu'ils soient décideurs ou opérateurs sur le terrain, et reste incontournable pour l'acceptation de l'IA dans les processus de décision et les



systèmes d'armes. C'est dans cet esprit que s'inscrit la récente création de l'Agence ministérielle pour l'IA de défense (AMIAD).

Par ses analyses et ses recommandations, ce livre blanc reflète la vitalité des réflexions menées au sein de la communauté de l'intelligence artificielle. Il s'adresse à tous ceux qui s'intéressent à l'IA appliquée à la défense et offre des repères pour construire des systèmes performants et dignes de confiance.

Nous adressons nos remerciements à l'ensemble des contributeurs pour leur engagement et leur expertise. Leur mobilisation illustre notre ambition collective de faire de l'IA un levier stratégique, au service de la sécurité et dans le respect des valeurs que nous partageons.

Karl Neuberger et Olivier Denti

Co-animateur du groupe de travail
IA de confiance dans la Défense

L'IA

un enjeu stratégique





L'IA dans les cyberattaques

Un multiplicateur de risques



Frédérique Douzet

Directrice de GEODE et de l'IFG Lab



Le 4 février 2025, à la veille du Sommet pour l'Action sur l'intelligence artificielle de Paris, Google annonce renoncer à son engagement de ne pas utiliser l'IA pour produire des armes. Quelques mois auparavant, en octobre 2024, la maison blanche publiait un mémorandum sur l'intelligence artificielle (IA) affirmant clairement que l'IA est un instrument au service de la défense et de la sécurité nationale, mais aussi au service de la conduite des opérations cyber offensives. Le mémo précise que des capacités seront développées afin d'être testées pour en évaluer les risques. Dans le contexte d'une offensive du vice-président des États-Unis contre la « régulation excessive » de l'IA, ces évolutions laissent présager un développement rapide de l'IA dans les opérations cyber, à des fins militaires ou criminelles.

Un rapport récent d'Open AI ^[1] alerte sur l'accélération de la génération automatisée de codes malveillants à l'aide de l'IA générative, augmentant la portée des cyberattaques. Ces exemples soulignent une réalité : l'IA agit comme un multiplicateur stratégique de puissance, mais aussi de risques, appelant donc une vigilance particulière et un comportement responsable des acteurs.

Multiplicité des risques :

Amplification des cyberattaques :

L'IA amplifie la complexité, la sophistication, la portée et la rapidité des cyberattaques. L'IA permet de gagner en rapidité et en efficacité, d'identifier et d'exploiter plus rapidement les vulnérabilités. Les algorithmes avancés peuvent détecter des failles « zero-day » très rapidement, laissant peu de temps pour réagir. Les conséquences sont immenses, tant dans la sphère civile que militaire. Des infrastructures critiques comme les hôpitaux, les

systèmes d'approvisionnement en eau potable ou les réseaux électriques peuvent être paralysés côté civil, et des systèmes d'armes pourraient être mis hors service ou être infiltré côté militaire. En outre, l'attribution des attaques devient de plus en plus complexe. Les attaquants utilisent l'IA pour masquer leurs traces. Cette difficulté entrave les efforts de riposte et crée un climat d'impunité pour les acteurs malveillants.

Empoisonnement des données :

Un autre risque croissant est celui de l'empoisonnement des données d'entraînement de l'IA. Cette manipulation peut fausser les processus décisionnels de l'IA dans des secteurs critiques tels que la santé, la finance et la défense. Ces résultats erronés peuvent entraîner des pertes financières, une sécurité compromise, voire des pertes humaines.

En parallèle, il est de plus en plus compliqué d'attribuer ces attaques. En effet, les attaques pilotées par l'IA peuvent aider les attaquants à masquer leurs traces, ce qui rend plus difficile l'identification des attaquants et de leurs motivations. La difficulté d'attribution complique les efforts de réponse et de récupération et limite la capacité à tenir les attaquants pour responsables, ce qui peut enhardir les acteurs malveillants.

Prolifération, contagion et risque systémique :

La prolifération d'outils offensifs d'IA augmente également le risque systémique. Les interconnexions complexes des systèmes d'information modernes signifient qu'une attaque ciblée pourrait déclencher des effets en cascade, touchant des réseaux entiers. Ces effets sont aggravés par la rapidité et l'échelle des attaques, rendant la gestion des crises encore plus difficile.



Comment limiter ces risques ?

Quelle voie à suivre ? :

Face à la prolifération des risques liés à l'IA, il est impératif de tracer une voie alliant cadre normatif, coopération internationale et sensibilisation des décideurs. Les pistes à explorer sont multiples et posent des questions difficiles. Comment garder l'humain dans la boucle en tant que responsable moral et légal ? L'IA nécessite-t-elle de préciser ou compléter le droit existant applicables aux cyber opérations ? Quelle est la responsabilité des plateformes privées qui fournissent la puissance de calcul et hébergent les données en cas d'incident ? Pour relever ces défis, trois priorités se dégagent :

Un cadre normatif renforcé : La première étape consiste à renforcer les mécanismes d'encadrement éthique et juridique de l'IA, en s'inspirant des initiatives existantes. Par exemple, la France est le premier pays au monde à s'être dotée d'un Comité d'éthique de la défense en France, qui a ouvert la voie en émettant des avis sur les SALIA (systèmes d'armes létaux intégrant de l'autonomie)^[2]. Ses réflexions méritent d'être étendues à l'ensemble des applications militaires de l'IA, notamment les cyber opérations offensives. La Déclaration politique sur l'utilisation militaire responsable de l'intelligence artificielle et de l'autonomie^[3], signée par 58 États dont la France, constitue également une base utile sur laquelle construire.

Parallèlement, il est crucial de réaffirmer les principes fondamentaux du droit international humanitaire (DIH)

applicable à ces opérations, notamment l'interdiction de cibler des infrastructures civiles. Enfin, ce cadre normatif doit permettre d'établir clairement les responsabilités morale et légale des acteurs impliqués en cas d'usage d'armes autonomes, afin de ne pas les laisser déployer des activités offensives à haut risque en toute impunité. Le 11 février 2025, 27 États ont signé la Déclaration de Paris sur le maintien du contrôle humain dans les systèmes d'armes utilisant l'IA.^[4]

Une coopération internationale accrue :

La France, forte de son engagement dans des initiatives internationales telles que le plan REAIM (Responsible AI in the Military Domain), a la légitimité pour organiser une réunion de haut niveau dédiée à l'encadrement de l'IA. Ce sommet international devra rassembler des États, des experts en intelligence artificielle, des représentants de l'industrie et des spécialistes en droit international, afin d'aborder les implications complexes de l'IA à usage militaire, notamment dans le domaine cyber. Une des priorités sera de traiter de manière conjointe les questions relatives à l'IA civile et militaire. La nature duale de cette technologie implique que des innovations civiles peuvent être détournées à des fins militaires, et vice-versa. Le sommet devra ainsi favoriser une approche globale et intégrée ces enjeux, sans éluder la question spécifique des opérations cyber offensives appuyées sur l'IA, comme cela a été trop souvent le cas jusqu'à présent dans les discussions multilatérales.

Pour retrouver l'intégration de l'avis du comité éthique sur les SALIA : https://www.defense.gouv.fr/sites/default/files/ministere-armees/20210429_Comite_d_ethique_de_la_defense_-_Avis_integrer_autonomie_systemes_armes_letaux.pdf

La souveraineté de l'IA pour la Défense : un impératif stratégique pour une IA de confiance



Olivier DENTI

Directeur Data / IA
Capgemini Invent
France



Camille MAINDON

Consultante cybersécurité
Capgemini Invent
France



Dans un contexte géopolitique marqué par une course technologique intense, la souveraineté en intelligence artificielle (IA) s'impose comme un enjeu crucial pour la Défense nationale. Pour la France, garantir cette souveraineté ne se limite pas à une simple indépendance technologique, mais constitue une condition essentielle pour développer une IA de confiance. Cet article explore les fondements, les défis et les opportunités liés à la souveraineté de l'IA dans le domaine de la Défense.

Qu'est-ce que la souveraineté en IA pour la France ?

La souveraineté en IA peut être définie comme la capacité d'un État à concevoir, déployer et maintenir des technologies à base d'IA de manière autonome, en préservant ses intérêts stratégiques et ses valeurs. Ce concept repose sur plusieurs piliers fondamentaux :

- **La maîtrise technologique** : qui revient à contrôler la chaîne de valeur complète. Cela inclut le développement d'algorithmes, l'exploitation des données et la gestion des infrastructures critiques comme les supercalculateurs.
- **L'indépendance stratégique** : qui vise à limiter toute dépendance vis-à-vis de fournisseurs technologiques non-européens par le biais de normes éthiques et de réglementations strictes. L'objectif est de réduire les risques d'espionnage, d'ingérence ou d'interruption des services critiques.
- **La prospérité économique** : qui s'appuie sur le développement de l'industrie, de la recherche et du marché autour de l'IA en France, garantissant une innovation sur le long-terme et le développement de compétences locales dans les nouvelles technologies.

Compte tenu des tensions géopolitiques et de la dépendance croissante de nos sociétés au numérique (et demain à l'IA), **être souverain est déterminant pour maintenir une confiance nationale en l'IA.**

Quels sont les enjeux d'une IA souveraine pour la défense ?

Pour la France, garantir une IA souveraine pour la Défense implique de maintenir et développer sa puissance stratégique et militaire au niveau international en :

Renforçant sa position sur le marché mondial de l'IA dans le secteur de la Défense et pour des cas d'usage propres aux opérations militaires. Aujourd'hui, ce marché est largement dominé par les Etats-Unis et la Chine qui captent respectivement 56% et 24% des capitaux risqués investis dans l'IA dans le monde ^[1].

- **Protégeant ses opérations militaires** et les systèmes critiques intégrant de l'IA en cas de tensions géopolitiques. Il est essentiel pour la France de pouvoir garantir un accès restreint aux systèmes de commandement et de contrôle, et de garder la main sur les informations stratégiques traitées (données de renseignement, données militaires sensibles). En outre, le pays doit de se prémunir contre des cyberattaques qui pourraient provenir de puissances étrangères vendant des solutions basées sur l'IA, compromettant des opérations militaires.
- **Garantissant la résilience des activités de la Défense**, grâce à une infrastructure nationale sécurisée, mais aussi le développement de solutions résilientes capables de fonctionner dans des environnements dégradés. En ce sens,

l'installation d'un supercalculateur dédié à l'IA au Mont-Valérien illustre cette volonté de renforcer le contrôle sur les données sensibles.

- **Permettant l'acceptabilité sociale** des technologies militaires et en préservant l'image internationale de la France via la promotion **d'une IA éthique et conforme aux valeurs nationales** (respect des droits humains et du droit international humanitaire, non-recours aux systèmes d'armes létales autonomes, etc.).

Le défi pour la France est donc **de maîtriser l'ensemble de la chaîne de production de l'IA**, de la donnée d'entrée à l'utilisateur final, en passant par l'infrastructure et la conception de l'algorithme, afin de garantir la confiance envers cette technologie.

Pour y faire face, le secteur de la Défense français peut capitaliser sur le développement d'un écosystème innovant. Une grande partie des innovations en IA provient du secteur civil (start-ups, PME, laboratoires). Il est donc impératif d'établir une collaboration étroite entre ces acteurs et les forces armées pour maintenir un avantage technologique tout en maîtrisant les systèmes critiques^[2].

Un équilibre entre souveraineté et coopération internationale

Bien que la souveraineté implique une certaine indépendance, elle ne doit pas conduire à l'isolement. La coopération internationale est essentielle pour accélérer les progrès technologiques et renforcer la résilience des systèmes. Dans cette optique, la France a tout intérêt à **développer des partenariats stratégiques et ciblés**, permettant de mutualiser les coûts de R&D sur des briques technologiques non critiques (ex : Fonds européen de Défense) et de collaborer sur des enjeux communs comme la (cyber) sécurité des solutions ou le partage de renseignements.

Cette coopération nécessite néanmoins **d'établir différents niveaux de partage** en fonction de la **sensibilité des informations ou des technologies** ciblées. En effet, si la France gagnerait à partager les bonnes pratiques avec ses voisins européens ou ses alliés de l'OTAN (ex : innovation et efficacité accrue), elle doit cependant maintenir le développement de certaines capacités critiques militaires au niveau national. Afin de faciliter les partenariats stratégiques entre les différentes puissances, une **gouvernance multilatérale** est nécessaire pour définir des normes communes – standards de sécurité, d'interopérabilité, normes éthiques, etc. – concernant le développement et l'usage militaire de l'IA. Conjointement, des **mécanismes de contrôle** mutuels doivent être élaborés tout en

maintenant le dialogue sur les risques et limites à ne pas franchir (ex : systèmes d'armes létales autonomes).

Ainsi, par la coopération internationale, la France peut également s'affirmer comme un acteur majeur du développement d'une IA de confiance dans le secteur militaire.

Conclusion : Une IA de confiance performante, éthique et souveraine

La souveraineté en IA pour la Défense représente un défi national majeur qui conditionne directement l'autonomie stratégique et la supériorité opérationnelle des forces armées françaises. Alors que l'IA devient un facteur déterminant de performance pour la puissance militaire, il est impératif que la France continue d'investir dans cette technologie tout en renforçant ses partenariats européens.

Pour garantir une IA de confiance, il ne suffit pas seulement de développer des solutions performantes ; elles doivent également être alignées sur les valeurs nationales et protégées contre toute ingérence extérieure. En adoptant une approche équilibrée entre autonomie nationale, maîtrise de la chaîne de production et coopération internationale, la France pourra non seulement préserver sa souveraineté technologique mais aussi affirmer son rôle de leader dans l'innovation pour la Défense du XXI^e siècle.

Coopération et confiance dans les projets complexes



Jean-Siri Luang Aphay

Docteur en sciences de l'information et de la cognition,
Directeur au sein du cabinet Formind



La réussite des projets complexes n'est jamais un acquis a priori. Et lorsque l'on considère la conduite des projets informatiques, les taux d'échec augmentent fortement. Si la grande majorité des projets aboutit à des livraisons, les critères amenant à considérer la satisfaction ou l'échec sont de natures diverses : coûts, délais, fonctionnalités, périmètres, cible de marché, et bien entendu état de sécurité dans la durée.

Ce constat augmente avec le temps, en se confrontant aux usages métiers, mais également à l'épreuve des exigences d'exploitation (maintenance, évolutions, ...). Ainsi le bon fonctionnement des systèmes d'information doit beaucoup à la cohérence de l'ensemble des exigences prises en conception. Cette exigence de cohérence est multipliée dans le cadre de l'intelligence artificielle, tant dans les dépendances technologiques, techniques et fonctionnelles, que dans les conséquences cognitives, fonctionnelles et sociales. Le principal levier de cohérence est la coopération des parties prenantes engagées. Par coopération, on entend l'engagement réciproque des acteurs et non uniquement des successions ordonnées de livrables. Bien que tous les standards de pilotage de projet s'assurent de l'agencement des contributions, l'échec des projets matérialise les divergences d'intérêts, de références, de méthodologies, de temporalité et de représentations des parties prenantes.

En termes de gouvernance, deux facteurs de risque infléchissent fortement la trajectoire des projets.

Le premier est le postulat des prédispositions constructives des acteurs comme exclusivement convergentes. Leurs disponibilités, approches

méthodologiques, critères de qualité et satisfaction, exigences technologiques, ... peuvent comporter des écarts avec les exigences du projet. Chaque partie prenante s'inscrit individuellement dans un jeu stratégique consistant à optimiser ses contributions et ses intérêts particuliers.

Ce jeu produit des arbitrages dont les effets de bord sont immédiats ou ultérieurs, perceptibles ou insidieux. Le second facteur est le postulat d'une confiance par délégation satisfaisante.

La confiance est un gage de fiabilité face au risque. Elle permet de s'engager dans une relation à risques sans remettre en question tous les termes de la maîtrise des risques. Elle se construit, intuitivement, méthodologiquement ou institutionnellement. Dans ce dernier cas, il s'agit d'une confiance par délégation. L'ensemble des conditions d'attribution de la confiance n'est pas maîtrisé, mais la confiance dans le système, les structures et les institutions qui fondent la relation nous semblent établis. Ainsi en va-t-il du chirurgien qui opère, du notaire qui conserve les titres de propriétés ou de la banque qui gère l'argent.

La confiance accordée aux systèmes d'informations repose sur la délégation de confiance à un système d'ingénieries successives ordonnées. Cette confiance comprend deux biais. Le premier est introduit par la dépendance aux services informatiques et l'absence de choix éclairé. Le second est l'idée a priori que ces systèmes sont judicieusement gérés et l'incapacité de discuter des garanties proposées. La complexité des systèmes d'information est parfois telle qu'aussi bien les dirigeants que les utilisateurs ne

peuvent discuter des termes de la confiance accordée. Tout système élaboré par un groupe d'experts dont les domaines d'application sont inaccessibles au commun peut donc aisément susciter la confiance puisqu'il n'existe aucun espace de questionnement de cette confiance.

Par leur ampleur, les projets d'intelligence artificielle justifient la mise en place de processus de coopération dédiés. La coopération ne limite pas à suivre un calendrier précis, mais porte sur la construction d'une collégialité et une compréhension commune entre tous les acteurs du projet de l'écosystème. Cette nécessité pour les systèmes d'information était renvoyée aux pratiques de management. L'expérience, du simple prisme de la cybersécurité, suggère que ces pratiques sont insuffisantes pour assurer la cohésion. Cette nécessité est formalisée pour la première fois de façon insistante dans le document du NIST sur l'IA ^[1]. Il anticipe la portée des risques technologiques et sociaux de l'intelligence artificielle et insiste sur la cohérence politique des projets – au sens des régulations de la communauté IA engagée.

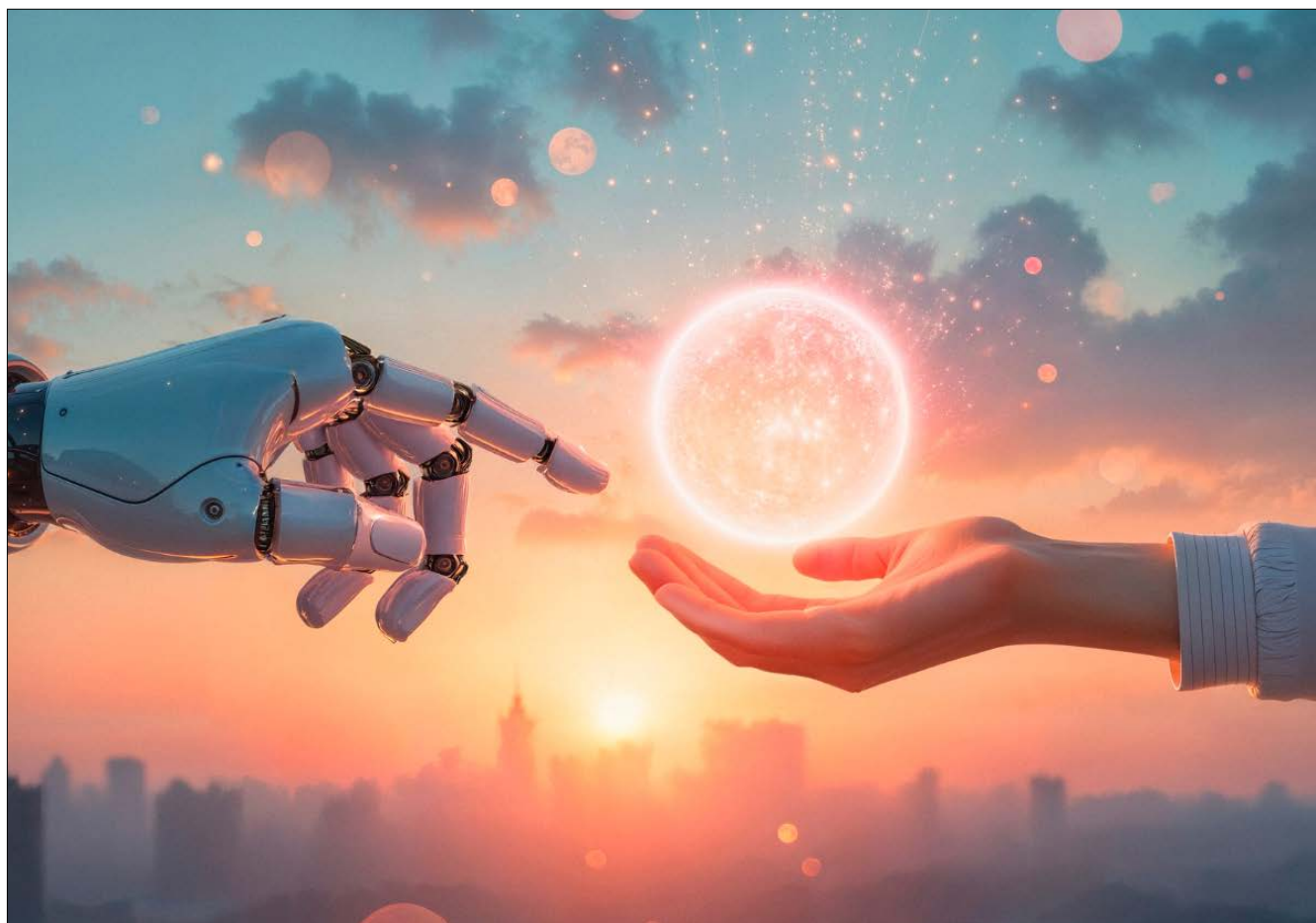
L'état d'un système – un projet de solution d'intelligence artificielle est un système socio-technique – est le résultat de la qualité de ses relations sociales et organisationnelles.

En résumé, un processus de coopération comprend trois axes. Le premier est la connaissance des écarts que les jeux stratégiques génèrent par rapport aux objectifs initiaux. Le second porte sur l'identification des leviers de régulations sociales. La hiérarchie, la contractualisation et les réunions ne sont pas des leviers suffisants, ils sont les supports de ce jeu stratégique.

Le troisième levier est la construction de représentations communes. Un socle commun d'évidences est l'ensemble des représentations partagées qu'il n'est pas nécessaire de questionner. Dans des projets mobilisant de très fortes expertises, le cloisonnement des représentations peut rapidement amener chacun à être confiant sur son propre champ d'application, sans pour autant que cela ne soit évident pour les autres personnes qui en dépendent.

Cet inventaire des forces se matérialise en une cartographie des objectifs individuels et partagés, des cadres de références et de valorisation de chacun, des contraintes opérationnelles respectives et des contournements possibles.

La confiance ainsi construite au sein du projet est alors un gage transmissible auprès des dirigeants et utilisateurs.



Les enjeux et perspectives de l'IA de défense à l'échelle



Majda Belhaj

Directrice Data et IA
Capgemini Invent



Florian Canderlé

Senior Manager
Capgemini Invent



Le développement de l'IA de défense à l'échelle représente la capacité à déployer des solutions d'IA non pas comme des prototypes isolés, mais comme des systèmes robustes, intégrés et utilisables de nombreux utilisateurs sur le théâtre d'opérations. Contrairement à des IA simples, implémentables rapidement dans un cadre restreint, l'IA de défense à l'échelle exige une industrialisation et une homologation des briques technologiques pour les rendre fiables face à des conditions réelles et changeantes, souvent imprévisibles. Cependant, ce déploiement nécessite de surmonter les contraintes spécifiques à ce secteur telles que la sensibilité, la disponibilité et le partage des données entre l'État et les industriels, la maturité des architectures de défense, le manque d'acteur technologique majeur Européen, l'intégration dans des systèmes complexes ou encore les contraintes de recrutement et les réglementations internationales. Nous proposons dans cet article des pistes d'accélération sur lesquels travailler pour sécuriser le passage à l'échelle.

Vision stratégique

Pour commencer, il est essentiel de définir une vision claire des grandes thématiques d'IA sur lesquelles investir afin de donner un cap à l'entreprise, par exemple autour des drones, de la guerre électronique, du renseignement, des centres de commandement, ou encore de la logistique. Impliquer les Comités de Direction (CODIR) et Comités Exécutifs (COMEX) dans ce choix permet de recueillir leur vision à court, moyen et long terme. La vision doit inclure des métriques pour évaluer les gains attendus et le retour sur investissement (ROI). Une trajectoire

budgétaire ambitieuse sur plusieurs années doit ensuite être sécurisée, couvrant les investissements en ressources humaines, technologies, accompagnement au changement, et justifié par les ROI escomptés.

Modèle opérationnel

La définition d'un modèle opérationnel (organisation, gouvernance, processus, indicateurs) adapté est crucial. Plusieurs critères doivent être pris en compte tels que la taille de l'organisation, le volume et la typologie de cas d'usage, le niveau de maturité en IA et la culture d'entreprise. Si le modèle Hub-and-Spoke^[1] est de plus en plus privilégié, il convient de noter que nous observons plutôt des modèles centralisés pour des IA factory dans le secteur de la défense, probablement du fait de la maturité et des contraintes évoqués précédemment.

Méthodologie agile appliquée à l'IA

Le développement dans un mode agile est essentiel pour co-construire avec les utilisateurs finaux, améliorer la performance des modèles par itérations successives et assurer un fonctionnement en boucle courte. Néanmoins, les contraintes d'homologation propres au secteur de la défense nécessitent de bien prévoir les niveaux de maturité attendus, par exemple sur la documentation des besoins et critères d'acceptation, les exigences de cybersécurité, la qualité du code ou encore les contraintes de performance.

Plateforme technologique souveraine

La plateforme technologique de développement IA dans la défense doit être souveraine pour limiter la dépendance des systèmes critiques envers des acteurs technologiques étrangers (par exemple la réglementation américaine ITAR ^[2]), de maîtriser les données et de sécuriser les systèmes d'information et de communication. Il convient d'évaluer la maturité des capacités technologiques actuelles et de définir une trajectoire d'évolution pour développer de nouveaux cas d'usage (par exemple autour de LLM, avec les contraintes technologiques associées), tout en intégrant les contraintes de cybersécurité dès la conception des solutions.

Gestion et gouvernance des données

L'accès et la gestion des données est un pilier central pour le succès des projets d'IA dans la défense, notamment autour de deux aspects. Premièrement, la collecte, le stockage et le partage dans des environnements contraints, nécessite de prendre en compte l'évaluation de leur sensibilité (conformément aux réglementations en vigueur, par exemple Spécial France ou Secret) et le traitement adéquat pour assurer leur sécurité. Deuxièmement, la qualité de l'annotation et la constitution de jeux d'entraînement et de test représentatif de la réalité opérationnelle dans un contexte frugal demeure primordial. Il est donc clé d'avoir un lien en boucle courte avec les utilisateurs finaux au sein des forces.

Formation et développement des compétences

L'IA est une révolution technologique qui effraie autant qu'elle passionne. L'identification des populations militaires et civiles à former et à sensibiliser demeure un facteur clé de succès. Des parcours de formation doivent être définis selon les besoins et profils pour faciliter l'adoption de cette technologie. Des sessions de sensibilisation aux technologies de l'IA doivent être mises en place pour tous, des opérationnels aux membres des CODIR et COMEX, dans les entreprises privées comme dans les forces armées. La formation doit également inclure des aspects spécifiques à la défense, tels que l'éthique de l'IA propre à la doctrine française, la cybersécurité ou encore la gestion des risques.

Conclusion

De nouvelles dynamiques sont en œuvre autour de l'IA de défense générant aussi davantage d'inquiétudes sur les risques associés : l'accélération et le passage à l'échelle ne sont pas antinomiques à la notion de confiance mais impose une rigueur accrue et une

industrialisation de l'IA de confiance tout au long de son cycle de vie. Des outils dédiés à l'IA de confiance existent et contribuent fortement à industrialiser l'audit et la mise sous contrôle les IA lorsque leur volume devient important. Étant donné le contexte politique actuel, le choix s'impose, les acteurs publics et privés de défense françaises et européennes doivent accélérer le déploiement de l'IA à grande échelle pour sécuriser l'avenir de la défense européenne souveraine.

L'enjeu n'est pas l'automatisation de la guerre mais la distanciation des combattants



Hubert ETIENNE

Chercheur en éthique de l'IA
Président de Quintessence AI
Président fondateur de la Paris Conference on AI & Digital Ethics
Enseignant à l' ESCP Europe et à HEC Paris



L'entrée de l'IA dans le domaine de la défense soulève des questionnements éthiques traditionnellement appréhendés sous le prisme de l'automatisation de la guerre. Exhumant l'antique théorie de la guerre juste, on se questionne ici sur l'acceptabilité morale qu'une machine assassine un homme au regard de la dignité humaine, là sur l'impact du déploiement de systèmes d'armes létales autonomes (SALA) sur le respect du droit humanitaire international.

D'aucuns soulignent qu'un drone de combat ne saurait torturer ni violer, d'autres qu'il peut perdre le contrôle et semer le chaos sans distinction entre les cibles. De facto, nos espoirs de guerres « propres » fondés sur une meilleure discrimination entre civils et combattants furent froidement douchés par les « Drones papers^[1] » dans lesquels on apprenait notamment qu'un sixième des personnes tuées par des frappes de drones américains aux Yémen en 2011 et 2012 étaient des civils.

Lorsque ces considérations éthiques se présentent sur la scène internationale, c'est sous la forme d'un débat entre ceux qui entendent interdire les SALA et les partisans de leur régulation. Derrière le jeu de dupes qui se joue à Genève, les débats opposent les États capables de tirer profit de ces armes à ceux qui en souffrent la menace. La vidéo d'autonomousweapons.org mettant en scène des essaims de drones tueurs à reconnaissance faciale^[2] divisa ainsi les audiences, provoquant l'horreur de certains autant que la fascination d'autres. C'est précisément d'émotions dont je souhaiterais traiter ici car, à défaut de trancher le débat sur l'utilisation des SALA, c'est l'accroissement de la distance affective

entre les combattants que je souhaiterais souligner dans le processus d'automatisation de la guerre.

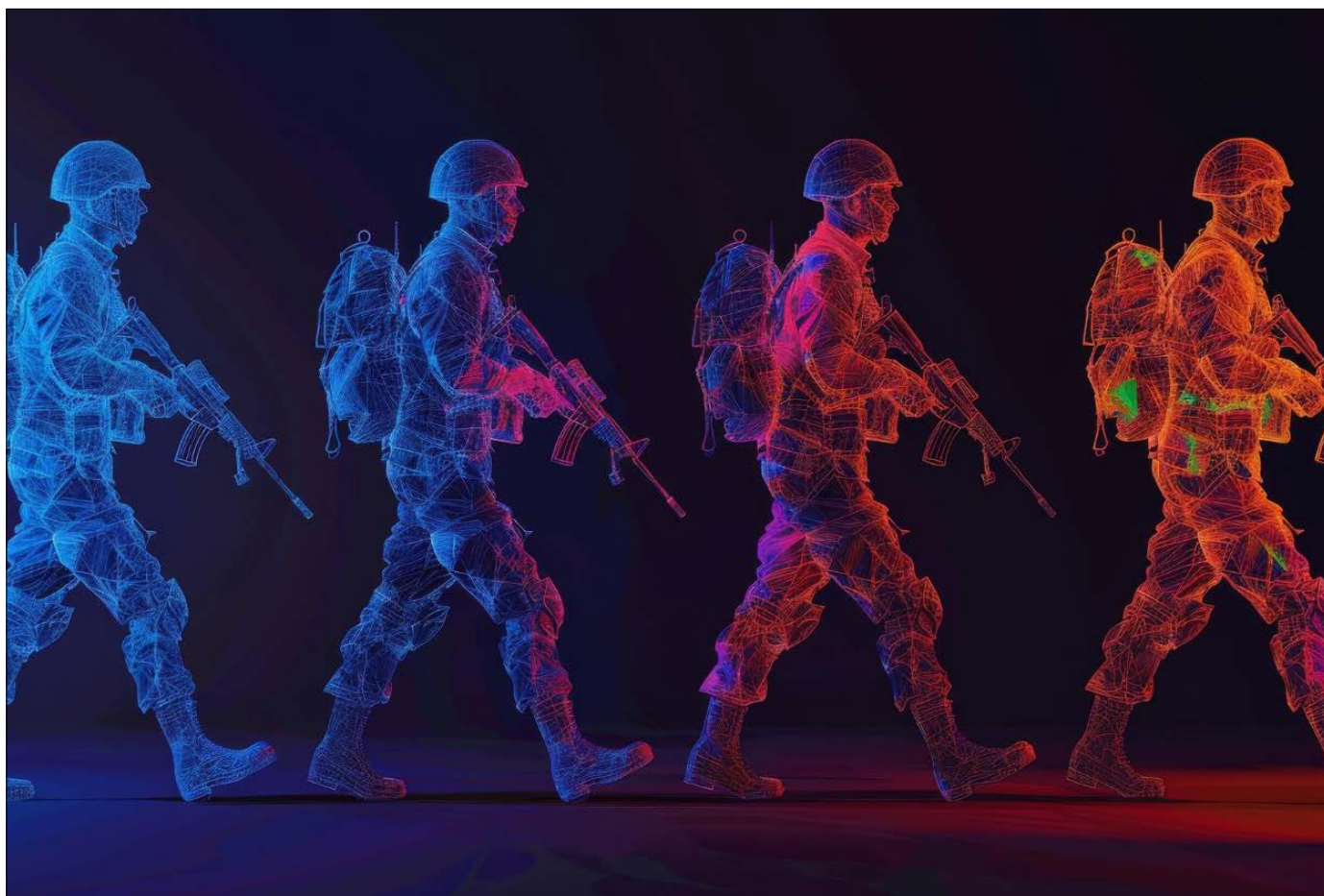
Il y a un monde entre le poilu qui enfonce sa baïonnette dans la poitrine d'un ennemi de tranchée et le pilote d'un drone semant le chaos du bout des doigts à douze mille kilomètres d'allonge. La première situation met l'homme face à l'homme et les affects sont si bien partagés que chacun reconnaît le choix critique qui s'impose à lui dans les yeux de son vis-à-vis : c'est lui ou c'est moi. Dans la seconde, un opérateur déclenche une pression de l'index sur l'ordre d'un supérieur hiérarchique. La scénarisation incite à scinder l'agent moral pour diluer sa responsabilité, celui qui déclenche l'action n'étant pas celui qui prend la décision. Le meurtrier ne voit pas le visage de sa cible dont la silhouette pixellisée lui est désignée par le recoupement de signaux tels que des modèles de reconnaissance vocale et faciale à un degré d'erreur près. Le général François Lecointre confiait récemment les propos suivants : « ce que j'ai expérimenté moi dans le combat au corps à corps, c'est que la proximité avec la personne que l'on doit tuer rend cet acte très difficile. Parce que vous voyez le visage de votre ennemi et que le visage, c'est ce que dit Levinas, vous dit que vous ne pouvez pas transgresser le tabou absolu de la mort de votre prochain [...] Le paradoxe c'est que ce qui humanise le combat c'est de se confronter à ce tabou et d'accepter de le transgresser mais que l'on ne parvient à le faire qu'en laissant se déchaîner la part de bestialité qu'on a en soi ^[3]. »

La distance dont il est question n'est pas seulement spatiale mais d'abord affective. Pour faciliter le meurtre

et en atténuer le traumatisme, on peut moduler la perception de l'acte afin qu'il ne soit pas assimilé par son auteur comme un meurtre de sang-froid. L'armée américaine n'hésite pas à gamifier la guerre, valorisant les compétences des gamers pour piloter ses drones et usant de jeux vidéo comme outils de recrutement^[4]. On peut aussi moduler la perception de l'ennemi pour que le combattant ne soit pas tenté d'y reconnaître un alter ego mais l'instance d'une autre race ou espèce à laquelle il ne saurait s'assimiler, donc projeter sur lui son empathie. L'épisode « Men Against Fire » de la série Black Mirror illustre la manière dont la technologie permet d'ériger une telle distance affective en dépit de la proximité spatiale : grâce à un implant cérébral, les soldats perçoivent leurs ennemis comme des monstres et leurs supplications comme des cris terrifiants.

Dès lors, il ne s'agit pas de glorifier la guerre barbare de nos ancêtres mais de saisir ce que la médiation technique introduit progressivement, du couteau au fusil et du fusil au drone. Plutôt que de débattre sur la

pseudo autonomie de la décision d'un tir léthal et de se rassurer du maintien de « l'humain dans la boucle », c'est le type d'humain qui émerge de cette boucle qu'il faut interroger et les répercussions anthropologiques de cette distanciation affective par-delà le champ de bataille.



L'IA de Confiance dans la Défense : Opportunités et Défis



Rachid El Haddari

Head of Defense and Technology department
École Centrale d'Électronique (ECE)



Autrefois considérée comme un actif stratégique sous-exploité, la donnée a connu une transformation radicale grâce à l'intelligence artificielle (IA). Aujourd'hui, la donnée, à travers l'IA, représente non seulement une opportunité fantastique de développement, mais aussi une arme puissante capable par exemple de manipuler les populations^[1] et conférer un avantage décisif dans les conflits armés.

L'intelligence artificielle (IA) de confiance dans le domaine de la **défense** est un concept essentiel qui vise à intégrer des technologies avancées tout en garantissant leur sécurité, leur fiabilité et leur éthique. Cette approche est cruciale pour améliorer l'**efficacité** des opérations militaires et assurer la **souveraineté** technologique des nations.

Introduction à l'IA de Confiance

L'IA de confiance se réfère à l'utilisation de systèmes d'intelligence artificielle qui sont non seulement **performants**, mais aussi **sûrs**, fiables et **éthiques**. Dans le contexte de la **défense**, cela signifie que les technologies d'IA doivent être capables de traiter des **données sensibles**, de prendre des **décisions critiques** et de fonctionner dans des environnements complexes, tout en **minimisant les risques d'erreurs** et en respectant les principes **éthiques**.

Traitement des Données

L'un des principaux avantages de l'IA dans la défense est sa capacité à traiter **rapidement** et **efficacement** de **grandes quantités de données**. Les sources de données peuvent inclure des satellites, des radars, des drones, des capteurs au sol, et bien d'autres. L'IA peut

analyser ces données pour extraire des informations pertinentes, identifier des modèles et fournir des renseignements exploitables aux décideurs militaires. Par exemple, dans une opération de surveillance, l'IA peut analyser les images capturées par des drones pour détecter des mouvements suspects ou identifier des cibles potentielles. Cette capacité à traiter des données en **temps réel** permet aux militaires de réagir plus rapidement et de manière **plus informée**.

Automatisation des Tâches

L'IA permet également d'**automatiser** certaines tâches, ce qui **libère** les soldats pour des missions plus complexes et **stratégiques**. Par exemple, des systèmes d'IA peuvent être utilisés pour la reconnaissance de cibles, la planification de missions, la gestion logistique, et même la maintenance prédictive des équipements. L'automatisation peut améliorer l'**efficacité opérationnelle** et réduire les risques pour les soldats en les éloignant des situations dangereuses. Par exemple, des robots équipés d'IA peuvent être envoyés pour désamorcer des explosifs ou explorer des zones dangereuses avant que les troupes humaines n'interviennent.

Sécurité et Fiabilité

La sécurité et la fiabilité sont des aspects cruciaux de l'IA de confiance. Les systèmes d'IA doivent être conçus pour minimiser les risques d'erreurs et **garantir que les décisions prises sont justes et éthiques**. Cela implique des tests rigoureux, des validations continues et des mécanismes de surveillance pour détecter et corriger les anomalies. Il est également essentiel de garantir que les systèmes d'IA sont résistants aux cyberattaques.

Dans le domaine de la défense, une cyberattaque réussie (en altérant par exemple ses données d'apprentissage) contre un système d'IA pourrait avoir des conséquences catastrophiques. Par conséquent, des mesures de sécurité robustes doivent être mises en place pour protéger ces systèmes contre les menaces potentielles.

Éthique et Transparence

L'**éthique** est un autre pilier de l'IA de confiance. Les décisions prises par les systèmes d'IA doivent respecter les **principes éthiques** et les **droits de l'homme**. Cela inclut la transparence dans le fonctionnement des algorithmes, la responsabilité des décisions prises et **la garantie** que l'IA ne discrimine pas ou ne cause pas de préjudices injustifiés. Cependant, l'utilisation de l'IA dans la défense soulève également des questions éthiques et des risques importants. Les systèmes d'armes autonomes, par exemple, posent des défis en matière de responsabilité et de contrôle humain [2]. Il est essentiel de garantir que ces technologies respectent les principes du droit international humanitaire et d'éviter toute dérive éthique [3].

Souveraineté Technologique

La **souveraineté technologique** est un aspect important de l'**IA de confiance dans la défense**. Il est crucial que les nations développent leurs **propres technologies d'IA pour ne pas dépendre d'autres pays**[4]. Cette indépendance technologique est essentielle pour **garantir** la sécurité nationale et **protéger les intérêts stratégiques**. Le développement de technologies d'IA nationales permet également de mieux contrôler la qualité et la sécurité des systèmes utilisés. Les pays peuvent ainsi s'assurer que leurs systèmes d'IA respectent les normes de sécurité et d'éthique qu'ils ont défini.

La France, par exemple, a mis en place une stratégie ambitieuse pour devenir un leader mondial de l'**IA de défense**, avec la création de l'Agence ministérielle pour l'IA de défense (Amiad [5]) et le développement de supercalculateurs[6] pour traiter les **données militaires**.

De l'importance de la formation

Enfin, la formation et l'éducation sont des éléments clés pour assurer une utilisation responsable et efficace de l'IA. Il est nécessaire de former les militaires et les professionnels de la défense aux technologies IA, mais aussi de sensibiliser la population aux enjeux de cette révolution technologique.

Des initiatives comme l'intégration de cours d'**IA dans les programmes scolaires** sont essentielles pour préparer les générations futures aux défis de l'ère numérique [7].

Conclusion

L'IA de confiance dans le domaine de la défense est une approche qui vise à intégrer des technologies d'intelligence artificielle avancées tout en garantissant leur sécurité, leur fiabilité et leur éthique. Cela implique le traitement efficace des données, l'automatisation des tâches, la garantie de la sécurité et de la fiabilité des systèmes, le respect des principes éthiques et la souveraineté technologique.

La formation et l'éducation joueront un rôle déterminant dans cette transition vers une défense intelligente et éthique.

Je terminerai en adressant un message à nos jeunes : formez-vous à l'IA ! La France a besoin de vous !

Vers une opérationnalisation de

L'IA

sur le champ de bataille



Industrialisation de l'IA, un pilier de la confiance



Aurélia NÈGRE et l'équipe IA

Head of AI Engineering
Safran.AI



L'émergence des systèmes d'IA fondés sur l'apprentissage profond dans des domaines critiques introduit de nouveaux défis en matière de vérification, de test et d'audit. Ces modèles, par leur nature même, présentent des comportements plus complexes à analyser que les systèmes experts traditionnels.

La confiance vis-à-vis d'un système d'IA nécessite deux piliers : itérer avec les utilisateurs finaux, depuis la conception jusqu'aux tests sur données réelles, et intégrer les meilleures pratiques du développement logiciel, avec une traçabilité rigoureuse des données et des expérimentations IA. Cette approche collaborative, couplée à des méthodes de travail industrielles, permet à la fois de renforcer la sécurité de ces systèmes et de garantir leurs performances opérationnelles.

Nous partageons ici cinq enseignements tirés de nos huit ans d'expérience sur des produits logiciels incorporant de l'IA en production dans des environnements contraints.

1. Placer l'utilisateur final au centre

Toute IA de défense est utilisée par des forces armées professionnelles, expertes dans leurs domaines respectifs. Pour prendre des décisions éclairées à partir des systèmes d'IA (SIA), il est fondamental que ces systèmes correspondent parfaitement au cas d'usage opérationnel et que les utilisateurs finaux en comprennent finement les capacités, le périmètre d'utilisation et les limites.

Pour cela, l'intégration des utilisateurs dès les phases initiales de conception est cruciale. Cette collaboration précoce permet de définir le périmètre de fonctionnement opérationnel, et oriente la création des jeux de données d'apprentissage et de test par les Data Scientists. Elle

permet également de cibler les cas d'usages où l'IA peut apporter de la valeur, faire comprendre à l'utilisateur final les points forts et les erreurs classiques des SIA. Au cours de leur durée de vie, les algorithmes feront l'objet de mises à jour régulières, en se fondant sur les retours des utilisateurs et sur les données réelles. En effet, il est rare de disposer lors de la phase de conception, de données exhaustives et représentatives des conditions finales d'utilisation. Cela est particulièrement vrai pour les cas d'usage défense. Ces boucles courtes d'itération permettent alors d'atteindre de façon pragmatique la meilleure performance opérationnelle, et sont des facteurs clé d'adoption.

Enfin, l'autonomie des utilisateurs dans l'évaluation des systèmes représente un facteur clé de confiance. Des protocoles et outils spécifiques doivent leur permettre de tester les modèles sur leurs propres données opérationnelles, associés à une formation sur l'interprétation des métriques d'évaluation.

2. Obtenir des données annotées de qualité

Les systèmes d'IA fondés sur l'apprentissage profond ont besoin d'importants volumes de données annotées pour être performants. Or, pour les cas d'usage défense, les données opérationnelles sont généralement classifiées, et peu disponibles.

Bien que l'utilisation de données en source ouverte soit un bon moyen de prototyper des premières versions d'algorithmes, un SIA en production nécessite d'obtenir des données réelles ou a minima très similaires. Cela implique d'investir dans des données commerciales d'intérêt et représentatives.

Elles seront utilisées pour constituer les jeux de données d'apprentissage, pour définir un processus d'annotation rigoureux incluant l'expertise métier des forces et pour les tests en conditions réelles.

Lors de la conception, une attention particulière doit être portée sur l'écart entre les données d'entraînement et les données réelles. L'utilisation de données souveraines classifiées suppose de disposer d'un système d'information homologué pouvant les héberger.

Lors du développement, la traçabilité constitue un pilier fondamental de l'IA de confiance. Chaque jeu de données doit être rigoureusement tracé et accessible, avec une séparation claire entre les données d'entraînement et celles servant de validation et de test.

3. Assurer une traçabilité des expérimentations

L'historique complet des expérimentations menées par les Data Scientists doit être conservé et indexé. Ces explorations, même infructueuses, font partie intégrante du processus de développement et peuvent s'avérer précieuses pour la compréhension des choix effectués et les travaux ultérieurs. Un registre centralisé des modèles, associant algorithmes et métadonnées, permet d'éviter toute désynchronisation. Cette documentation exhaustive garantit la reproductibilité de chaque algorithme, condition indispensable pour :

- Conduire des audits approfondis ;
- Prévenir les contaminations entre jeux de données ;
- Évaluer l'impact d'anomalies découvertes ultérieurement ;
- Permettre la reproduction des algorithmes ;
- Faciliter l'amélioration des modèles.

4. Industrialiser les processus de développement

L'automatisation des tâches répétitives et manuelles réduit le risque d'erreur et améliore la vélocité de conception des systèmes d'IA. Elle incite à la standardisation et permet au Data Scientist de concentrer son effort sur la démarche scientifique. Il reste chargé de la sélection et de la préparation des données, de définir la stratégie algorithmique, d'analyser qualitativement les résultats et d'explorer de nouvelles méthodes ou outils permettant de repousser les limites de performances.

Dans le domaine de la défense, la robustesse du code est cruciale, Étant donné la complexité et le coût élevés des redéploiements (les solutions sont souvent hébergées on-premise), il est essentiel de mettre en place, en amont, une stratégie de tests exhaustifs couvrant l'intégralité des composants logiciels.

5. Garantir la performance opérationnelle

Le secteur de la Défense exige des performances optimales. Les protocoles de test, définis en collaboration étroite

avec l'utilisateur final, constituent un facteur de confiance essentiel. Ces tests, automatisés et diversifiés, incluent :

- L'évaluation des métriques de qualité (F1, précision, rappel, etc.)^[1] sur une large base de test représentative de l'utilisation opérationnelle de l'outil ;
- La vérification des cas limites et des jeux de test spécialisés par contexte d'utilisation ;
- Le contrôle des ressources matérielles (a minima RAM^[2] et VRAM ^[3]) et des temps de calcul ;
- Des évaluations qualitatives ciblées.

Chaque amélioration de l'algorithme d'IA fait l'objet d'une évaluation sur le nouveau périmètre ainsi que de tests de non-régression. Ces cinq enseignements sont au cœur de notre méthodologie de développement, et irriguent à la fois les expérimentations algorithmiques, mais aussi les investissements dans l'outillage et nos échanges avec les utilisateurs. Ils permettent ainsi de fournir des systèmes d'IA à forte valeur ajoutée et offrant une grande confiance à l'utilisateur.

Construire la confiance en l'IA : la nécessité de la certification. Le point de vue d'un CESTI



Jean-Léon Cusinato

Evaluateur CESTI logiciel
Amossys



Un CESTI (Centre d'Évaluation de la Sécurité des Technologies de l'Information) est un laboratoire d'évaluation de logiciel agréé par l'ANSSI. L'objectif de cette structure est de vérifier que les produits analysés sont résistants à un attaquant tel que défini dans le schéma de certification au travers d'une évaluation. Cette vérification, appelée évaluation, est une tâche complexe qui devient encore plus ardue lorsqu'il s'agit de produits intégrant des réseaux de neurones – une branche de l'Intelligence Artificielle actuellement en plein essor. Ces derniers sont de plus en plus utilisés dans divers domaines, allant de la reconnaissance d'images à la prise de décision autonome.

En tant que CESTI, évaluer ces produits présente aujourd'hui des défis uniques. Chaque fonction d'un logiciel accessible par un attaquant représente un risque pour la sécurité. Par ailleurs s'assurer du niveau d'alignement ^[1] et de l'absence de biais d'une intelligence artificielle est également un sujet majeur mais restent actuellement un sujet de recherche. Cet article traite de l'adaptation du processus d'évaluation pour tenir compte de cette évolution des produits intégrant de l'intelligence artificielle, sans pour autant nécessiter de vérifier l'alignement des modèles.

Problématique

Intégrer une intelligence artificielle dans un logiciel requiert de déléguer une partie du traitement des informations à un réseau de neurones. Cette approche permet d'étendre les fonctionnalités du produit, mais entraîne l'opacité du traitement aboutissant aux résultats. Un réseau de neurones doit être fiable sur plusieurs points en ^[2] :

- Retournant des résultats qui ne reproduisent pas les biais présents naturellement dans les données

produites par l'être humain ;

- Prenant des décisions en accord avec les objectifs et l'éthique des utilisateurs qui manipulent ce réseau de neurones.

Malgré ces problématiques, l'évaluation des produits contenant des réseaux de neurones est cruciale pour garantir leur sécurité et leur fiabilité. En identifiant les potentielles vulnérabilités, les CESTI contribuent à la création de systèmes plus robustes et moins susceptibles de subir des attaques. De nombreuses industries sont soumises à des réglementations strictes en matière de sécurité des technologies de l'information et la mise en place d'une méthodologie d'évaluation est une étape nécessaire dans la réalisation des analyses de ces nouveaux produits basés sur des réseaux de neurones.

Positionnement de l'évaluation

L'intelligence artificielle est généralement intégrée dans des programmes de multiples usages : prendre des décisions, générer du contenu, interpréter du langage naturel, etc. Les résultats produits par une intelligence artificielle sont, à un moment donné, traités par un logiciel appelant, et c'est précisément à cette étape que se situe le plus grand risque. En effet, le programme ne peut pas considérer les données générées comme étant entièrement fiables, en raison du caractère non déterministe des résultats. Même en l'absence de biais intentionnel dans le produit, des contrôles peuvent être effectués en amont, mais ils sont impérativement nécessaires en aval du réseau de neurones.

Solutions actuelles

Face aux défis complexes posés par l'évaluation des produits contenant des réseaux de neurones, plusieurs solutions sont abordées pour améliorer leur sécurité et leur fiabilité. Des recherches sont actuellement en cours^[3] développer des méthodes rendant les modèles plus transparents et interprétables. Cette approche implique une introspection du réseau de neurones afin de quantifier ses capacités d'interaction, sans pour autant décomposer son fonctionnement interne. Une telle analyse est essentielle dans le cadre de l'évaluation effectuée par un CESTI sur ces technologies.

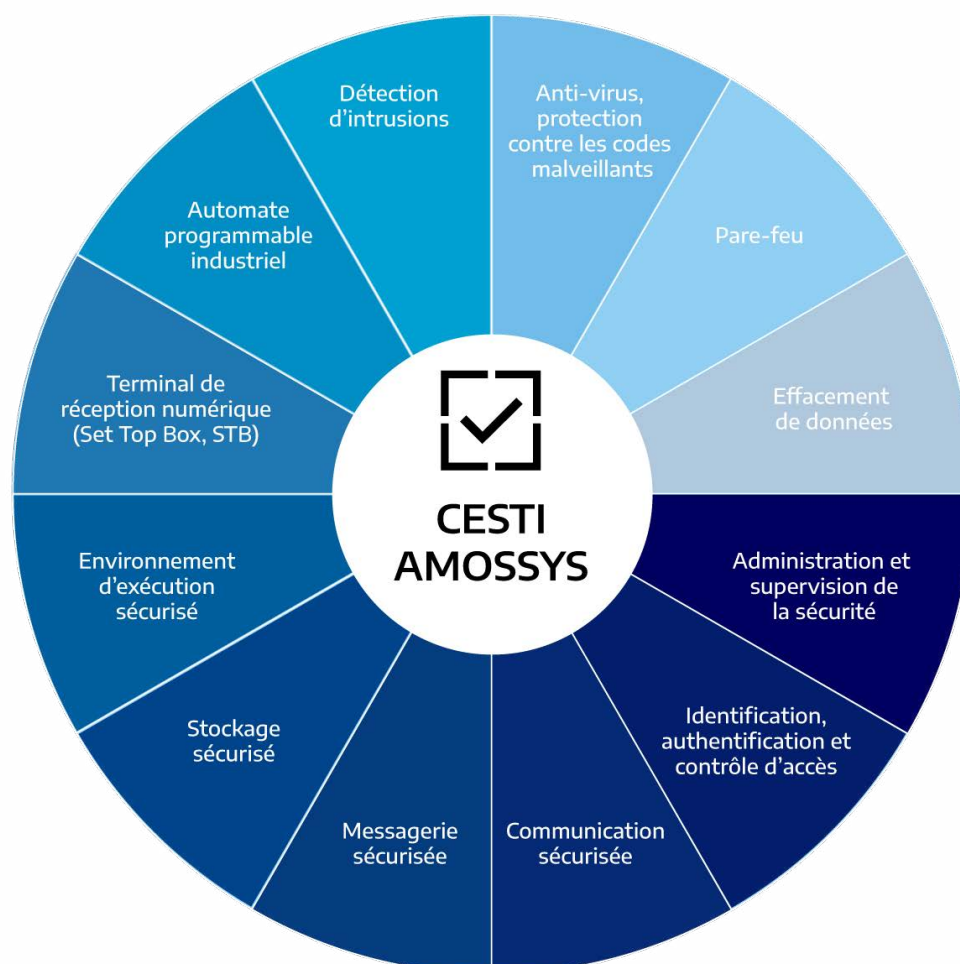
La construction de plateformes de simulation d'attaques permettra, à l'avenir, de tester la robustesse d'un produit face à des perturbations du moteur IA. L'axe de ces tests vise à exhiber les possibilités d'exploitation malveillante d'un réseau de neurones, tout en travaillant dans un environnement contrôlé. Enfin, la normalisation de l'utilisation des réseaux de neurones et l'élaboration de bonnes pratiques constituent une approche abordée par l'ANSSI et les CESTI pour identifier les risques sécuritaires. Ces méthodologies incluent des guides détaillés sur les tests à effectuer et les mesures de sécurité à mettre en place. Leur adoption permettra de garantir une évaluation cohérente et

fiable des produits contenant des réseaux de neurones – à l'instar des méthodologies sur l'analyse de code.

Conclusion

Les CESTI ont pour mission de certifier ou qualifier des produits afin d'instaurer la confiance des utilisateurs finaux. Celle-ci est particulièrement cruciale pour les produits intégrant des fonctionnalités d'intelligence artificielle. Bien que l'IA ne transforme pas fondamentalement le processus d'évaluation, elle introduit un nouveau profil d'attaquant potentiel, rendant ainsi la certification encore plus essentielle. À l'avenir, les CESTI pourraient collaborer avec des laboratoires spécialisés dans l'alignement des IA pour fournir une évaluation globale des produits contenant des réseaux de neurones. Ces laboratoires se concentrent sur l'amélioration de la sécurité, de l'éthique et de la fiabilité des systèmes d'IA et permettent de garantir les résultats.

Cette collaboration permettrait de combiner les connaissances approfondies et l'expertise technique des CESTI avec celles des laboratoires spécialisés, offrant ainsi une évaluation complète et robuste des produits intégrant de l'intelligence artificielle, et donc un niveau de confiance bien plus élevé



Transformations de la guerre : les enjeux de la confiance dans les applications militaires de l'IA



Edouard Josse
Managing Consultant
Secteur Public
Capgemini Invent



Cerine Ferrah
Strategy & Transformation
Consultant, Public Sector
Capgemini Invent



À l'heure où nous écrivons ces lignes, l'IA transforme déjà en profondeur les opérations militaires sur le terrain ukrainien, où précision, rapidité et autonomie jouent un rôle crucial. L'IA de confiance se distingue par sa capacité à répondre aux exigences spécifiques de robustesse, d'explicabilité, de transparence et de contrôle humain. Déployée dans des contextes militaires, elle doit garantir des performances fiables tout en minimisant les risques liés aux biais ou aux défaillances.

L'IA de confiance va ainsi au-delà des enjeux relevant de l'innovation technologique, et des défis stratégiques et opérationnels sous-jacents, en intégrant des considérations éthiques et juridiques afin de s'aligner avec les règles d'engagement des armées.

L'IA de confiance concourt aussi à une parfaite maîtrise des enjeux d'embarquabilité et de robustesse dans les systèmes d'armes. En France, cette notion a été très tôt prise en compte par le ministère des Armées. Dès 2022, le document d'orientation de l'innovation de défense a souligné la nécessité de développer des solutions d'IA explicables, frugales et auditables.

Les récentes initiatives françaises, telles que la création de l'AMIAD (Agence ministérielle pour l'IA de Défense), la hausse des financements et la mise en place d'« usines IA » par les industriels, illustrent l'ambition de développer des systèmes IA à la fois performants et alignés sur ces principes.

Ces solutions, qu'elles soient intégrées dans systèmes embarqués, des systèmes de ciblage ou de

commandement et conduite (C2), s'inscrivent dans une approche où la confiance est au cœur de la stratégie, offrant une supériorité tactique et décisionnelle dans des environnements complexes et évolutifs.

La guerre en Ukraine a mis en exergue la multiplication des applications militaires de l'IA. Premier conflit de haute intensité à l'ère de l'IA, le champ de bataille ukrainien a consacré de grands bouleversements dans la conduite des opérations, au niveau opérationnel comme tactique, et dans la manière d'intégrer l'innovation en boucle courte dans les forces. Cette guerre préfiguratrice de la conflictualité moderne a ainsi montré l'importance de la masse à faible coût associée à l'autonomie dans les capacités opérationnelles et systèmes d'armes. Il s'agit donc de matériels augmentés par l'IA embarquée – intelligence et autonomie – peu dispendieux et livrés au rythme du combat ; qui sont considérés par les deux belligérants comme du consommable.

Les systèmes autonomes concourent à l'obtention de la supériorité des feux, un avantage tactique décisif. L'omniprésence du feu de précision, direct comme indirect, grâce aux systèmes d'IA rend le champ de bataille toujours plus transparent et les cibles détectées sont donc détruites en quelques secondes ; d'où l'utilisation de drones à bas coût.

Dans cette transparence du champ de bataille, la supériorité opérationnelle repose ainsi sur la supériorité décisionnelle, soit la capacité à avoir une compréhension tactique en temps réel ou quasi réel, rendue possible par la fusion de données multi capteurs, et une fluidité

accrue dans l'exécution de la manœuvre. L'IA doit servir à analyser la masse d'informations disponibles pour les restituer de manière pertinente au commandement, en limitant la surcharge cognitive. C'est pourquoi les cas d'usage d'aide à la planification et à la conduite des opérations, et à l'exploitation du renseignement, servent l'initiative et la liberté d'action du chef.

L'IA offre plusieurs avantages opérationnels clés : vitesse, en accélérant planification et conduite des opérations, coordination et boucle de décision ; masse, mobilisant une puissance de travail accrue ; qualité, face à la complexité croissante des opérations, garantissant des décisions fiables et précises ; résilience, en renforçant l'autonomie du commandement en effectifs contraints et dispersés. Ces avantages concourent à la supériorité décisionnelle et à la précision des feux, tout en réduisant la vulnérabilité du soldat débarqué.

Plusieurs acteurs européens ont émergé ces dernières années pour porter ce type de cas d'usage IA à haute valeur. Que ce soit dans l'analyse géospatiale pour le

renseignement (GEOINT) à travers des technologies de vision par ordinateur appliquées à des images satellite, dans l'aide à la décision et l'automatisation des processus de commandement et de contrôle (C2), ou encore dans l'IA embarquée pour des systèmes de défense augmentés. Ces acteurs ont apporté une proposition de valeur qui répond à des besoins opérationnels nouveaux. Cependant, l'heure est désormais au passage à l'échelle, de la transformation d'idées en capacités, afin de rendre ces solutions IA véritablement éprouvées au combat. Les systémiers-intégrateurs^[1] de la base industrielle et technologique de défense (BITD) sont naturellement indispensables pour conduire cette industrialisation, cette intégration dans les opérations d'armement, mais doivent s'appuyer sur un large écosystème pour y parvenir. Les entreprises de services numériques (ESN) en font partie, alliant expertise technique, agilité et innovation.

Vers la sécurisation et la frugalité des fonctions IA embarquées



Gaëtan Le Guelvouit
Direction Performance
et Innovation
b<>com



Stéphane Paquelet
Direction Performance
et Innovation
b<>com



Les fonctions d'intelligence artificielle se sont imposées dans de nombreux domaines, permettant d'accélérer des tâches répétitives, d'améliorer l'expérience utilisateur et même de contribuer à la création. Jusqu'à récemment, ces fonctions étaient essentiellement exécutées dans le cloud afin de pouvoir accéder aux puissances de calcul et d'énergie nécessaires. Mais on voit apparaître depuis peu des architectures dédiées à l'IA, les NPU (Neural Processing Unit), pendents neuronaux des classiques CPU et GPU.

Du fait de leur conception et de leur spécialisation, les NPU permettent l'exécution locale de fonctions IA avec une consommation énergétique réduite. Ces composants sont désormais présents sur la plupart des smartphones haut de gamme, dans les téléviseurs ou dans les ordinateurs portables.

L'exécution locale représente un atout en matière de confidentialité et d'intégrité. Toutefois, les capacités des NPU restent souvent limitées, notamment en termes de mémoire – réduisant ainsi la taille des modèles – et ne prévoient généralement pas de mécanismes intégrés de protection.

Nous décrivons dans cet article deux approches permettant de construire des réseaux compacts de traitement du signal et d'assurer leur sécurité.

Algorithmes frugaux et interpolateurs neuronaux

Le traitement du signal en temps réel repose sur un compromis délicat entre optimalité statistique et simplicité algorithmique. Par exemple, l'exploration

exhaustive d'une fonction de vraisemblance dans le cadre de problèmes d'estimation statistique présente souvent une complexité combinatoire difficilement compatible avec les matériels limités en mémoire et en puissance de calcul. Dans ces cas, il est parfois nécessaire de faire des compromis en sacrifiant une partie de la précision statistique pour répondre aux contraintes d'exécution.

Parallèlement, les réseaux neuronaux se sont imposés grâce à leur propriété d'approximation universelle. Ils permettent d'approcher des fonctions non linéaires complexes pour des tâches variées, mais ne répondent pas toujours aux exigences statistiques, comme la réduction des biais ou la résistance au bruit.

Pour concilier ces deux approches – optimalité statistique et traitement en temps réel –, nous proposons d'utiliser systématiquement des interpolateurs neuronaux peu profonds (shallow neural networks) pour calculer automatiquement les critères statistiques comme le maximum de vraisemblance.

Ces architectures permettront de :

- réduire la complexité de calcul tout en maintenant des performances satisfaisantes,
- et répondre aux contraintes des systèmes embarqués, où la capacité de calcul et la consommation énergétique sont limitées.

La conception de ces réseaux peut s'appuyer sur des données simulées, en combinant des modèles théoriques du signal et des mécanismes d'apprentissage automatique. Cela pourrait marquer

une rupture technologique dans l'éco-conception en rendant possible une IA embarquée explicable.

Sécurisation des réseaux neuronaux

La durabilité de l'éco-conception passe également par une IA de confiance, capable de résister aux cyberattaques telles que le piratage ou l'empoisonnement. Les réseaux neuronaux doivent être protégés pour éviter toute exploitation frauduleuse, comme la contrefaçon ou le détournement de données. Pour cela, nous proposons l'utilisation de mécanismes de sécurisation incluant :

- Le bruitage des poids du réseau. Les poids des réseaux sont perturbés par l'ajout d'un bruit contrôlé, rendant leur lecture ou utilisation frauduleuse inefficace sans clef dédiée.
- Le chiffrement des sorties. Une clef de déchiffrement est nécessaire pour interpréter correctement les résultats produits par le réseau, ce qui protège à la fois les données et l'intégrité des algorithmes.

Cette combinaison permet de construire une protection à la manière d'une boîte blanche cryptographique : l'algorithme IA s'exécute dans un environnement embarqué non sécurisé, ce qui rend impossible toute rétro-conception du traitement et toute récupération des données utilisées.

C'est la direction prise par la startup Skyld^[1], qui protège toute fonction d'inférence contre le vol de modèles de réseaux de neurones. Cependant, pour renforcer sa robustesse, il est nécessaire de l'adapter aux contraintes liées à la quantification des opérations, un domaine encore peu exploré dans la littérature. Par ailleurs, l'ajout d'une fonctionnalité garantissant l'intégrité des réseaux de neurones demeure un défi non résolu.

Exemples d'applications

Les récepteurs 5G utilisent des méthodes d'estimation des paramètres de synchronisation temporels et fréquentiels des signaux. Dans une de nos publications^[2], nous montrons comment un réseau de neurones compact permet de réduire de deux à trois ordres de grandeur le nombre d'opérations élémentaires par rapport à une exploration exhaustive sur la grille des paramètres.

Cette approche est théoriquement généralisable grâce aux travaux fondamentaux de Kolmogorov-Arnold^[3], revisités et adaptés aux techniques neuronales en 2024^[4], suggèrent qu'il est possible de réduire considérablement la taille des représentations neuronales en intégrant des splines.

Dans un autre domaine, nous avons développé une

fonction de conversion vidéo en temps réel^[5] permettant d'améliorer le rendu des images affichées. Initialement sous forme analytique, nous avons adapté le modèle sous forme de réseau de neurones afin de l'intégrer aux décodeurs vidéo équipés d'un NPU. Un travail autour de la frugalité a été effectué par élagage et quantification afin de rentrer dans la mémoire réduite du composant. Cette fonction a pour vocation d'être disponible dans des terminaux grand public et sera utilisée pour manipuler des contenus de valeur (films, retransmissions d'événements sportifs). Si l'on ajoute la propriété industrielle de l'algorithme lui-même, on se rend compte qu'il est indispensable d'y ajouter des fonctions de sécurité.

Le travail envisagé avec Skyld consiste alors à appliquer leur technologie de protection, à l'adapter aux contraintes du temps réel et à définir le meilleur compromis entre sécurité et frugalité. Nous disposerons ainsi d'une sécurisation prouvée, renouvelable et automatisée.

Conclusion : les outils pour l'IA embarquée et sécurisée

D'une manière générale, systématiser la mise en œuvre d'algorithmes sous une forme neuronale présente de multiples avantages. D'abord, elle permet de standardiser les mises à jour des algorithmes (qui consistent simplement à modifier les poids du réseau tout en restant agnostique par rapport à la cible). De plus, grâce à la disponibilité croissante des NPU, l'exécution est frugale, y compris dans l'embarqué. Enfin, grâce aux travaux sur une protection générique des modèles (confidentialité des poids et des données, intégrité), leur sécurisation sera grandement facilitée sans que les performances en pâtissent.

Face à une IA source de crises, la confiance comme solution ?



Justin PONCET

Founder
OPSCI.AI



Clément BÉNESSE

Head of AI research
OPSCI.AI



L'incursion de l'intelligence artificielle dans tous les domaines, à toutes les étapes des chaînes de valeurs et de manière distribuée, remet en cause un certain nombre de pratiques au nom des gains de productivité ou du passage à l'échelle massif. Mais le moteur de cette cinquième révolution industrielle n'est pas seulement un objet technique, c'est un objet culturel, fait de modèles entraînés sur de vastes corpus de données sociales. Ce « produit de la société » en catalyse donc les risques et les dysfonctionnements, tel un propagateur potentiel de biais et de désinformations (cf. éditions précédentes du livre blanc).

Chaque modèle, en particulier pour l'IA générative et dans le cas des LLM, est un prisme de valeurs d'une société, de ses valeurs, de ses idées, de ses idéologies ; et chaque utilisation du modèle contribue à leur propagation. Les grands modèles sont logiquement sous scrutin, l'on pense par exemple aux pudeurs de DeepSeek sur Tian'anmen ou Hong Kong, très largement commentées dans la presse. Les risques ne se limitent pas à ces cas les plus évidents.

Les pipelines agentiques, une dissimulation des risques

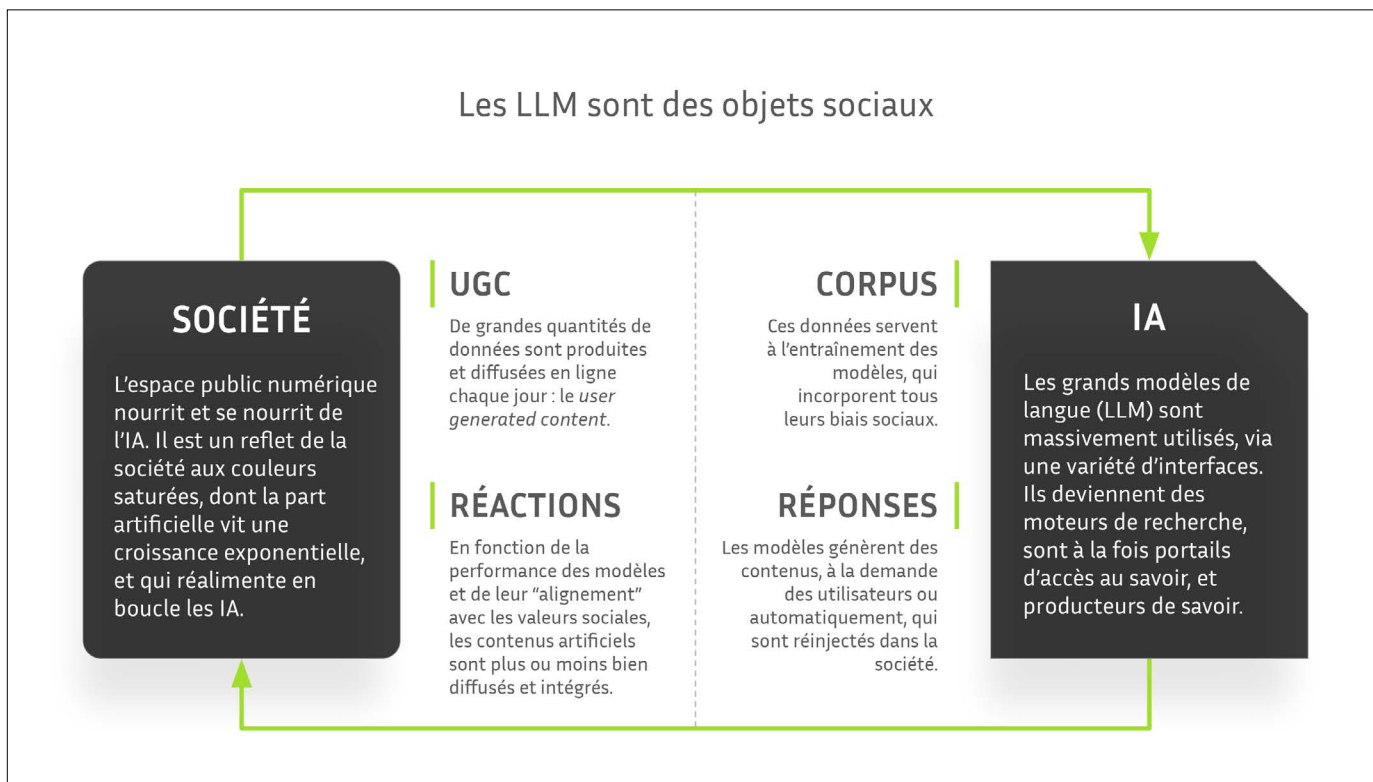
Si 2024 était l'année durant laquelle les modèles open-source ont rattrapé les modèles fermés, 2025 sera - est déjà ! - placée sous le signe des « agents ». Cette

approche, par l'appel répété de modèles qui « décident » par eux-mêmes comment utiliser, formater et présenter les données, contribue à une obfuscation des biais culturels et idéologiques. En sortant le modèle du devant de la scène et en le faisant apparaître, à tort, seulement comme une brique technique à impact minimal, ce nouveau paradigme floute l'impact à grande échelle que peut avoir un modèle aligné selon certaines valeurs.

Dans un contexte géopolitique en proie à des tensions particulièrement vives entre blocs capables de produire des modèles à hautes performances (frontier models), la question de la souveraineté se pose plus que jamais, et nous manquons de ces nouveaux outils de soft-power ô combien nécessaires. Mais force est de constater que le pas de côté opéré par l'Europe avec l'emphase sur « l'IA de confiance » est probablement le bon. En effet, le besoin de sécurité éprouvé par tous les utilisateurs face à une IA potentiellement source de crises, accidentelles ou intentionnelles, devient la condition sine qua none de son acceptabilité sociale, institutionnelle et industrielle.

L'IA pour auditer l'IA, sous contrôle humain

La nécessité du contrôle de la machine par l'homme se heurte aux contraintes de volumes de données et d'échelles justement posées par la technologie. Il apparaît dès lors impossible de ne pas l'utiliser elle-même pour ses capacités de triage, et in fine comme juge. Les



outils d'IA générative, au-delà de leur fonctionnalité première, apportent des méthodes de clustering latent^[1]. Utilisant les mêmes types d'architecture, ces LLMs as a judge^[2] sont une sorte de « remède dans le poison » permettant un audit à l'échelle pour garantir des comportements alignés avec les attentes fixées. L'approche agentique précédemment évoquée devient alors une solution d'audit, où chaque agent déployé se spécialise et apporte un prisme d'analyse additionnel et complémentaire aux précédents. L'humain n'est pas laissé en dehors de la boucle algorithmique, comme il se positionne en « contremaître », disposant d'une vue panoptique assistée par IA.

Se pose en parallèle la question de la mise à jour des outils analytiques des crises créées par des acteurs humains (e.g. Disarm, Stix)^[3] : comment les adapter aux accélérations technologiques et garantir leur efficacité à l'échelle ?

Des données, toujours des données

L'accès à des données de qualité, et conséquemment la création synthétique de jeux de données adaptés et représentatifs, est un enjeu majeur pour le développement des solutions de contrôle. Par exemple, comment s'assurer qu'un modèle est bien aligné avec nos valeurs, sans données pour les caractériser ? Comment entraîner des agents spécialistes de prompt-injection^[4], sans exemples pour les entraîner ? Comment être plus

efficace et plus rapide (i.e. des modèles spécialistes plutôt que des modèles généralistes), sans données spécialisées ? A contrario de la promesse algorithmique de ces dernières années (« donnez-moi tout Internet et je vous rendrai un modèle qui a réponse à tout »), les données de haute qualité n'ont jamais été aussi précieuses, en ce qu'elles font mouche à chaque fois. Elles sont la base de toute possibilité de génération de scénarios de synthèse pour garantir la robustesse des modèles, en allant explorer en profondeur les failles potentielles, sur des sujets précis dont les modèles sont moins familiers, mais tout autant attendus au tournant. Cette augmentation des capacités d'audit, dans une optique de red-teaming^[5] / blue-teaming^[6] assistés par l'IA, est une des conditions de mitigation des crises à venir.

Standardisation et industrialisation des tests de cyber résilience des IA



Nicolas KALMANOVITZ

Chief Operating Officer
Yogosha



Christophe MARNAT

SVP Sales Europe & Africa
Yogosha



Ces dernières années, la vitesse d'évolution et d'adoption des technologies d'IA générative dépasse largement la capacité des organisations à déployer des protocoles de sécurité adaptés. Cela engendre une augmentation significative des risques et des menaces. Dans ce contexte, comment assurer la **cyber-résilience des IA** dans des domaines sensibles ? Nous pensons que ce défi peut être relevé grâce à une collaboration ouverte et une standardisation rapide, soutenues par des plateformes numériques qui accélèrent la pratique des checklists de tests de sécurité.

L'histoire et l'importance des checklists

Depuis l'Antiquité, les sociétés utilisent des **checklists** pour standardiser des savoirs complexes, comme en architecture ou en médecine. Au XXe siècle, leur efficacité les a rendues indispensables dans des secteurs clés de notre modernité tels que l'aviation, la médecine, l'industrie, et l'exploration spatiale.

Avec l'explosion de l'Internet au début du XXIe siècle, la sécurité des applications web est devenue critique. C'est à ce moment-là que la fondation **OWASP** a révolutionné le domaine en mettant en place une approche collaborative pour identifier les vulnérabilités, prioriser les risques et produire des ressources open-source, comme le célèbre **Top 10 OWASP**. Par son action, l'OWASP a amélioré la sécurité des systèmes, standardisé les pratiques et influencé des normes internationales (ISO, PCI DSS) tout en facilitant l'intégration de la sécurité dès la conception des systèmes (Security by Design).

Depuis 2023, OWASP a élargi ses travaux à l'intelligence artificielle avec l'**AI Security and Privacy Guide** et, plus

récemment, la checklist **OWASP Top 10 LLM (2025)**, qui est devenue une référence pour atténuer les risques des IA génératives. Cependant, ces standards généraux doivent être adaptés à la variété des assets et des contraintes des systèmes d'information qui les abritent, y compris les plus sensibles, tout en créant les conditions de l'industrialisation des tests tout au long de leur cycle de vie.

Expérimentations récentes : une méthodologie éprouvée

En 2023 et 2024, deux initiatives ont émergé dans des contextes variés afin de répondre à un même enjeu : l'absence de **tests de sécurité standardisés** pour certains assets critiques. La première initiative menée par le **Groupe ADP** (Aéroport de Paris) et **Yogosha** visait à industrialiser les tests de sécurité des équipements de sécurité (scanners de personnes et d'équipements, détecteurs d'explosifs, etc.) tandis que la seconde entre le **CNES** et **Yogosha** avait pour objectif d'industrialiser les tests de sécurité sur les systèmes spatiaux (stations sols, systèmes spatiaux, systèmes de liaisons, etc).

Une démarche structurée en plusieurs étapes a été expérimentées dans les deux cas :

1. Déploiement d'une plateforme digitale pour piloter les tests offensifs, centraliser les vulnérabilités et s'adapter aux contraintes d'hébergement (SecNum Cloud, on-premise connecté ou déconnecté).
2. Collaboration interdisciplinaire : réunir des experts internationaux et des membres de l'organisation pour tester les systèmes pendant plusieurs jours à distance ou sur site.
3. Analyse collaborative : partager les scénarios

- d'attaque, identifier les vulnérabilités les plus importantes et élaborer une checklist reproductible tenant compte des standards, des spécificités des assets et de la priorisation des risques.
4. Industrialisation : Intégrer la checklist dans une plateforme de sécurité offensive sous la forme d'un guide de test digitalisé pour systématiser les tests offensifs tout au long du cycle de vie des assets.
 5. Partage et mise à jour : lorsque cela est possible, diffuser publiquement les checklists pour renforcer les compétences de la communauté des chercheurs, diffuser les bonnes pratiques dans les organisations et faciliter une mise à jour régulière au fur et à mesure de l'évolution des menaces.

Ces deux expérimentations ont atteint leurs objectifs et cette méthodologie a prouvé dans deux contextes différents qu'il est possible d'industrialiser les tests de sécurité en seulement quelques mois, même dans le cadre de systèmes sensibles. Elle peut être appliquée aux tests de sécurité et de résilience des IA, notamment LLM.

Une recette simple pour sécuriser les IA

Pour résumer, il apparaît que les leviers de l'industrialisation des tests de sécurité des IA sont les suivants :

- Plateforme de tests de sécurité offensive intégrant un système de checklists et adaptable aux contraintes d'hébergement.
- Intégration des standards existants (ex: Checklist OWASP Top 10 LLM 2025 pour les IA Générative)
- Collaboration interdisciplinaire : mobiliser des experts variés internes et externes pour combler les lacunes des standards.
- Création et systématisation des checklists : développer de nouvelles checklist pour les assets spécifiques et intégrer les tests tout au long de leur cycle de vie.
- Partage des connaissances : diffuser en interne et, si possible, publiquement les résultats pour renforcer le partage des bonnes pratiques et le développement de nouveaux standards.

Une sécurité résiliente dès la conception

C'est en généralisant cette approche collaborative et méthodique que nous pourrons regagner du terrain face aux technologies émergentes et développer les IA de confiance dans le domaine de la Défense. Elle est à même de garantir une sécurité renforcée et une résilience des écosystèmes IA grâce à une approche "Security by Design", qui protège les systèmes tout au long de leur cycle de vie.

L'IA de confiance est une illusion : construisons des produits de confiance



Florian MAILLARBAUX

Directeur Développement
Commercial
Comand AI



Violette NICAISE

Responsable de projets - Dévelop-
pement Commercial & Stratégie
Comand AI



1. L'IA de confiance, une illusion technique et opérationnelle ?

Le monde de l'intelligence artificielle a connu plusieurs évolutions majeures au cours des dernières décennies. À ses débuts, l'automatisation reposait sur une algorithmique déterministe, fournissant des preuves formelles de résultats pour des instructions données. Toute erreur intrinsèque au fonctionnement d'un algorithme était rare et identifiable. Ce cadre de prévisibilité s'est érodé avec l'arrivée du Machine Learning, où les sorties générées par l'IA se font sur des bases probabilistes et non algorithmiques. Bien que les résultats ne soient plus totalement explicables, il reste encore possible d'évaluer quantitativement la performance de ces modèles.

L'arrivée des modèles génératifs, en particulier des grands modèles de langage, a amplifié cette incertitude. Dans de nombreux cas, il est désormais impossible de mesurer objectivement la fiabilité et la performance des sorties des LLMs. Comment, par exemple, mesurer les biais d'un LLM qui rédige des résumés de textes ? Si les conclusions générées par des LLMs sont donc toujours vraisemblables, leur exactitude, le respect des consignes données, et donc leur fiabilité, ne peuvent pas être objectivement établies.

Cette difficulté à évaluer la fiabilité des modèles s'étend aux conditions dans lesquelles ils sont développés et exploités. Un modèle peut être biaisé ou manipulé à trois niveaux critiques : l'infrastructure, l'inférence et l'embedding^[1]. À ce dernier niveau, de subtiles altérations dans les premières couches du modèle permettent

d'intégrer des instructions cachées, potentiellement malveillantes et presque indétectables^[2].

L'idée d'une 'IA de confiance' repose ainsi sur une hypothèse erronée : celle qu'un modèle pourrait être fiable, prévisible et explicable. Non seulement les modèles ne sont plus fiables à 100 %, mais la question même de la mesure de la fiabilité devient délicate. Étant donc entendu qu'il n'existe pas de méthode permettant de garantir la confiance dans les réponses générées par l'IA, il apparaît illusoire de vouloir se doter d'une IA de confiance.

Par ailleurs, la possibilité même de labelliser une IA comme fiable introduirait un risque majeur dans un contexte opérationnel où la rapidité de décision est essentielle : celui d'une délégation aveugle à une machine perçue comme infaillible. Plus un système est estampillé de confiance, moins ses résultats seront interrogés, augmentant le risque d'une acceptation automatique de ses conclusions sans analyse critique. L'IA n'est autre qu'un conseiller éclairé, un outil d'aide à la décision dont l'humain reste l'ultime responsable—tant sur le plan juridique que moral.

Pour autant, l'impossibilité d'établir la fiabilité de ces modèles n'enlève rien à la puissance de ces outils lorsqu'il s'agit d'analyser de grandes quantités de données, d'automatiser des tâches répétitives et de dégager des tendances exploitables. Plutôt que de chercher une IA infaillible, l'enjeu est d'intégrer ces modèles dans des systèmes qui en maximisent la valeur sans en déléguer aveuglément le contrôle.

2. Une alternative pragmatique : le produit de confiance

L'IA ne doit donc pas être perçue comme une entité autonome à laquelle on délègue des décisions, mais bien comme un outil d'analyse et un amplificateur de capacités humaines. À l'humain la confiance et à l'IA la performance.

Le produit doit permettre cette juste articulation, en intégrant des points de contrôle et d'analyse clés aux moments opportuns de l'expérience utilisateur. Il doit également être simple d'usage, y compris en environnement dégradé. Pour concevoir ce produit de confiance, il convient de se prémunir de différents écueils.

Le premier serait de limiter l'intervention humaine au rôle de 'pousse-bouton'. L'exemple de l'utilisation du logiciel de ciblage Lavender par Tsahal à Gaza est à ce titre édifiant. Des erreurs factuelles, comme l'absence des cibles aux emplacements indiqués ont été ignorées lors des décisions de frappe prises par les opérateurs, alors même que l'outil d'intégration de données leur permettait de tester et de confronter les analyses^[3]

Plusieurs facteurs expliquent cette dérive : une surcharge d'informations à traiter en un temps réduit, une confiance excessive dans la machine et un tempo opérationnel soutenu qui dissuade l'opérateur de confronter l'analyse fournie à son expertise. Mais ces erreurs relèvent également d'un défaut de conception du logiciel, qui devrait intégrer des mécanismes intégrant des étapes de validation critique, même dans un tempo opérationnel soutenu.

Le deuxième serait de ne pas permettre aux opérateurs d'en reprendre immédiatement son contrôle, comme ce fut le cas des pilotes des Boeing 737 MAX, dont deux crashes successifs sont advenus en 2018 et en 2019. Le système de stabilisation intégré, le MCAS, pouvait s'activer en réponse à un seul capteur, sans en avertir les pilotes ni leur permettre de reprendre la main sur la trajectoire^[4]. Le problème ne résidait pas dans la technologie d'assistance au pilotage en soi, mais dans une conception logicielle qui privait l'humain d'un droit de regard et d'une capacité d'action immédiate. Enfin, le troisième écueil serait un produit qui impose une validation humaine à chaque micro-bloc d'IA. Une supervision excessive, imposant un contrôle systématique à chaque étape du raisonnement, même les plus minimes, deviendrait contre-productive et rendrait l'avantage stratégique constitué par l'IA caduque. Le temps de traitement exploserait, l'IA perdrait alors tout intérêt opérationnel.

La solution réside donc dans un produit judicieusement conçu pour l'utilisateur, où la performance des briques logicielles ne court-circuite pas les capacités humaines mais les amplifie. Un produit de confiance doit intégrer des points de contrôle aux moments clés de l'expérience utilisateur, sans alourdir inutilement le processus. Surtout, il doit être d'une simplicité d'usage telle que ces validations restent intuitives et facilement réalisables, même dans des environnements dégradés. C'est bien sur l'interaction entre l'opérateur et la machine que porte la confiance, non sur l'IA en elle-même.

C'est pourquoi nous plaidons pour un produit de confiance : un système conçu pour intégrer l'IA comme un amplificateur de capacités humaines, non comme un substitut à la réflexion et à la responsabilité. Plutôt que de poursuivre la quête d'une IA infaillible, l'Europe et la France doivent investir dans de bons produits de confiance, où l'IA est un levier stratégique maîtrisé.

Défis techniques de

L'IA





L'IA de confiance pour la défense : de la conception à l'évaluation



Sébastien HENWOOD

Expert IA, PhD
AMIAD



Dominique CHAUVEAU

Manager partenariats et référent
cyber
AMIAD



L'IA de confiance, quels enjeux

L'Agence Ministérielle pour l'Intelligence Artificielle de Défense (AMIAD) établie courant 2024 introduit une capacité en IA au ministère des Armées. Dans la logique historique d'équiper nos forces armées de matériels sûrs pour les utilisateurs et sécuritaires pour les infrastructures, l'IA de confiance apparaît comme un objectif évident.

Si la révolution du numérique du début du siècle a pu bousculer les processus de conception usuels de matériels défense, les IA de type apprentissage automatique apportent un lot d'enjeux venant à nouveau perturber les cycles de conception habituels. En particulier, les performances d'un modèle d'IA sont directement dépendantes de la quantité et de la qualité des données disponibles lors d'une phase d'entraînement.

Toutefois, ces données sont généralement acquises de manière incrémentale voire même en permanence : il convient alors d'itérer rapidement sur de nouvelles versions d'un modèle d'IA, meilleur - ou plus adapté à un nouveau contexte d'emploi - que son prédécesseur. Dans ce scénario, l'évaluation du caractère de confiance n'est pas circonscrite à une phase de développement en amont de l'usage par les forces d'un système embarquant une composante d'IA, mais s'étend dans tout le cycle de vie de l'IA.

Chercher à se soustraire à cette facette itérative de l'IA serait risquer un retard de phase sur d'importantes évolutions techniques ce qui induit des risques pour l'accomplissement des missions régaliennes. En conséquence, tous les processus

de spécification, conception et évaluation sont à adapter à la nature évolutive de l'IA.

Par ailleurs, on peut constater que les outils et méthodes permettant de mesurer la confiance en un système d'IA concentrent les efforts d'une frange conséquente de l'industrie et du domaine académique. En témoignent les nombreuses initiatives visant à établir une hygiène du développement de l'IA et les risques à anticiper : AI Act, Confiance.AI, DEEL, etc. Les organismes normatifs comme l'ISO proposent des normes^{[1][2]} pour le domaine de l'IA et certains domaines industriels se dotent aussi de normes appropriées à leurs spécificités.

Concevoir de l'IA de confiance

On constate ainsi que le domaine de l'IA de confiance cristallise l'intérêt de nombreux acteurs en France et à l'international. En particulier, on remarque que la prise en compte de cette problématique et l'anticipation des questions relatives à l'évaluation des briques d'IA se doit de faire partie intégrante, au fur et à mesure des différentes étapes - potentiellement itératives - du processus de conception d'un système, dès lors qu'il intègre des fonctions adressées par de l'IA. Sur la base des éléments observés dans les travaux relatifs à l'IA de confiance, nous identifions au minimum quatre caractéristiques essentielles d'un processus à même de produire de l'IA de confiance :

- Multi-compétences : le processus doit aborder une approche pluridisciplinaire de conception de l'IA de défense. Les champs d'expertise doivent y échanger sur les aspects techniques permettant d'aboutir à une IA performante et de

- confiance à chaque phase de son cycle de vie. Ces compétences incluent sans s'y limiter l'expertise en IA, mais également l'étude des facteurs humain, des problématiques d'embarquabilité, des aspects sécurité de l'IA ou encore de MCO [3],
- Agile : le processus de conception doit intégrer facilement toute évolution technologique permettant de bonifier les performances de l'IA de confiance dans une logique d'amélioration continue,
 - Itératif : le processus de conception est conçu pour permettre des incréments selon la mise à disposition de nouvelles données, le changement de cadre d'emploi, ou tout autre évènement déclencheur dans le cycle de vie de l'IA,
 - Traçable : dans une démarche scientifique, le processus de conception est documenté et les différents choix de conception sont traçables au travers d'une ou plusieurs expériences permettant à tout acteur de juger et comparer de la pertinence des choix effectués.

Évaluer de l'IA pour la confiance

Ce processus de conception est à mettre en relief avec le processus d'évaluation. Ce dernier partage certaines caractéristiques du processus de conception, et peut s'y intégrer. D'une part, en étant au fait de l'état de l'art des outils et méthodes permettant de concevoir et de mesurer la confiance en un système ayant une composante d'IA. D'autre part, en participant aux étapes de spécifications tant du processus de conception d'un composant que des caractéristiques de ce dernier pour en augmenter les attributs de confiance : qualité attendue des données, métriques, ou encore sûreté et sécurité sont des éléments à considérer tout au long de la mise en place du système afin d'aider à l'évaluation de la confiance à accorder aux briques d'IA. Cet échange transparent et constant entre les acteurs tout au long du développement doit nourrir le processus de conception : le concepteur perçoit clairement les attentes du client et ajuste ainsi son flot de conception ; d'autre part l'évaluateur peut identifier les écueils émergents durant le projet et apporter une expertise complémentaire dans l'obtention d'une IA de confiance. Ce procédé s'inscrit dans le cycle de vie de l'IA : dans son développement initial, mais également lors de la phase d'utilisation lorsque des incréments, réentraînements ou autres changements de domaines d'emploi sont à considérer.

Le projet en ressort gagnant : mieux spécifié, mieux conçu, mieux évalué. Les inévitables évolutions à apporter au système pour satisfaire les évaluations sont également maîtrisées au plus tôt. Toutefois, ce fonctionnement collaboratif entre le processus de conception et l'évaluation d'un composant d'IA bouleverse les habitudes des différentes parties prenantes, notamment pour les systèmes les plus critiques. En comparaison avec une traditionnelle

campagne d'essai, l'évaluateur ne saura se contenter de résultats sur un jeu de données d'évaluation ou de qualification (qui demeurent néanmoins essentiels). Dans l'optique de mesurer la confiance, il souhaitera observer et juger sur pièces les jeux de données d'entraînement, protocoles d'entraînement et d'ajustement, et autres éléments pertinents à la conception d'un composant d'IA : c'est le concept d'assurance d'apprentissage.

L'IA de confiance, une révolution pour le ministère des armées

Le développement d'une IA de confiance n'est donc pas seulement limité aux évolutions techniques, mais également à une évolution du paradigme de conception et d'évaluation du système tout au long de sa vie. Cette évolution induit une collaboration plus étroite entre client, concepteur et évaluateur afin de concourir à l'obtention d'une IA de confiance. En d'autres termes, les processus organisationnels de ces parties prenantes doivent s'inscrire dans une logique de transparence pour la conception du composant d'IA. Cette transparence participe à la création d'un processus de conception fiable, aboutissant à des IA de confiance. Cette évolution organisationnelle sous-tend les efforts du ministère des Armées dans la démarche d'adoption de cette technologie de rupture, démarche nécessaire à la mise en place d'une IA de défense qui participe activement à l'accomplissement des missions régaliennes.

Vers des IA vérifiables



Arnaud VALENCE

Enseignant-chercheur en cybersécurité au laboratoire Confiance Numérique et Sécurité ESIEA



Les systèmes d'intelligence artificielle (SIA) sont de plus en plus intégrés à notre vie quotidienne, et sont même désormais en mesure de nous fournir des conseils grâce aux IA génératives. Cependant, à mesure que ces systèmes gagnent en importance, les appels à la transparence se multiplient. En effet, s'il est légitime de protéger les fournisseurs de SIA contre la violation de leurs secrets de fabrication, il est tout autant légitime de demander à ces mêmes fournisseurs de garantir l'intégrité de leurs modèles. Cette tension entre les exigences d'intégrité et de confidentialité est inhérente aux SIA « fermés ». Pourtant, elle peut être surmontée par le dernier cri des technologies cryptographiques : les preuves à divulgation nulle de connaissance.

Une preuve à divulgation nulle de connaissance (ZKP, Zero Knowledge Proof) est un protocole cryptographique permettant à une entité dite « prouveur » de démontrer à une autre entité dite « vérificateur » qu'une information est vraie, sans révéler aucune information à son sujet (autre que sa véracité). Apparues dans les années 80, les ZKP sont aujourd'hui incontournables sitôt qu'existe une double exigence de transparence et de confidentialité des données. C'est la raison pour laquelle on les trouve implémentées dans des domaines aussi différents que le web, l'internet des objets, le vote électronique, la Défense, la santé ou encore les cryptomonnaies. Aujourd'hui, les ZKP font irruption dans les SIA au regard des enjeux éthiques, juridiques et financiers, et il est possible qu'elles figurent un jour parmi les exigences imposées aux IA de confiance (a minima aux IA stratégiques).

Si les ZKP sont aussi efficaces, c'est parce qu'elles

fournissent deux propriétés idéales : il est impossible pour le vérificateur de refuser une preuve valide et, réciproquement, il est impossible pour le prouveur de convaincre sans preuve le vérificateur que l'information est vraie. Techniquement, on parle respectivement de complétude et de robustesse. Mais ce qui fait la force des ZKP, c'est que ces propriétés essentielles sont obtenues sans divulgation du « témoin » de la preuve. Là réside, dans cette troisième propriété, l'aspect révolutionnaire des ZKP.

Des preuves d'inférence à base de ZKP

L'intérêt du monde de l'IA pour les ZKP est récent et c'est encore essentiellement une affaire de chercheurs. Les ZKP posent en réalité deux défis aux fournisseurs de SIA privés :

1. Comment peuvent-ils certifier le calcul de leurs modèles ?
2. Comment peuvent-ils certifier les calculs effectués pendant l'entraînement de leurs modèles ?

La première question conduit à la notion de « preuve d'inférence à base de ZKP ». Les preuves d'inférence sont apparues récemment, notamment depuis les travaux sur les safetynets (2017) ^[1] et VeriML (2021) ^[2]. Au départ, l'objectif était de vérifier l'exactitude des calculs effectués (par un serveur ML-as-a-service) sans exigence particulière sur la confidentialité du modèle du serveur. C'est cette exigence manquante que visent les preuves d'inférence à base de ZKP, pour garantir le secret du calcul tout en conservant la garantie que le calcul a bien été effectué par le modèle putatif. Les ZKP permettent aux fournisseurs de SIA de produire une preuve qu'un calcul demandé s'est produit honnêtement, c'est-à-dire qu'une matrice de

ponds spécifique a été utilisée sur les entrées du modèle, sans demander la publication desdits poids (qui n'est pas exigible aux systèmes fermés dès lors que prévalent des questions commerciales, éthiques ou régaliennes).

Au début des années 2020, cette solution a d'abord été envisagée sur des petits modèles tels que les modèles convolutifs adaptés aux collections MNIST (70 000 images en noir et blanc de 28x28 pixels) ou CIFAR-10 (60 000 images en couleurs de 32x32 pixels). Les derniers travaux permettent aujourd'hui le passage à l'échelle sur des collections de données réalistes telles qu'ImageNet (1.5 million d'images) ou le langage naturel. Les cas d'utilisation des SIA vérifiables sont nombreux. Citons par exemple les « audits sans confiance », l'authentification biométrique sans confiance ou encore les notes de crédit sans confiance.

En complément des ZKP, il peut être naturel d'exiger la confidentialité des données transmises par le client au serveur, afin de protéger la vie privée de l'utilisateur. Cette exigence peut être satisfaite par une autre technique cryptographique bien connue : le chiffrement homomorphe^[3].

Vers des preuves d'apprentissage

La seconde question pousse l'exigence de transparence vers ce que l'on appelle des « preuves d'apprentissage » (PoL, Proof of Learning). On vise maintenant à garantir le de l'apprentissage, car un serveur cloud pourrait être tenté de réduire le coût de calcul dans la phase d'entraînement. L'idée des PoL est de prouver que les serveurs cloud ont effectué les tâches de calcul assignées pendant la phase d'apprentissage. Ce qui est actuellement proposé est d'observer les données du calcul (par exemple les informations accumulées par l'algorithme d'apprentissage lors de la descente de gradient stochastique^[4]). Mais il n'existe pas encore à ce jour de procédure garantissant la confidentialité de l'apprentissage, i.e. sans divulgation de connaissance. À l'instar des preuves d'inférence, il est également naturel de demander la confidentialité des données d'entraînement. Plusieurs techniques cryptographiques peuvent ici jouer leur rôle, dont les circuits brouillés (Garbled Circuits), la confidentialité différentielle ou le chiffrement homomorphe. Cette dernière technique semble à nouveau la plus efficace car elle garantit que le serveur cloud n'apprend rien sur les données (puisqu'elles lui sont transmises une fois chiffrées).

Conclusion

La vérifiabilité des IA est la dernière-née des propriétés que l'on attend d'une IA de confiance. Elle consiste en l'hybridation entre l'IA et les techniques cryptographiques, afin de surmonter l'apparente contradiction entre

certification et préservation du secret du calcul (soit entre intégrité et confidentialité du calcul). Ce dépassement est rendu possible par l'utilisation des ZKP, qui sont déjà expérimentées dans les phases d'inférence et ne demandent qu'à l'être dans les phases d'entraînement. Les ZKP ne sont en revanche d'aucune utilité pour garantir le secret des entrées des modèles. Cette autre propriété est réglée par le chiffre homomorphe, qu'on peut voir à ce titre comme le compagnon de jeu idéal des ZKP, pour protéger la vie privée des utilisateurs.

L'influence de l'IA sur la cybersécurité : défis et opportunités du point de vue de la recherche



Marc-Oliver PAHL

Titulaire Chaire Cybersécurité des Infrastructures Critiques
CyberCNI
Directeur de recherches
IMT Atlantique



L'intelligence artificielle (IA) a évolué rapidement ces dernières années, impactant de nombreux secteurs, dont la cybersécurité. Les avancées récentes en apprentissage automatique, IA générative et modèles de langage permettent aux attaquants d'automatiser et d'améliorer les campagnes de phishing, d'exploitation de vulnérabilités et de désinformation. En parallèle, les défenseurs de la cybersécurité utilisent l'IA pour détecter les anomalies, prédire les menaces et réagir de manière dynamique, offrant ainsi une adaptabilité et une évolutivité accrues aux systèmes de sécurité.

L'IA permet d'analyser des données complexes et d'adapter les stratégies face à des scénarios inconnus, similaire de l'apprentissage humain. L'IA offre la capacité d'analyser des données complexes et d'ajuster les stratégies face à des situations inédites, à l'image de l'apprentissage humain. En parallèle, elle exploite son potentiel pour traiter de vastes volumes de données en un temps quasi instantané. Cependant, elle présente aussi des limites, comme la vulnérabilité aux attaques adverses^[1] et des biais dans les données d'apprentissage, ce qui nécessite une mise en œuvre responsable. Cet article explore les applications principales de l'IA en cybersécurité, en d'un point de vue de la recherche autour de la cybersécurité des infrastructures critiques^[2] détaillant ses capacités défensives et son potentiel d'utilisation dans les cyberattaques.

Les futures cyberattaques alimentées par l'IA

Les cyberattaques pilotées par l'IA ne sont plus une menace lointaine, mais une réalité émergente qui redessine le paysage de la cybersécurité. Les avancées actuelles démontrent que l'IA permet désormais de concevoir des systèmes d'attaque entièrement

automatisés et adaptatifs, capables d'apprendre et d'améliorer en temps réel. Les logiciels malveillants auto-apprenants sont déjà capables de modifier de manière autonome leurs tactiques pour contourner les mécanismes de défense avancés, tels que les systèmes de détection d'intrusion (IDS) et la défense contre les cibles mobiles (MTD), rendant ainsi les défenses traditionnelles inefficaces. En outre, les attaques pilotées par l'IA deviennent de plus en plus sophistiquées, avec la capacité d'exploiter dynamiquement les vulnérabilités sans intervention humaine, en utilisant des techniques telles que l'ingénierie sociale automatisée et le spear-phishing^[3] à grande échelle. De plus, l'IA démocratise les cyberattaques, permettant même à des personnes ayant une expertise technique limitée de lancer des campagnes très efficaces. Cette tendance, qui comprend déjà l'utilisation de l'IA pour le déchiffrement de mots de passe, le déploiement de logiciels malveillants et même la génération de deepfake pour les attaques de phishing, laisse présager un avenir où les cyberattaques seront plus intelligentes, plus adaptatives et plus difficiles à prédire.

Il est donc essentiel d'utiliser l'IA du côté du défenseur, ce qui est un objet important dans la recherche en cybersécurité.

L'IA pour la modélisation semi-automatique de systèmes dans les jumeaux numériques

Une application de l'IA dans le domaine de la cybersécurité est l'aide à la génération de modèles virtuels de systèmes. Ces « jumeaux numériques » permettent aux chercheurs et opérationnels d'étudier et d'analyser leurs systèmes dans un environnement contrôlé - sans affecter directement l'infrastructure productive. Dans le cas d'une chaîne de production, un jumeau numérique peut être utilisé pour

tester les effets des attaques sur le système virtuel afin d'obtenir des informations sur le système réel sans avoir besoin de le compromettre ou d'arrêter son fonctionnement. L'apprentissage automatique permet également la détection des anomalies, une identification rapide des vecteurs d'attaque potentiels avant que les adversaires ne les exploitent, ce qui améliore en fin de compte les mécanismes de défense.

L'IA pour l'analyse des données floues et la sécurité basée sur les intentions

Les approches traditionnelles de la sécurité reposent sur des ensembles de règles prédéfinies, qui sont souvent rigides et ne s'adaptent pas à l'évolution des menaces. L'analyse de données floues pilotée par l'IA offre une nouvelle alternative, permettant aux systèmes de fonctionner sur la base d'une intention de haut niveau plutôt que sur l'application de règles spécifiques. Ce changement de paradigme permet aux utilisateurs de définir des objectifs de sécurité, tandis que l'IA adapte dynamiquement les mesures de sécurité en fonction des observations en temps réel. L'IA aide les analystes de la sécurité en automatisant les tâches de surveillance de routine, en filtrant de grandes quantités de renseignements sur les menaces et en fournissant des informations en temps réel qui seraient insurmontables pour des équipes humaines seules. Elle améliore la prise de décision en proposant des analyses prédictives et en suggérant des stratégies d'atténuation optimales, ce qui permet aux analystes de se concentrer sur les réponses stratégiques plutôt que sur la détection manuelle des menaces.

Désinformation générée par l'IA et prébunking

L'une des principales menaces émergentes en matière de cybersécurité est l'utilisation de l'IA pour produire et diffuser de la désinformation. La désinformation désigne des informations délibérément fausses ou trompeuses créées pour tromper les individus ou manipuler l'opinion publique. Des modèles d'IA générative sophistiqués peuvent produire des récits très réalistes mais faux, exacerbant la difficulté de distinguer la vérité de la fiction et sapant la confiance dans les écosystèmes d'information. Alors que les efforts de recherche se concentrent sur la détection et la prévention, les mesures proactives telles que le "prebunking" (sensibilisation des utilisateurs aux tactiques trompeuses avant qu'elles ne soient exposées) sont très prometteuses. Notre application web, JudgeGPT^[4], fournit une plateforme pour tester différents modèles d'IA dans la génération de désinformation et analyser leur impact en direct sur soi-même. Elle donne un bon exemple pour le potentiel de l'IA générative en 2025.

Confiance

L'IA est une technologie qui est clairement plus que la somme de ses parties. Même si chaque étape du fonctionnement peut être bien expliquée, la qualité des résultats et les possibilités inédites de ses résultats sont encore surprenantes, y compris pour les chercheurs. L'IA restera une technologie dominante dans le domaine du génie logiciel. Par conséquent, c'est un défi important pour la recherche que de susciter la confiance dans cette technologie. Expliquer comment l'IA est parvenue à ses conclusions et rendre ses opérations plus transparentes et compréhensibles pour les humains sont donc des éléments clés pour accroître la confiance dans la technologie de l'IA.

Conclusion

Les progrès rapides de l'IA présentent à la fois des défis et des opportunités en matière de cybersécurité. Alors que les attaquants exploitent de plus en plus l'IA pour améliorer leurs tactiques, les défenseurs peuvent tirer parti de la même technologie pour renforcer leurs réponses aux cybermenaces croissantes. Toutefois, il est essentiel d'utiliser l'IA de manière éclairée et responsable. Sans une compréhension approfondie des implications de l'IA, son déploiement peut introduire de nouvelles vulnérabilités difficiles à détecter. À l'avenir, une combinaison de progrès technologiques et d'expertise humaine sera essentielle pour sécuriser les écosystèmes numériques contre le paysage évolutif des menaces alimentées par l'IA.

Pour nos recherches à la chaire cybersécurité des infrastructures critiques, l'est clairement le développement central de la dernière décennie qui joue un rôle clé dans toutes nos recherches en tant qu'outil important et menace.

Recommandations de sécurité pour la mise-en œuvre de plateformes d'IA générative dédiées à l'ingénierie des systèmes critiques.



Adrien BECUE

Expert en IA et Cybersécurité
Thales



L'ingénierie des systèmes critiques englobe la conception, le développement, la vérification et la maintenance de systèmes dont la défaillance pourrait entraîner des conséquences graves, telles que des pertes humaines, des dommages environnementaux ou des pertes économiques significatives. L'introduction de l'IA générative offre des opportunités d'automatisation et d'amélioration de ces processus d'ingénierie. Cependant, elle introduit également des risques spécifiques qui peuvent compromettre la sécurité intrinsèque des systèmes. Par exemple, l'interaction entre un utilisateur et une IA générative peut exposer des informations techniques sensibles relatives à ces systèmes. Par ailleurs le risque d'hallucination des modèles génératifs peut entraîner des erreurs de conception et des défaillances graves. L'utilisation de l'IA générative dans l'ingénierie des systèmes critiques requiert donc une approche rigoureuse en matière de sécurité, de souveraineté et d'explicabilité. Afin d'assurer l'intégrité des systèmes et de protéger les données sensibles, nous recommandons la mise en place des mesures suivantes :

1. Protection de l'Infrastructure Matérielle

Il est essentiel de déployer l'infrastructure matérielle sur site, en évitant les services cloud non maîtrisés, afin de garantir la souveraineté des données. L'architecture matérielle doit être renforcée en utilisant des processeurs et GPU dédiés, accompagnés d'une segmentation réseau stricte pour limiter les risques de compromission. L'interconnexion avec internet ou avec une infrastructure de niveau de sécurité différent doit être sécurisée par l'usage de passerelles et de pare-feu. Des contrôles d'accès rigoureux, incluant une authentification forte et une gestion fine des permissions utilisateur,

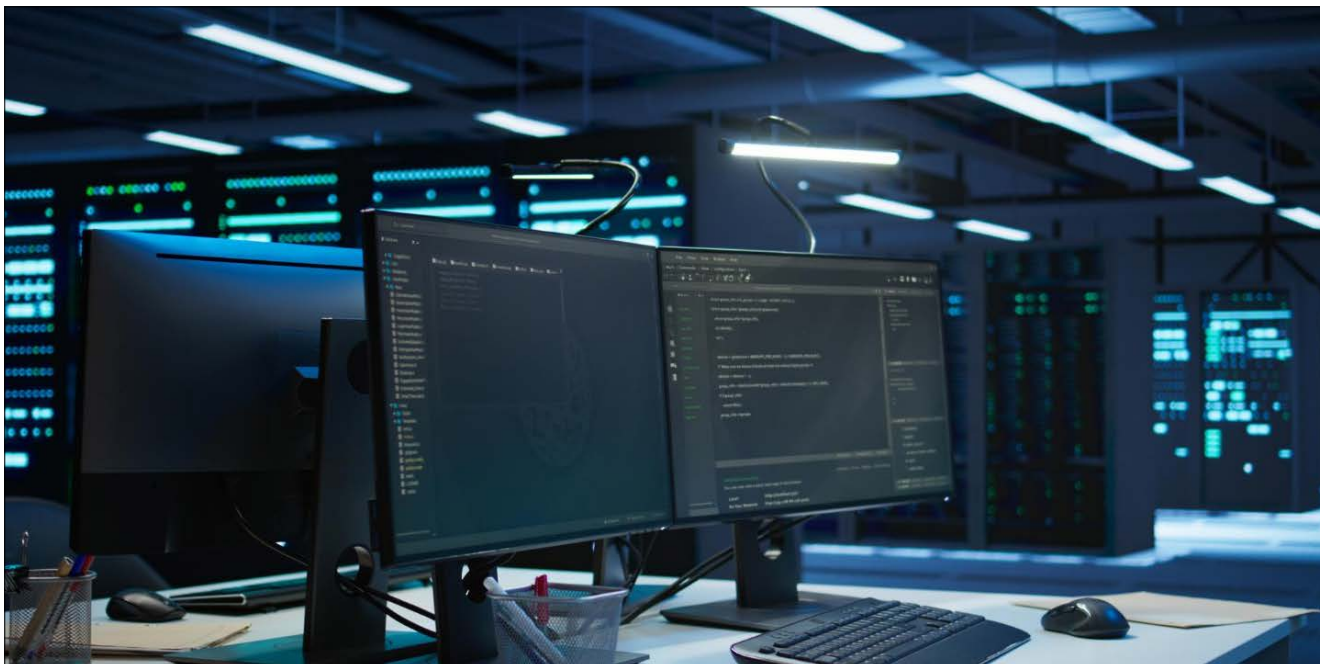
doivent être mis en place. Une surveillance continue est nécessaire, avec l'enregistrement des actions critiques et la collecte d'événements qui seront remontés en centre de supervision de sécurité pour être analysés. Enfin, la plateforme doit subir des tests de pénétration réguliers et une homologation au niveau de sécurité requis par les projets adressés.

2. Sécurisation des interactions homme-machine

Pour prévenir les risques liés à l'exploitation malveillante des modèles d'IA générative, il est recommandé de mettre en place une interface de prompts sécurisée. Cette interface doit filtrer et valider les entrées afin d'éviter les biais et hallucinations des modèles, contrôler la fréquence des requêtes pour limiter les abus et détecter les usages suspects, et enregistrer les interactions pour assurer la traçabilité et prévenir les injections malicieuses. De plus, le développement d'une bibliothèque de prompts pré-validés est conseillé. Ces prompts certifiés encadreront l'utilisation des modèles et seront mis à jour continuellement pour répondre aux nouveaux besoins et sécuriser les interactions. Ils offriront le double avantage d'améliorer la productivité des équipes d'ingénierie et réduire le risque d'erreurs liées à la qualité des prompts.

3. Contrôle des sources et vérification des sorties

Pour renforcer la cohérence et la fiabilité des résultats produits par l'IA, il est recommandé d'intégrer un mécanisme de génération augmentée par récupération (RAG). Cela implique l'utilisation d'une base de connaissances contrôlée pour améliorer la précision des réponses, le filtrage des mises à jour pour éviter les attaques de type empoisonnement, une indexation et recherche sécurisées pour garantir un accès rapide aux informations pertinentes sans compromission.



Parallèlement, la mise en place d'un mécanisme de validation neuro-symbolique est recommandée. Ce système effectuera une vérification automatique des résultats par des techniques d'IA symbolique tels que des systèmes à base de règles ou des graphes de connaissance. Il attribuera un score de fiabilité basé sur des modèles alternatifs, et vérifiera la conformité aux normes de sécurité pour éviter toute dérive du modèle principal.

4. Maîtrise des modèles et amélioration continue

Un ou plusieurs grands modèles de langage (LLM) pourront être utilisés selon les cas d'usages, sachant que des modèles généralistes seront plus souples à l'usage mais moins précis que des modèles spécialistes, notamment pour la production de code ou d'architectures techniques. On privilégiera des modèles souverains ou open sources contrôlés pour limiter les dépendances à des solutions propriétaires non maîtrisées. Leur précision à l'usage pourra être améliorée par des techniques d'affinement ou de renforcement. L'intégration d'un mécanisme de récompense par feedback humain permettra d'alimenter une boucle d'amélioration continue. Toutefois, l'empoisonnement des données d'affinage ou des récompenses sont des menaces internes à surveiller. Des mécanismes de signature, de détection et de correction des biais du modèle peuvent y remédier.

5. Supervision et gestion des incidents

Une détection proactive et une réponse rapide aux menaces sont essentielles. Les personnels du Centre d'Opérations de Sécurité (SOC) doivent être formés à reconnaître et traiter des attaques spécifiques à l'IA, telles que l'empoisonnement, l'évasion, l'inversion de modèle ou l'injection de prompt. Une analyse en

temps réel des journaux système et du comportement des utilisateurs est recommandée afin d'identifier toute anomalie, gérer centralement les incidents de cyber-sécurité et mettre en place des contre-mesures adaptées. Une veille systématique sur les menaces et les vulnérabilités spécifiques à l'IA devra être assurée. De plus, il conviendra de surveiller les évolutions réglementaires et d'assurer une mise en conformité continue de la plateforme avec les normes de sécurité.

Conclusion

L'intégration de l'IA générative dans l'ingénierie des systèmes critiques est une avancée majeure qui soulève des défis de sécurité, de souveraineté et d'explicabilité. Le présent article propose une réponse à ces enjeux à travers des préconisations : 1) Protection de l'Infrastructure Matérielle, 2) Sécurisation des interactions homme-machine, 3) Contrôle des sources et vérification des sorties, 4) Maîtrise des modèles et amélioration continue, 5) Supervision et gestion des incidents. Une application des bonnes pratiques en vigueur en matière de sécurité informatique traite déjà une grande partie des risques identifiés. Néanmoins, une compréhension des menaces et vulnérabilités spécifiques à l'IA, et notamment à l'IA générative, est nécessaire. Pour cela, nous orientons le lecteur vers quelques solides références : ENISA^[1], ANSSI^[2], MITRE ATLAS^[3], NIST AI 600-1^[4], OWASP^[5], CWE AI WG^[6].

IA raisonnante pour les systèmes décisionnels critiques



Stéphane DELAYE

Directeur technique délégué à
l'Intelligence Artificielle
Eviden



Zyed ZALILA

PDG-Fondateur & Directeur R&D
INTELLITECH
Professeur de mathématiques du flou et
d'Intelligence Artificielle
Université de Technologie de Compiègne



Cet article expose certaines des propriétés fondamentales que doit valider une Intelligence Artificielle (IA), lorsqu'elle est employée pour développer et déployer des Systèmes Décisionnels (SD) critiques.

Intelligibilité et Explicabilité des Systèmes Décisionnels

Une IA de confiance est définie par plusieurs critères essentiels qui garantissent sa fiabilité et sa conformité aux valeurs éthiques et aux exigences des Forces Armées, dans un cadre juridique international et européen qui se précise tous les jours. Après le premier sommet REAIM en 2023 (Pays-Bas), le sommet de 2024 (Corée du Sud) a permis de confirmer une feuille de route. Soutenue par 61 États, elle suggère des principes et un cadre pour la gouvernance future, en soulignant qu'être responsable implique de se conformer au droit international, de tenir les humains responsables et comptables de leurs actes, de garantir la fiabilité de l'IA, de maintenir une implication humaine appropriée et d'améliorer l'explicabilité de l'IA.

L'une de ses principales caractéristiques est d'être « centrée sur l'humain ». L'humain ne doit pas seulement être « dans la boucle » comme un point de contrôle obligatoire dans un processus, mais l'IA doit être façonnée pour être une extension de son action. L'autonomie demandée aux SD étant grandissante, il faut donc être parfaitement conscient du domaine de responsabilité délégué et être en mesure de l'assumer. Cette nécessité de contrôler la délégation plus formellement met en avant deux autres propriétés essentielles : l'intelligibilité (ou transparence) du SD et l'explicabilité de ses prédictions. Au-delà de la validation statistique et tests d'échantillons d'un SD opaque (« boîte noire »), un SD

transparent subit une validation préalable formelle et réfutable sur l'ensemble de son cadre d'emploi. Il s'agit de vérifier l'intégralité des connaissances embarquées dans le SD, formant sa logique interne décisionnelle, et si elles sont adaptées et suffisantes au cadre d'emploi.

Si le SD est intelligible, il sera nécessairement explicable : chacune de ses décisions sera expliquée grâce aux connaissances du SD, préalablement auditées et certifiées avant son déploiement. Ainsi, et grâce aux propriétés d'intelligibilité et d'explicabilité, la phase d'audit du SD permet la détection d'erreurs dans l'ensemble de données d'entraînement et de biais éventuels^[1] : un préalable pour un usage éthique et responsable des SD. La relance automatique de l'induction (du modèle de décision) sur cet ensemble de données mis à jour permet de découvrir des SD plus performants et exempts de biais.

De l'intérêt d'expliquer les connaissances d'un SD pour en comprendre son adéquation au cadre d'emploi

Imaginons qu'au sein d'un essaim de drones, l'un d'eux soit chargé de missions de reconnaissance et de classification, tout en recevant des informations transmises, en temps réel, par les autres drones de l'essaim. Afin d'assurer un premier niveau de contrôle des informations transmises à ce drone qui devra « décider », nous lui donnons une capacité embarquée de contrôle des types de données, comme avec l'outil open-source Magika^[2], afin de vérifier la fiabilité et l'authenticité des informations communiquées et de déceler d'éventuelles compromissions.

Comment s'assurer que le réseau neuronal embarqué est

bien adapté à ce cadre d'emploi et qu'il ne comprend ni portes dérobées, ni stratégies décisionnelles malveillantes ?

La solution proposée consiste à faire appel à une IA de type Raisonnante Générale^[3] (IARG) pour induire automatiquement un SD transparent, composé de règles SI...ALORS à logiques continues, et révélant le comportement décisionnel caché du SD opaque, ici Magika. L'expert métier est alors en mesure d'auditer les règles induites pour débusquer un éventuel comportement malveillant. Cet ensemble de connaissances est facilement embarquable et actionnable avec une frugalité et une efficacité énergétique de premier ordre.

Au cœur de ce processus de conception de SD, la collaboration unique homme-machine forme une boucle vertueuse :

1. L'expert pose le problème de modélisation à résoudre (définition des prédicteurs potentiels, collecte et caractérisation des données) ;
2. L'IARG augmente l'expertise humaine en découvrant, par perfectionnement évolutif de ses stratégies d'induction, des connaissances robustes, tout en les exprimant de manière compréhensible, sous la forme de règles SI...ALORS à logiques continues (boucle d'induction évolutive).
3. L'expert corrige les erreurs et les biais détectés par l'IARG et relance l'induction pour découvrir des connaissances plus robustes (boucle d'induction corrective).

En outre, l'intelligibilité (i.e. l'interprétabilité) d'un SD à base de règles permet de faire face à un problème prégnant, notamment dans le monde de la défense : les données « opérationnelles » peuvent être partielles, incomplètes, entachées d'erreur, sans annotation, non représentatives, etc.

Dans ce contexte, comment prendre des décisions critiques sur la base de données lacunaires ?

L'intelligibilité est encore une fois le critère décisif. Elle apporte des avantages suivants :

- L'explicabilité de chaque décision permet la détection explicite des erreurs du modèle (biais, mauvais prédicteurs...) par l'expert métier.
- L'interprétabilité globale du modèle facilite la certification.
- En cas de changement de comportement du processus étudié, l'IARG émettra des refus de décision. L'expert métier devra alors analyser les nouveaux cas non reconnus pour les qualifier. La nouvelle base de données est aussitôt soumise à l'IARG pour découvrir le nouveau SD modélisant le nouveau comportement du processus (boucle d'induction adaptative).
- L'expert peut intégrer des règles métier saisies manuellement et demander à l'IAG d'évaluer leur robustesse en interaction avec les autres règles induites automatiquement. Cette approche permet l'annotation rapide de données non

annotées et le contrôle de la qualité de cette annotation par l'IARG.

Conclusion

La possibilité de travailler au niveau de la connaissance, et pas uniquement au niveau de la donnée, permet d'augmenter l'intelligence globale du SD et de l'expert métier. À partir d'un grand ensemble de données, les IA non-linéaires (réseau de neurones, arbre boosté, forêt aléatoire) sont capables de produire des SD robustes. Toutefois, ces SD sont opaques : ils ne peuvent être audités, ni certifiés. A l'inverse, certaines techniques d'IA sont capables de produire des SD intelligibles. Toutefois, ces SD sont non-robustes. Répondant, par conception, aux contraintes inhérentes à l'intégration de l'IA dans les systèmes critiques, les SD induits par l'IARG, sont à la fois robustes et intelligibles, autorisant leur usage éthique et responsable.

De la sécurité du Machine Learning



Teddy FURON

Directeur de Recherche au Centre de Recherche Inria de l'Université de Rennes / IRISA
Responsable de l'équipe-projet ARTISHAU



Étudier la sécurité intrinsèque de l'apprentissage automatique (Machine Learning, ML, en anglais), un des piliers de l'Intelligence Artificielle, est d'une grande importance. Le ML possède d'excellentes performances en termes de généralisation et de robustesse, ce qui nous donne un faux sentiment de sécurité. Or, généralisation, robustesse et sécurité sont trois concepts souvent confondus :

- La généralisation est la capacité du ML à fonctionner sur des données de test non vues mais de même nature que les données d'apprentissage.
- La robustesse est la capacité du ML à fonctionner sur des données de test de nature légèrement différente que les données d'apprentissage.
- La sécurité est la capacité de fonctionner sur des données délibérément perturbées par des attaquants, ou au moins de sentir que les conditions ne sont pas réunies pour fonctionner en toute sécurité.

L'intention des attaquants, mais surtout la recherche d'attaques efficaces tirant parti de leur connaissance du système ciblé, fait une grande différence entre la sécurité et la robustesse. La littérature récente fait état d'une multitude de dangers. Être un expert en sécurité de l'IA implique de connaître l'ensemble de ces menaces.

Le ML consiste à apprendre un modèle à partir de données d'entraînement et à appliquer ce modèle à des données de test pour produire un résultat. Les données d'entraînement, le modèle et les données de test sont des actifs qui doivent être protégés pour que le résultat soit fiable. Protéger signifie défendre certaines dimensions classiques de la Cybersécurité : la confidentialité, la privacy et l'intégrité.

Ainsi, 3 actifs × 3 valeurs donnent lieu à 9 familles de scénarios d'attaque.

1. Protection des données d'apprentissage

Les données d'entraînement possèdent une grande valeur en raison des efforts considérables investis dans leur collecte et leur annotation.

Confidentialité.

Comment apprendre un modèle quand on a des données sensibles mais pas les compétences ML nécessaires ? Le calcul multipartite^[1] ou le chiffrement homomorphique^[2] protègent les données tout en permettant des calculs. Cela revient à apprendre un modèle sur des données chiffrées. Une autre alternative est l'apprentissage fédéré où plusieurs propriétaires apprennent collectivement un modèle sans jamais révéler leurs données.

Privacy

Le modèle est le résultat d'un apprentissage sur les données d'entraînement, il dépend donc de ces données et laisse potentiellement filtrer des informations sur celles-ci. Une procédure d'apprentissage avec confidentialité différentielle^[3] offre un compromis entre l'utilité du modèle (par exemple, sa précision) et la confidentialité des données (moins de fuite d'informations). Ainsi le modèle ne divulgue pas (ou seulement dans une faible mesure contrôlable) si une donnée particulière faisait partie de l'ensemble d'apprentissage.

Intégrité

Un attaquant peut introduire une porte dérobée dans le modèle en altérant des données d'entraînement. En manipulant de manière cohérente certaines données associées à une classe cible, le modèle apprend, lors de son entraînement, à associer ces modifications spécifiques à cette classe. Ainsi, toute requête modifiée selon le même schéma activera la classe cible en sortie du modèle compromis. En résumé, l'attaquant parvient à prendre le contrôle du modèle à notre insu.

2. Protection du modèle

Confidentialité

Les réseaux neuronaux profonds fonctionnent sur des GPU, qui offrent une grande efficacité de calcul, mais présentent de sérieuses lacunes en matière de sécurité. Cela rend le vol d'un modèle relativement simple. Cette vulnérabilité constitue une menace majeure pour les systèmes embarquant de l'IA, qu'il s'agisse d'applications mobiles ou de véhicules militaires équipés de capteurs et d'intelligence artificielle. Une solution prometteuse consiste à obfusquer les paramètres du modèle et à mettre en place un protocole entre le GPU et un circuit d'exécution sécurisé, chargé de révéler uniquement le résultat final.

Privacy

Est-il possible d'identifier un modèle enfermé dans une boîte noire (accès via une API) simplement en l'interrogeant ? C'est possible dans une certaine mesure et cette information s'avère précieuse pour un attaquant cherchant à mener une attaque d'évasion, car la divulgation du modèle cible améliore grandement l'efficacité des attaques. Du point de vue du défenseur, l'identification de modèle pourrait également détecter le vol de modèles.

Intégrité

De nombreux modèles sont disponibles en open-source. L'utilisateur doit avoir la garantie qu'ils sont exempts de portes dérobées. Une autre menace est un cheval de Troie par micro-modifications des poids d'un modèle pour injecter un comportement dangereux.

3. Protection des données de tests

Confidentialité

L'inférence sur une requête sans la « voir » est possible. C'est le problème dual de l'entraînement sur des données chiffrées où c'est maintenant la requête est chiffrée et le modèle pré-appris. Dans ce contexte, deux approches principales sont en concurrence : le calcul multipartite et le chiffrement homomorphe. La première repose sur l'hypothèse que les différentes parties impliquées ne collaborent pas de manière malveillante, tandis que la seconde présente l'inconvénient d'être très lente.

Privacy

L'utilisateur souhaite inférer une donnée pour un problème de classification spécifique. Cependant, il ne souhaite pas que d'autres informations soient extraites de cette donnée. Est-il possible de « dépouiller » la donnée en ne soumettant au modèle que le strict minimum pour mener à bien son inférence ? Ce problème est très peu étudié dans le monde académique.

Intégrité

C'est le domaine des exemples adverses. L'attaquant ajoute une petite perturbation souvent indétectable à la requête pour tromper un classificateur. Ce sujet est la partie émergée de l'iceberg de la sécurité du ML, avec plus de 6 000 publications scientifiques au cours

des quatre dernières années. Et pourtant, aucune ligne de défense simple n'a été trouvée.

La sécurité du ML considère ce panorama riche de scénarios d'attaques. Les défenses visent à protéger de bout en bout l'apprentissage et l'utilisation d'un modèle. La sécurité du ML est partie intégrante de l'IA de confiance surtout lorsque l'IA est opérée dans un milieu hostile.

Intelligence Artificielle responsable et de confiance pour la Défense : vers de nouvelles pratiques d'ingénierie ?



Michel BARRETEAU

Expert en Intelligence Artificielle de confiance de bout-en-bout

Thales cortAlx Labs France



Patricia BESSON

Responsable du laboratoire Analyse et Raisonnement dans les Systèmes Complexes

Thales cortAlx Labs France



Fateh KAAKAI

Expert Safety & Chercheur en IA de Confiance

Thales cortAlx Labs France



Intégrant depuis plusieurs décennies des algorithmes de Recherche Opérationnelle, la Défense n'a pas échappé à l'impact de la dernière vague d'algorithmes d'Intelligence Artificielle (IA) considérant notamment l'apprentissage machine qui s'appuie principalement sur des données et moins sur des règles pensées par des humains. Du fait de risques opérationnels parfois plus importants que dans le monde civil (ex : aéronautique), une ingénierie d'IA de confiance digne de ce nom est requise pour qualifier ou certifier ces systèmes militaires critiques.

Le point de départ est le partage des principes (avec une base éthique) que l'on doit prendre en compte pour atteindre la confiance visée. Si le groupe d'experts de haut niveau européen en IA^[1] avait initié cette voie en 2019 en ciblant 4 principes éthiques (« respect for human autonomy, prevention of harm, fairness, explicability ») déclinés en 7 exigences clés (apparentés à des critères de confiance), l'OTAN a quant à elle annoncé ses 6 principes d'utilisation responsable de l'IA^[2] en 2021 : « Lawfulness, Responsibility and Accountability, Explainability and Traceability, Reliability, Governability, Bias mitigation ». Les ministères de la Défense des Etats-Unis ou du Royaume Uni par exemple ont choisi des principes du même ordre.

Ces principes vont guider la façon dont on va mesurer (quantitativement ou qualitativement) la confiance d'un système militaire intégrant de l'IA. Ces systèmes étant de nature complexe, ils nécessitent des processus de développement rigoureux adaptés à l'IA et intégrant aussi des mécanismes de surveillance du comportement de l'IA après déploiement. En effet, les différents types d'IA ne peuvent adresser le même niveau de risque. Ainsi, on ne peut imaginer s'appuyer

sur une IA générative de type grand modèle de langage (LLM) embarquée en phase d'exécution de mission opérationnelle, compte tenu de la forte probabilité actuelle de générer des hallucinations. Par contre, utiliser un algorithme de détection ou classification de cibles militaires basé sur de la vision peut être envisageable à condition de développer cet algorithme en assurant la confiance de bout-en-bout de son cycle de vie. Cela suppose une ingénierie qui considère les spécificités de ces IA.

On pense souvent à l'IA comme une application logicielle mais il ne faut pas oublier qu'elle s'exécute sur un composant matériel qui lui aussi doit être de confiance. La communauté RISC V^[3] contribue à cet effort en délivrant des architectures ouvertes. D'une manière générale, considérer des IA de type « boîte blanche » (connaissance des entrées et sorties du modèle, ainsi que du modèle ou de l'algorithme lui-même et de la façon dont il a été construit) accroît nativement la confiance. Néanmoins des méthodes ou processus outillés de confiance existent pour des IA « boîtes noires », améliorant par exemple leur explicabilité.

Ainsi, l'IA doit être considérée d'un point de vue ingénierie système (agnostique de toute implémentation IA), logicielle (en s'appuyant aussi sur des outils de développement matures et appropriés aux besoins de l'IA), matérielle (souvent avec des contraintes d'embarquabilité), mais aussi et surtout algorithmique à base d'IA et c'est là la nouveauté majeure.

Les récents travaux (confiance.ai^[4] rassemblant notamment plusieurs industriels de systèmes critiques, EASA et SAE G34 / EUROCAE WG114^[5] avec le futur standard ED-324/ARP6983) décrivent



les activités d'IA de confiance au sein du cycle de développement de l'IA se concentrent dans un premier temps sur une IA du type apprentissage machine supervisé. Cela signifie que les données sont au cœur du problème et nécessite en conséquence des activités de confiance relatives à leur viabilité, pertinence, représentativité ...

Comparativement au cycle de vie en V classique [6], ces travaux considèrent désormais un cycle en W : le premier V se concentre justement sur l'algorithme d'IA (conception, développement, évaluation) en considérant dans un premier temps une plateforme hôte ; le second V est plus classique puisqu'il correspond à l'intégration, vérification, validation de cet algorithme au sein d'une plateforme cible (souvent embarquée). Une autre différence caractéristique clé de ce nouveau cycle de développement de l'IA touche aux données puisqu'on est dans le cas d'un apprentissage machine. On reprend alors une notion introduite par le monde automobile il y a quelques années : le domaine opérationnel de conception. Il permet de décrire les conditions dans lesquels le comportement du système d'IA sera assuré (ex : conditions météorologiques). Il faut néanmoins caractériser cette enveloppe de fonctionnement multidimensionnelle afin de vérifier que le système

d'IA fonctionnera nominalement quel que soit le cas de figure prévu au sein de cette enveloppe ou à proximité. A terme, ce nouveau cycle de vie devra être étendu pour considérer d'autres types d'IA comme par exemple l'apprentissage par renforcement, l'IA hybride, distribuée ou générative.

Le challenge de l'adaptation de l'ingénierie de l'IA pour la Défense ne s'arrête pas là : cette dernière doit maintenant faire face à une accélération de son processus de qualification d'IA. En effet les stratégies militaires sont très ingénieuses pour tromper l'ennemi au travers de multiples idées de camouflage. Une IA qui doit détecter des armements ou forces ennemis doit être fréquemment remise à jour pour intégrer de nouvelles photos montrant ces dissimulations. A cela il faudrait ajouter les problématiques de confiance en termes d'autonomie. Mais ce ne sont là que quelques exemples de challenges que l'ingénierie de l'IA pour la Défense devra adresser sans attendre.

Fiabilité de l'IA en contexte militaire : Trouver l'équilibre entre autonomie et supervision humaine



Vincent BLOT

Doctorant en incertitude des modèles d'IA
Capgemini Invent, LISN, CNRS



Avec l'augmentation de la quantité de données recueillie sur les terrains militaires (que ce soit via une augmentation du nombre de capteurs ou par l'amélioration de la qualité des données), se pose la question de leur analyse et de l'équilibre entre traitement automatique et traitement humain.

Prenons l'exemple d'un sous-marin. Lorsqu'il est en mer, il enregistre en permanence tous les bruits qui l'entourent. Cette quantité massive de données peut représenter jusqu'à plusieurs pétaoctets par jour.

Les opérateurs, appelés « oreilles d'or », analysent ces sons pour identifier des menaces potentielles mais ils ne sont plus en mesure de tout analyser par eux-mêmes. L'intelligence artificielle (IA) entre alors en jeu pour aider à traiter ces flots d'informations. Il demeure essentiel de savoir distinguer les situations où l'on peut se fier à l'IA de celles où l'intervention humaine est indispensable. Un contrôle totalement automatique peut-être sujet à un trop grand nombre d'erreurs tandis qu'une analyse totalement humaine est impossible.

Comment, alors, faire en sorte de concentrer l'expertise humaine sur les analyses les plus complexes afin d'optimiser l'analyse de l'IA en la focaliser sur des exemples plus simples et moins sujets à erreur ?

Dans quelles situations utiliser l'IA ?

L'IA excelle dans des tâches où la rapidité et la constance sont essentielles. Contrairement aux humains, elle ne se fatigue pas, n'oublie rien et peut analyser des données à une vitesse époustouflante. Dans le domaine militaire, cela signifie identifier des menaces ou anomalies bien plus rapidement qu'un humain ne pourrait le faire seul. De plus, une IA bien conçue peut réduire certaines erreurs causées par des biais ou des distractions humaines.

Cependant, aucune IA n'est parfaite ou omnisciente. Elle reste confinée à son domaine de validité, c'est-à-dire le domaine restreint qui est représenté par les données d'entraînement. Prenons l'exemple de la détection de chars. Lorsqu'un algorithme apprend cette tâche, cela ne signifie pas qu'il sera capable de la réaliser quel que soit le terrain dans lequel il est utilisé. Si les données d'entraînement de ce modèle sont composées à 90% de chars dans le désert et les 10% restants dans d'autres types d'environnements, il est fort probable que la performance de l'algorithme soit moindre dans une région enneigée. De plus, même si l'algorithme est exploité dans son domaine de validité, cela ne l'empêchera pas de commettre des erreurs.

Une question survient alors : comment faire en sorte qu'une IA puisse rendre le contrôle à l'humain lorsqu'elle n'a pas la capacité de prendre la bonne décision, et ce quel qu'en soit la raison ?

La quantification d'incertitude en IA

L'un des principaux défis des IA actuelles réside dans leur incapacité à évaluer efficacement leur propre incertitude. Cette dernière est due à deux facteurs principaux : l'entraînement du modèle (incertitude épistémique : le modèle n'a pas vu assez de données durant l'entraînement) et la qualité des données sur les terrains d'opération (incertitude aléatoire : images de moindre qualité ou obstruées). On peut considérer que l'incertitude épistémique peut théoriquement être réduite à zéro en entraînant le modèle avec plus de données, ou en utilisant des modèles plus performants, néanmoins, la quantité de données nécessaire pour la supprimer est considérable. Quant à l'incertitude aléatoire, elle provient de certaines conditions réelles qui compliquent la réalisation de prédictions : des pluies fortes en mer qui ajoutent un bruit de fond pour les capteurs des sous-marins, du brouillard pour la détection de chars sur des images...

Finalement l'incertitude aléatoire est aussi celle à laquelle les opérateurs humains doivent faire face.

Pour répondre à ce besoin, plusieurs méthodes ont été développées afin de quantifier l'incertitude associée aux prédictions des modèles d'IA. Parmi ces méthodes, les prédictions conformes (Conformal Prediction) constituent une approche prometteuse qui fournit des intervalles de confiance valides sous des hypothèses minimales. Ces techniques permettent d'encadrer les prédictions du modèle en fournissant des garanties formelles sur leur fiabilité : c'est-à-dire des garanties que la vraie valeur se situe dans l'intervalle dans 90% des cas.

Quand l'humain doit-il reprendre le contrôle ?

La décision de laisser une IA opérer seule ou d'impliquer un humain dépend du niveau d'incertitude prédit par le système et du contexte. Dans les scénarios critiques, où une erreur pourrait avoir des conséquences graves, il est essentiel que l'humain reste dans la boucle.

Pour cela, l'expertise humaine joue un rôle central. Une IA ne doit pas remplacer les experts, mais plutôt les compléter. Par exemple, un analyste pourrait utiliser l'IA pour pré-trier des données et se concentrer sur les cas les plus incertains. Cette collaboration nécessite également des méthodes rigoureuses d'évaluation. Les seuils d'alerte – quand l'IA demande l'intervention d'un humain – doivent être définis avec précision et ajustés en fonction des résultats sur le terrain. Intervient alors un compromis entre la précision du système d'IA (ses réponses sont-elles toujours correctes ?) et le rappel (Quel est le pourcentage de données traitées par l'IA sans validation humaine ?). Ces deux objectifs sont généralement complexes à concilier simultanément, et il est essentiel de trouver un équilibre adapté, tenant compte à la fois de la capacité humaine à gérer les données et de l'efficacité de l'IA dans l'exécution de cette tâche.

Dans un scénario de maintenance prédictive trois différents niveaux de prédictions peuvent être mis en œuvre :

- Les pièces pour lesquelles l'IA est certaine à 99 % qu'un remplacement rapide est nécessaire : les appareils sont immédiatement arrêtés pour permettre le changement des pièces défectueuses.
- Les pièces pour lesquelles l'IA est certaine à 99 % qu'aucune maintenance n'est requise : les appareils concernés peuvent continuer à fonctionner normalement sans intervention.
- Toutes les autres pièces sur lesquelles l'IA ne peut pas fournir une prédiction suffisamment confiante : un contrôle humain des données effectué, et les opérateurs décident ou non d'immobiliser les appareils.

Conclusion

L'intelligence artificielle apporte une puissance et une rapidité incomparables dans de nombreux domaines, y compris le secteur militaire. Cependant, il est illusoire de croire qu'une IA atteindra un jour la perfection. Les erreurs font partie inhérente des systèmes d'IA, et il

est crucial d'accepter cette réalité pour développer des solutions adaptées.

Dans un environnement aussi sensible que le domaine militaire, un contrôle aveugle laissé à une IA serait irresponsable. En combinant des méthodes avancées de quantification d'incertitude avec l'expertise humaine, il est possible de concevoir des systèmes hybrides fiables. L'objectif n'est pas de supprimer l'erreur, mais de la minimiser et de la rendre acceptable. En trouvant le bon équilibre entre technologie et intervention humaine, nous pouvons construire des IA de confiance capables d'assister efficacement les opérateurs dans des missions critiques.

Les principales vulnérabilités des LLMs face aux attaques malveillantes



Louis JOURDAIN

Ingénieur IA/NLP
Groupe Chapsvision



Christian LANGEVIN

Chief Product Officer IA/NLP
Groupe Chapsvision



L'intégration de l'IA générative (GenAI), et notamment des LLMs (Large Language Models), dans les entreprises et grandes institutions offre un potentiel immense : automatisation des tâches complexes, gain de temps pour les utilisateurs, et innovation des services. Cependant, ces outils introduisent de nouvelles vulnérabilités dans les systèmes d'information. En effet, qu'il s'agisse des données traitées, des modèles eux-mêmes, voire des logiciels incorporant des composants de GenAI, ces innovations ouvrent des fenêtres d'attaque exploitées par des acteurs malveillants. Ces derniers peuvent perturber le fonctionnement des services ou exfiltrer des données sensibles. Identifier et mettre en œuvre des contre-mesures efficaces est désormais une priorité majeure pour la cybersécurité.

A. Qu'est-ce qu'un système d'IA fiable ?

L'intégration des LLMs dans des systèmes complexes ouvre de nouvelles voies pour des attaquants potentiels, capables de perturber le bon fonctionnement d'un service. Un système génératif est attendu sur quatre aspects fondamentaux :

- **Fiabilité** : il doit exécuter avec précision et qualité la tâche assignée.
- **Respect des formats prescrits** : essentiel pour assurer l'interopérabilité dans des chaînes d'agents IA autonomes.
- **Confidentialité** : Un système d'IA doit garantir la sécurité des données personnelles de l'utilisateur et de l'institution.
- **Performance** : il doit maintenir un temps de réponse acceptable, même sous forte charge.

Toute altération compromettant ces attentes — qu'elle se manifeste par des résultats incohérents, des latences accrues, ou des formats non conformes — peut perturber le système avec des conséquences variables, allant d'une simple inefficacité à des dysfonctionnements critiques.

B. Les différents types d'attaques

Les experts en cybersécurité ont développé une taxonomie des attaques visant les systèmes génératifs, notamment selon les techniques utilisées par les attaquants. Cette taxonomie est bien sûr en constante évolution avec l'apparition de nouvelles menaces. A date les principaux types d'attaques répertoriées sont :

- **jailbreaking** : Tentative par des techniques de prompting spécifiques de désaligner le LLM de son paramétrage initial (production de contenu toxique, de biais...) ou de sa mission actuelle (« oublie toutes les instructions précédentes et ...»), voire de générer un contenu dangereux si exécuté, comme des redirections vers des url frauduleuses. Ce genre d'attaque impacte la fiabilité des systèmes.
- **prompt leaking** (fuite de prompt) : Technique visant à « sonder » le système génératif en lui faisant avouer son system prompt. Souvent phase de repérage pour dévoiler des secrets industriels ou pour avoir de meilleurs outils pour améliorer les autres formes d'attaques pas prompt injection. Le but est d'exfiltrer des données confidentielles.
- **attaque par code exécutable** : L'attaquant pousse le LLM à générer du code malicieux puis à l'exécuter ; nota : seulement dans des environnements où le LLM a la permission de générer et d'exécuter du code (coding agents). Cela impacte la fiabilité mais aussi l'intégrité même du système.
- **data poisoning (empoisonnement des données)** : Modification des données d'entraînement, de fine tuning d'un modèle ou des données qu'il requête lors de l'exécution de ses tâches de façon à influencer son comportement général (une modification de 0.1 % du corpus d'entraînement

peut avoir des effets notables et discrets, ce qui est dangereux dans des systèmes de classification utilisés pour la prise de décision par exemple). La fiabilité peut être drastiquement impactée.

- **attaque par DDOS (Distributed Denial of Service ou attaque en déni de service)** : Faire tourner un système d'IA est coûteux et les institutions disposent de ressources de calcul limitées. Dès lors requêter trop souvent les services ou forcer les LLM à générer du texte à l'infini peut fortement perturber les performances pour tous les utilisateurs.

Là où auparavant, attaquer une infrastructure nécessitait des compétences informatiques avancées et la maîtrise de plusieurs langages de programmation et des protocoles réseaux, les attaques par injection de prompts ne nécessitent que peu de compétences techniques et on peut en élaborer à partir de ressources sur le prompt engineering que l'on trouve en ligne. Les attaques plus avancées comme le data poisoning ou l'injection de prompt indirecte requièrent toutefois une meilleure connaissance de l'infrastructure attaquée (accès aux points d'indexation des données et à l'architecture du service qu'on veut déstabiliser). Ces attaques sont donc plus accessibles à des menaces internes ou à des acteurs ayant les capacités techniques de sonder en profondeur l'architecture data de la cible.

Cette taxonomie n'est évidemment pas exhaustive. D'autres types d'attaques plus techniques existent, telles que la falsification de requêtes serveur ou l'exploitation des sandbox^[1] dans des contextes d'agents autonomes. Par ailleurs, cet article se concentre sur les méthodes d'attaque et ne prend pas en compte les vulnérabilités intrinsèques des modèles, comme les hallucinations^[2] ou leurs problèmes d'alignement avec les préférences humaines.

C. Face à la malveillance, quelles solutions ?

Pour se prémunir des attaques, il est essentiel de mettre en place des mécanismes de filtrage appliqués à la fois :

- Aux entrées utilisateur, pour s'assurer que les interactions respectent les attentes et prévenir les comportements perturbateurs.
- Aux sorties du système, pour détecter tout détournement potentiel de son comportement.

Ces filtres peuvent être basés sur des tests automatisés, inspirés des tests unitaires en programmation, utilisant des règles, des expressions régulières ou des modèles d'IA spécialisés dans la détection d'anomalies. Par exemple, l'injection de prompt peut être identifiée par des modèles de classification entraînés à reconnaître des instructions suspectes.

En revanche, l'injection de prompt indirecte est plus complexe à détecter, notamment lorsque des données corrompues ou erronées sont insérées dans des bases de données. Ces attaques peuvent être subtiles, comme l'ajout de valeurs incorrectes dans une application financière. Pour limiter ces risques, plusieurs mesures doivent être prises :

1. **Contrôle strict des droits d'accès**: Chaque insertion dans la base de données doit être autorisée et effectuée par des utilisateurs disposant des privilèges appropriés.
2. **Surveillance active** : Mettre en place un système de monitoring permettant d'identifier rapidement les documents problématiques ou les comportements suspects.
3. **Validation rigoureuse des données** : Les nouvelles entrées doivent être systématiquement vérifiées, notamment par des processus automatisés ou manuels pour s'assurer de leur conformité.

Ces pratiques combinées permettent de réduire les surfaces d'attaque et de mieux protéger les systèmes génératifs contre des tentatives malveillantes, qu'elles soient directes ou indirectes. Elles ne sont bien sûr pas exhaustives car chaque type d'attaque demande sa contre mesure spécifique, mais constituent un ensemble de bonnes pratiques à adopter pour sécuriser les systèmes d'IA générative.

La cybersécurité de l'IA au cœur du champ de bataille



Lucie DUCHATEAU



Nicolas CHARBONNIER



Hugo MANIA



Division industrie et technologies
Agence nationale de la sécurité des systèmes d'information (ANSSI)

L'Agence nationale de la sécurité des systèmes d'information (ANSSI) déploie une politique globale de cybersécurité et en assure la coordination à l'échelle interministérielle. Les travaux de l'ANSSI portent notamment sur la sécurisation des systèmes d'IA pour renforcer le niveau de cybersécurité des organisations publiques et privées. Dans le cadre de ce livre blanc, nous appliquons les axes de réflexion et recommandations de cybersécurité de l'ANSSI au contexte de la défense.

L'IA est déjà un acteur incontournable du champ de bataille et ses capacités sont désormais pleinement intégrées aux stratégies des armées sur le terrain – notamment pour ce qui est de l'analyse de signaux (images, sons, signaux électromagnétiques), de l'aide à la décision, ou encore de la gestion de systèmes d'armes complexes. Dans ce contexte opérationnel particulièrement délicat car porteur de multiples contraintes d'environnement, il est particulièrement important de s'assurer de la cybersécurité des systèmes d'IA (SIA), c'est-à-dire de l'ensemble de composants matériels et logiciels visant à intégrer et exécuter un ou plusieurs modèles d'IA dans un système d'information.

Nous étudierons, dans cet article, les éléments pouvant avoir un impact sur la cybersécurité des SIA possédant les caractéristiques suivantes :

- implémentation dans des composants matériels locaux (on parle d'« IA embarquée ») et exposition à des contraintes opérationnelles fortes pouvant impacter leur fonctionnement : températures extrêmes, vibrations, interférences électromagnétiques ou limites de bande passante réseau (exemple : drones de combat) ;
- gestion de capacités temps-réel, notamment pour fournir une aide à la décision rapide sur le champ de bataille (exemple : systèmes de détection d'objets et de cartographie du terrain).

Disponibilité du SIA

La disponibilité du SIA est un réel enjeu de cybersécurité qui amène à se poser les questions suivantes :

- Quels sont les composants du SIA qui vont réaliser le calcul d'inférence^[1]? Où sont-ils localisés physiquement ? Le calcul est-il distribué ?

Recommandation : Cartographier l'ensemble des composants logiciels et matériels des SIA ainsi que les dépendances entre ces composants^[2].

- L'architecture du SIA est-elle complètement autonome ou bien s'appuie-t-elle sur une interconnexion avec des services centralisés situés hors de l'environnement contraignant ?

Recommandation : Sécuriser les interconnexions entre les éléments de terrain et les services centralisés. Assurer une redondance des réseaux de transport avec des technologies distinctes.

- La continuité des applications est-elle assurée, en cas d'indisponibilité ou d'altération de son fonctionnement ? En cas de réponses erronées du modèle, est-il prévu un système de secours ?

Recommandation : Mettre en place des modes dégradés des SIA en cas de dysfonctionnement. Prévoir un mode dégradé sans IA.

Protection physique du matériel

En second lieu, la protection physique du matériel est également un enjeu à plusieurs titres :

- La protection contre le vol des équipements : les modèles d'IA peuvent être considérés comme des éléments critiques à protéger comme des secrets industriels. Il faut considérer les attaques d'extraction de modèles^[3] en cas de perte ou de vol ;

Recommandation : *Mettre en place si nécessaire des mécanismes de protection en confidentialité des SIA (chiffrement, effacement sécurisé à distance, etc.).*

- La protection contre des attaques par canaux auxiliaires^[4] : même si elles ne sont pas propres aux SIA, celles-ci doivent être prises en compte dès la conception des modèles d'IA^[5] afin de prémunir des risques d'extraction de modèles ;

Recommandation : *Implémenter des fonctions de sécurité pour réduire les risques d'extraction des modèles d'IA via des attaques par canaux auxiliaires.*

- La protection contre des attaques en déni de service^[6] : celles-ci peuvent saturer la puissance de calcul utilisée par les modèles d'IA et augmenter les délais de réponse, ce qui est problématique pour des systèmes temps-réel.

Recommandation : *Mettre en place un contrôle des entrées envoyées aux modèles d'IA. Dédier des composants physiques à des calculs considérés comme critiques.*

Conception et entraînement des modèles d'IA

La protection contre des attaques sur la chaîne d'approvisionnement est un impératif critique. Des risques importants spécifiques aux SIA sont à prendre en compte lors de la phase d'entraînement des modèles d'IA, par exemple l'empoisonnement des jeux d'entraînement, l'implémentation de portes dérobées, etc.

Par ailleurs, la phase d'entraînement des modèles est souvent effectuée dans des environnements de développement, par définition plus ouverts et potentiellement plus vulnérables aux cyberattaques. Il convient donc de se poser les questions suivantes :

- Qui dispose des droits d'accès sur le modèle d'IA lors de l'entraînement ?
- Comment tester et évaluer les jeux de données d'entraînement ?
- Comment s'assurer de la pleine intégrité des modèles, une fois ces derniers entraînés ?
- Comment réaliser un déploiement sécurisé de ces modèles sur les équipements finaux ?
- Comment vérifier régulièrement l'intégrité des modèles d'IA, une fois ceux-ci en production, afin de s'assurer qu'ils n'ont pas été compromis ?

Recommandations :

- *Appliquer le principe de moindre privilège dans la gestion des accès aux SIA, incluant la phase de développement et d'entraînement des modèles.*
- *Évaluer le niveau de confiance des jeux de données d'entraînement (provenance, qualité, exhaustivité, représentativité, etc.) et s'assurer de leur intégrité avant toute opération.*
- *Implémenter des fonctions de contrôle réguliers de l'intégrité des SIA et une remontée d'alerte en cas de modification non légitime.*

Le maintien en condition opérationnel et de sécurité des SIA doit être discuté dès la phase de conception, pour pouvoir être en mesure de déployer des correctifs de manière réactive, en cas de dysfonctionnement. Or, ce déploiement peut poser des contraintes lorsqu'il est réalisé à distance, depuis des services centralisés. Il faut donc prévoir un processus de réentraînement des modèles d'IA qui s'inscrit en cohérence avec les contraintes physiques des environnements. Les opérateurs doivent pouvoir être en mesure de réaliser une supervision de sécurité des SIA (locale ou distante) et éventuellement être en mesure de faire des retours arrière vers des versions précédentes du modèle d'IA en cas de dysfonctionnement après une mise à jour.

Conclusion

Si la cybersécurité de l'IA consiste principalement à traiter des problématiques de cybersécurité standard, elle induit également de nouveaux risques^[7] notamment liés à une possible compromission lors de la phase d'entraînement. Le caractère probabiliste des réponses apportées par les modèles d'IA implique également une grande vigilance, et la mise en place des processus métier incluant des contrôles humains réguliers, ainsi qu'une évaluation des risques en cas d'incident. Dans des environnements de défense contraignants, ces risques peuvent être parfois difficiles à traiter, la disponibilité des composants de calcul ou des interconnexions réseaux n'étant pas toujours garantie. Afin d'assurer la confiance dans les SIA, il peut être opportun de mener une évaluation de cybersécurité au regard des risques et menaces identifiés.

Difficulté de la régulation de

L'IA

dans la défense



Systemes d'Armes Létales Autonomes

– aspects juridiques



Anne-Laure GAILLARD

Avocate en droit du numérique passionnée par les sujets d'éthique et de souveraineté
Withlaw Avocats



« Le déploiement de l'IA est donc un défi multidimensionnel, qui doit donc être compris pour ce qu'il est : l'instrument de la puissance au XXI^e siècle »^[1]. L'Europe s'est récemment dotée d'un règlement établissant des règles harmonisées concernant l'intelligence artificielle. Toutefois, ce règlement n'a vocation à régir que les usages civils de l'IA au sein de l'Union européenne et ne s'applique pas aux systèmes d'IA utilisés à des fins militaires, de défense ou de sécurité nationale.

Se pose alors la question du régime juridique applicable aux usages militaires de l'IA. Le droit international humanitaire, notamment les Conventions de Genève (1949) et leurs protocoles additionnels, impose des règles applicables à toutes les armes, y compris les SALA. Ces textes établissent des principes fondamentaux, tels que :

- Le principe de distinction, qui impose aux belligérants de distinguer les combattants et la population civile, ainsi que les biens civils et les objectifs militaires ;
- Le principe d'humanité, qui vise à alléger et, dans la mesure du possible, éviter les maux superflus engendrés par le recours à la force ;
- Le principe de discrimination ou de précaution, qui intervient lorsqu'une opération militaire présente des risques pour les populations civiles ;
- Le principe de proportionnalité impliquant que les attaques ne doivent pas causer des dommages excessifs aux civils par rapport à l'avantage militaire attendu ;
- Le principe d'interdiction des maux superflus, qui interdit de causer des dommages ou des souffrances qui ne sont pas nécessaires pour atteindre des buts strictement militaires et l'affaiblissement du camp adverse.

Si, d'un point de vue éthique et juridique, de nombreux systèmes d'IA utilisés par l'armée peuvent

probablement être régis par des principes similaires à ceux qui sont retenus, selon le cas, dans le domaine civil ou, s'agissant d'armes, par le droit international humanitaire, des systèmes tels que les Systèmes d'Armes Létales Autonomes (SALA), ou "killer robots", eux, soulèvent à l'échelle mondiale des questions beaucoup plus spécifiques.

Malgré l'existence de nombreuses initiatives et forums internationaux de discussions et de négociation sur la régulation des SALA, à commencer par les travaux du groupe d'experts gouvernementaux (GEG) au sein de la Convention sur certaines armes classiques (CCAC)^[2], aucun accord contraignant n'a encore été adopté. En effet, la principale difficulté reste l'absence de consensus entre des pays ayant des intérêts stratégiques divergents.

Ainsi, le concept de « contrôle humain significatif » est une exigence relativement consensuelle dans les débats internationaux sur les SALA. Cependant, ce consensus apparent masque des différences significatives dans leur interprétation et application par les pays, influencés par leurs doctrines militaires, leur niveau technologique et leurs objectifs stratégiques.

L'autonomie dans les SALA se réfère à la capacité d'un système à effectuer des tâches sans intervention humaine directe, en utilisant des algorithmes pour prendre des décisions basées sur des données reçues en temps réel, tandis que le contrôle humain significatif implique qu'un opérateur humain doit pouvoir superviser et intervenir dans le processus décisionnel critique, en particulier lors de l'identification et de l'engagement des cibles.

A titre d'exemples :

- Le Parlement européen adopte une position stricte, exigeant que toutes les décisions critiques

(comme le choix des cibles) restent sous contrôle humain direct^[3]. Cela exclut les SALA capables d'engager des cibles de manière entièrement autonome ;

- Cependant, des États membres tels que la France estiment que « Le débat sur les SALA ne doit pas parasiter les efforts entrepris dans le domaine de l'autonomie des systèmes d'armes, au risque d'un déclasserement technologique, industriel et stratégique »^[4] et retient une définition très restrictive des SALA, insistant sur leur caractère légal et autonome, qu'elle distingue notamment des drones armés^[5]. En 2020, la diplomatie française affirmait ainsi que les SALA, définis comme étant « des systèmes d'armes qui seraient capables de recourir à la force létale sans aucune forme de supervision humaine », n'existaient pas^[6]. Le Comité d'éthique de la défense fait quant à lui une différence entre les systèmes d'armes létaux autonomes et les systèmes d'armes létaux « intégrant de l'autonomie » mais demeurant sous maîtrise de l'humain (les « SALIA ») et rejette fermement l'usage des SALA^[7] ;
- à ce jour, et sans préjuger des évolutions de leur doctrine liées au retour de Trump à la présidence, les États-Unis mettent en avant l'idée de « contrôle humain significatif » mais interprètent ce concept de manière flexible. Selon la Directive 3000.09^[8], ce contrôle peut inclure des « niveaux appropriés de jugement humain » à différents stades (conception, déploiement ou activation), plutôt qu'une intervention humaine directe en temps réel.

Les guerres en Ukraine ou dans la Bande de Gaza ont popularisé l'image des drones comme outils indispensables sur le champ de bataille, mais elles ont aussi alimenté des craintes liées à la prolifération et à l'utilisation irresponsable des SALA. Ces conflits, où des drones et systèmes semi-autonomes ont été largement déployés, ont mis en lumière à la fois les avantages stratégiques et les défis éthiques de ces armes. Ils modifient potentiellement les approches des grandes puissances en matière de régulation et de développement de ces technologies^[9].

Cela pourrait diviser davantage les grandes puissances, entre partisans d'une régulation stricte et ceux privilégiant la supériorité stratégique (notamment les États-Unis, la Chine et la Russie).

Trust by Design : comment être éthique et un leader européen de l'IA



François MATTENS

VP Affaires publiques et partenariats stratégiques
XXII



L'intelligence artificielle est en train de redéfinir le champ de bataille moderne, remodelant les stratégies de défense et de sécurité. L'Europe, et en particulier la France, a l'opportunité de s'imposer comme un leader technologique à condition de concilier innovation et éthique. La vision par ordinateur, domaine d'excellence français, illustre parfaitement cette dynamique. Intégrer la confiance dès la conception est un impératif pour garantir un usage maîtrisé et accepté par la société.

La vision par ordinateur, un Game changer pour la défense et la sécurité

La vision par ordinateur transforme radicalement la capacité d'analyse et de décision des opérateurs de défense. En analysant et interprétant en temps réel des flux visuels complexes, ces systèmes permettent une réaction rapide et adaptée à des situations critiques. Son application dans la défense s'impose comme un levier stratégique, avec des usages tels que la surveillance d'infrastructures sensibles, l'identification automatique des menaces et l'optimisation des opérations militaires. La détection d'intrusions, l'analyse avancée des images satellites et le soutien aux forces de terrain illustrent son potentiel pour améliorer la protection des soldats et des installations.

L'intégration de l'IA dans la reconnaissance faciale et le traitement d'images peut aussi aider à identifier rapidement des individus dangereux ou à localiser des cibles stratégiques. Sur le terrain, l'exploitation des flux vidéo permet d'améliorer la surveillance en milieu urbain ou en zones de conflit, offrant aux commandements une vision plus fine et réactive des menaces. L'efficacité des drones autonomes équipés de vision par ordinateur pour les missions de reconnaissance ou de soutien logistique démontre encore une fois l'impact de cette technologie. L'enjeu est clair : maximiser la réactivité et la précision tout en réduisant l'exposition humaine aux risques.

Arrêtons de réguler par la peur : pour une approche équilibrée de la réglementation

L'Europe veut encadrer l'IA pour en garantir un usage responsable, ce qui est légitime. Toutefois, à force de multiplier les contraintes, il existe un risque de freiner l'innovation et d'handicaper les acteurs européens face à la concurrence internationale. Une régulation excessive pourrait créer un déséquilibre avec les acteurs américains et chinois, qui avancent avec moins de barrières. Trop souvent, la réglementation est motivée par une crainte disproportionnée des risques potentiels. Or, il est impossible d'anticiper et d'éliminer tous les dangers sans limiter l'innovation. Plutôt qu'un cadre rigide, l'Europe doit adopter une régulation adaptative, fondée sur des audits réguliers et une évaluation des impacts réels.

Les craintes autour de l'IA – biais algorithmiques, risques de surveillance abusive, erreurs décisionnelles – ne doivent pas mener à une paralysie technologique. La France est passée de la 13^e à la 5^e place dans le Global AI Index^[1], illustrant une montée en puissance rapide de son écosystème. Cet essor doit être soutenu par une régulation qui accompagne l'innovation au lieu de la freiner, comme le souligne le rapport interministériel sur l'IA^[2] qui plaide pour une attractivité renforcée et des mesures d'accélération pour l'industrialisation des technologies. Il est impératif d'accompagner l'innovation plutôt que de la brider. La transparence et l'explicabilité doivent être au cœur de la régulation, mais sans imposer des freins inutiles au développement. Un dialogue constant entre les industriels, les institutions publiques et les citoyens est essentiel pour instaurer une approche équilibrée. Encourager la certification des algorithmes et le contrôle des biais dans un cadre collaboratif permettrait une adoption plus sereine et efficace.

L'intelligence artificielle n'est pas une menace en soi ; elle devient un atout lorsqu'elle est maîtrisée et encadrée intelligemment.

La question n'est pas de restreindre son développement, mais de garantir que son utilisation respecte les valeurs fondamentales de la société, notamment en matière de protection des libertés individuelles et de souveraineté technologique.

Une IA éthique et responsable au service de la souveraineté technologique

L'adoption d'une approche proactive pour garantir la fiabilité et l'acceptabilité des solutions IA est essentielle. Cela implique l'intégration dès la conception de principes fondateurs tels que la bienfaisance, la non-malfaisance, l'autonomie, la justice et la transparence. Investir massivement dans la recherche et la sensibilisation des parties prenantes est un moyen de garantir une IA exempte de biais et alignée avec les normes éthiques européennes. Le respect des règles du RGPD et du futur AI Act doit être vu non pas comme une contrainte, mais comme un atout stratégique pour assurer la confiance du public et des institutions.

Le développement d'outils d'explicabilité est également un axe central. Les solutions mises en œuvre doivent permettre une traçabilité claire des décisions algorithmiques, assurant ainsi que chaque utilisateur – des forces de l'ordre aux citoyens – puisse comprendre et vérifier l'utilisation des systèmes d'IA. De plus, un audit continu des performances et des limites de ces technologies est nécessaire pour éviter toute dérive ou distorsion de résultats.

L'intelligence artificielle est une révolution comparable à l'invention de la poudre à canon ou de l'atome. Comme le rappelait Benjamin Franklin : « Un peuple prêt à sacrifier un peu de liberté pour un peu de sécurité ne mérite ni l'une ni l'autre et finit par perdre les deux. » Si l'Europe veut en tirer parti, elle doit trouver un équilibre entre régulation et innovation. La vision par ordinateur est un des leviers clés de cette transformation, et une approche responsable et performante est possible. Il est temps d'adopter une posture proactive, pragmatique et ambitieuse pour faire de la France et de l'Europe des leaders mondiaux de l'IA de confiance. L'avenir de la souveraineté technologique européenne dépend de la capacité à allier innovation et responsabilité, sans céder à une vision trop restrictive ou purement commerciale de l'intelligence artificielle.

Comment garantir qu'une IA respectera les principes de proportionnalité et de distinction



Olivia BREYSSE

Experte indépendante en IA et cybersécurité
Collège Numérique France 2030



L'éthique de la Défense est un équilibre subtil entre impératifs opérationnels, respect des droits humains et des droits internationaux. Un équilibre actuellement remis en question par la généralisation de l'usage de l'intelligence artificielle dans les systèmes de défense qui gagnent certes en précision, en efficacité et en autonomie, mais qui doivent en retour faire face à des risques accrus de dérives telles que :

- Des erreurs opérationnelles pouvant causer des dommages collatéraux et des pertes civiles,
- Une violation des droits humains en discriminant ou en ciblant involontairement des populations spécifiques,
- Une exacerbation des conflits ou la création de nouvelles tensions, en raison d'une mauvaise identification des comportements adverses.

Un contrôle humain indispensable

Aussi, face aux dommages que pourrait causer un système d'IA doté d'autonomie décisionnelle, le Comité d'Éthique de la Défense (COMEDDEF) préconise de maintenir systématiquement un contrôle humain sur les décisions critiques. Surtout, lorsqu'il s'agit de systèmes d'armes létaux^[1].

Néanmoins, une autonomie même partielle des systèmes d'IA questionne encore les principes de proportionnalité et de distinction du droit international. En effet, comment vérifier que le système n'entraînera pas des conséquences excessives pour les populations civiles, par rapport à l'avantage militaire attendu ? Comment s'assurer qu'un système identifiera correctement une cible militaire et ne commettra pas une erreur de classification menant à une attaque contre des civils ?

Des biais ancrés dès la conception

Les biais algorithmiques sont difficilement décelables

et même inévitables au moment de la conception d'un programme. Nous en avons pris conscience, bien avant l'utilisation commerciale des modèles d'IA, avec de simples systèmes automatisés qui amplifiaient déjà des discriminations opérées dans la réalité^{[2], [3]}. Mais, l'absence de neutralité des données historiques n'est pas la seule source de biais dans les systèmes d'IA :

- Un mauvais échantillonnage des données d'apprentissage peut entraîner ensuite une mauvaise généralisation du modèle, comme le ciblage d'un rassemblement de civils suivant des indicateurs fallacieux^[4],
- Un mauvais choix des variables représentatives peut créer des corrélations erronées, comme le fait d'associer une menace à un certain type de bâtiment ou d'équipement^[5],
- Des objectifs mal formalisés peuvent conduire l'algorithme à optimiser des métriques techniques au détriment des impératifs éthiques ou stratégiques. Par exemple, un drone militaire qui associe le contrôle de son opérateur à une menace l'empêchant d'atteindre son objectif et qui décide de se retourner contre lui^[6].

Un faux sentiment de maîtrise

Malgré la vigilance portée à la qualité des bases de données d'entraînement des modèles d'IA, aux hypothèses implicites et aux choix des équipes de conception^{[7], [8]}, c'est souvent à l'usage que les algorithmes révèlent leurs défaillances. Le risque zéro n'existe donc pas. D'autant que les biais des utilisateurs peuvent aussi aggraver l'impact de ces défaillances. Notamment le biais d'automatisation^[9] qui fait que nous avons naturellement tendance à privilégier les suggestions d'un système automatisé, même en présence d'informations contradictoires évidentes, le biais d'action^[10] qui nous pousse à agir,

quand bien même l'inaction aboutirait logiquement à un meilleur résultat ou le biais de confirmation^[11] qui nous amène à accorder moins d'importance aux informations qui contredisent nos croyances.

L'IA est principalement utilisée pour sa capacité à accélérer le traitement de l'information et les prises de décision^[12]. Or, nos performances cognitives se dégradent très vite sous l'effet de l'instantanéité. Nous tirons des conclusions hâtives, simplistes et parfois erronées, sous l'effet d'un sentiment d'urgence. Le contrôle humain aura donc quoi qu'il arrive ses limites dans les systèmes d'IA et il ne suffira pas à garantir que les principes de proportionnalité et de distinction seront toujours respectés.

Le rôle clef des comités d'éthique

A moins que ce contrôle ne s'appuie sur un cadre qui autorise les débats contradictoires et le jugement à s'exercer pleinement et en continu. Les dilemmes éthiques sont complexes et leur résolution nécessite un temps plus long et une expertise pluridisciplinaire. Une partie de la solution consiste ainsi à créer un comité d'éthique dans chaque structure qui développe en interne des systèmes d'IA ou qui en achète à un fournisseur tiers. Ce comité doit contenir à minima un membre compétent en :

- Éthique, pour évaluer les risques et faciliter les délibérations,
- Stratégie opérationnelle, pour décider des meilleures tactiques de défense à adopter,
- Droit international ou humanitaire, pour veiller à la conformité juridique des décisions,
- Technique, pour déterminer les solutions qu'il est possible d'implémenter,
- (Neuro)psychologie, pour réduire l'impact des biais.

Réconcilier by design nos deux modes de pensée

En somme, pour reprendre l'analogie de Kahneman, il s'agit de redonner toute sa place au mode de pensée plus réfléchi du système 2, pour que celui du système 1 plus impulsif ne soit pas le seul maître à bord des systèmes d'IA.^[13]

Notre cerveau fonctionne en effet suivant deux modes de pensée distincts qui interagissent au quotidien. Le Système 1 est dominant et toujours actif en arrière-plan, pour prendre en charge la majorité des décisions simples et rapides qui s'imposent à nous, sans effort conscient. Ce n'est que lorsque nous sommes confrontés à une situation complexe que le Système 2, plus lent et coûteux en énergie, est sollicité. Mais, il reste très influençable par les associations d'idées et des schémas préétablis du Système 1 qui peuvent vite le conduire à des conclusions totalement erronées.

Conscient de ces limites intrinsèques, on attendra donc aussi du comité d'éthique qu'il soit en mesure de fiabiliser les décisions humaines dans le respect des objectifs opérationnels fixés, en suggérant l'implémentation de garde-fous cognitifs dès la

conception des systèmes d'IA. En d'autres termes, il est temps de ne plus voir les systèmes d'IA seulement comme des technologies à réguler avec une supervision humaine stricte, mais aussi comme des leviers d'amélioration de notre propre rationalité. Une co-régulation pourrait être le plus court chemin vers une IA de confiance.

Quelle régulation de l'IA de défense ?



Brunessen BERTRAND

Professeure de droit à l'Université de Rennes
Chaire Jean Monnet



La régulation de l'IA de défense peut s'appréhender comme l'identification des règles juridiques contraignantes qui viennent spécifiquement encadrer le recours à des IA- qui, en dépit de l'évolutivité de la définition retient au moins la capacité d'inférence (déduction, prédiction) -, dans les systèmes mis sur le marché ou utilisés à des fins militaires, de défense ou de sécurité nationale. L'identification est assez rapide puisqu'il n'existe aucune règle spécifique sur l'utilisation de l'IA à des fins militaires.

Le cadre juridique reste général, les normes applicables à la défense nationale (droit français et droit international humanitaire) s'appliquant plutôt aux effets (dommages collatéraux par exemple). À l'heure où les premières régulations ont été adoptées pour l'utilisation de l'IA en droit de l'Union européenne, la question peut se poser de l'opportunité de prévoir des règles spécifiques au domaine militaire.

L'exclusion de la défense nationale des règles européennes sur l'intelligence artificielle

L'article 2 de l'AI Act^[1] évoque deux exclusions. D'une part, ce texte ne s'applique pas aux systèmes d'IA si et dans la mesure où ils sont mis sur le marché avec ou sans modifications exclusivement à des fins militaires ou de défense. D'autre part, ce règlement ne s'applique pas aux systèmes d'IA qui ne sont pas mis sur le marché dans l'Union, lorsque les résultats produits par ces systèmes d'IA sont utilisés dans l'Union exclusivement à des fins de défense.

Si cette exclusion paraît claire, le juge pourrait interpréter de façon étroite cette restriction et opérer une différenciation entre les activités qui relèvent intrinsèquement des opérations militaires et celles qui leur sont dissociables, comme l'utilisation de l'IA à des fins de formation, de recrutement ou d'activités plus administratives.

Le cas particulier des biens à double usage

L'AI Act distingue trois hypothèses. Premier cas : si un système d'IA utilisé à des fins de défense est utilisé en dehors de ce cadre pour d'autres finalités (par exemple, civiles, humanitaires, répressives ou de sécurité publique), alors ce système d'IA sera soumis à l'AI Act.

L'obligation est la même dans la seconde hypothèse, qui concerne les systèmes d'IA mis sur le marché à des fins exclues (militaires, défense) et à une ou plusieurs fins non exclues (fins civiles ou répressives). Ces IA à double usage relèvent du règlement IA. Par conséquent, les fournisseurs de ces systèmes doivent veiller à appliquer les règles européennes.

A l'inverse, et c'est là le troisième cas, un système d'IA mis sur le marché à des fins civiles ou répressives qui est utilisé à des fins de défense n'est pas soumis à l'AI Act. En pratique, l'AI Act sera de fait appliqué, notamment pour les exigences qui s'appliquent en amont du déploiement de l'IA. On songe en particulier aux règles de gouvernance ou de robustesse de données. Ce ne sera pas le cas en revanche des règles qui concernent l'utilisation du système (tenir des journaux, un registre etc).

Une régulation adaptée à la diversité des usages de l'IA à des fins militaires

L'IA de défense peut être utilisée dans des finalités extrêmement hétérogènes qui n'appellent pas toutes le même niveau d'encadrement juridique ou même éthique. Une aide à la préparation opérationnelle n'implique pas le même niveau de risque qu'une IA utilisée comme aide au renseignement ou au ciblage par exemple. Deux cas particuliers présentent un niveau de risque spécifique.



D'abord, les IA qui seraient utilisées comme armes, autonomes ou non. Dans ce cas, le droit des conflits armés, tel qu'il résulte des Conventions de Genève de 1949^[2], à vocation à s'appliquer. S'agissant plus spécifiquement des systèmes d'armes autonomes, le Haut représentant de l'Union européenne pour la politique étrangère tente de promouvoir l'adoption de principes directeurs^[3]. Le Parlement européen entend défendre la nécessité d'une « position commune sur les systèmes d'armes létales autonomes qui garantisse un véritable contrôle humain sur les fonctions critiques des systèmes d'armes, y compris pendant le déploiement »^[4].

Ensuite, le soldat augmenté. Anticipant les questions qu'elle pourra soulever, le ministère des Armées a proposé une approche éthique et juridique^[5] de l'augmentation des performances du militaire^[6]. La doctrine militaire française prévoit que l'augmentation des performances du soldat devra respecter les lois de bioéthique et le droit des conflits armés. D'une part, le respect des lois de bioéthique implique le respect de la dignité de la personne humaine et de ses droits fondamentaux tel que le droit à son intégrité physique et mentale ou l'inviolabilité du corps humain (en amont de l'augmentation mais aussi pendant les phases d'augmentation et post-augmentation). Elles reposent également sur le principe du consentement libre et éclairé du patient. Le cas échéant, les dispositions code de la santé publique sur les essais cliniques impliquant la personne humaine seront applicables^[7]. D'autre part, le contrôle de licéité au droit des conflits armés (notamment les principes de nécessité, de distinction, de proportionnalité, de précaution et d'humanité) sera nécessaire pour les « moyens d'augmentation, invasifs ou non, qui peuvent être considérés comme des nouveaux moyens et méthodes de guerre »^[8].

Approche éthique ou approche juridique ?

Si le respect de principes éthiques semble nécessaire, la pertinence de la juridicisation de ceux-ci peut se poser au regard de la spécificité du domaine militaire, qui implique une agilité dans son encadrement. Au-delà, il existe un certain alignement entre la conceptualisation des principes éthiques applicables en matière militaire et les principes juridiques évoqués de l'AI Act. On trouve ainsi, parmi les exigences proposées en matière militaire^[9], des objectifs de conformité qui font très largement écho aux exigences de fiabilité de l'AI Act et de transparence qui reprennent les dispositions sur la gouvernance des données et de traçabilité, de responsabilité, ainsi qu'un principe d'autonomie humaine qui est au cœur du règlement européen. L'opportunité de juridiciser ces principes n'est pas évidente dès lors que ceux-ci sont de fait appliqués, en particulier par une conduite de la guerre guidée par des valeurs démocratiques. La réflexion française sur la gouvernance des données va même plus loin sur la nécessité de garantir une autonomie et une souveraineté des données, qui tiennent compte des impératifs d'interopérabilité. Plus généralement, il semble prématuré de vouloir réglementer un ensemble de technologies en forte évolution, sans avoir de recul sur ses usages militaires réels.

SALA - SALIA, une distinction « de confiance » ?



Ysens de FRANCE

Docteur en droit international
Chargée de mission IA Gendarmerie nationale



L'affirmation française d'un Terminator qui ne défilerait pas au 14 juillet^[1] trouva sa force morale, en 2021, dans une distinction réalisée entre les Systèmes d'Armes Létaux Autonomes (SALA) et les Systèmes d'Armes Létaux Intégrant de l'Autonomie (SALIA)^[2]. C'est une position de principe (et d'experts !) fondée sur les capacités d'une machine à définir seule sa cible, c'est-à-dire à disposer d'une faculté de choix (d'où le terme autonomie) sans toutefois en être responsable (c'est là que l'autonomie technique se distingue de celle humaine). C'est une réponse au « syndrome Terminator » cette crainte de la domination, par les armes, de la race robotique sur la race humaine, qui fut mis en avant dès 2012.

La réponse française n'est pas isolée puisqu'elle fait écho aux discussions menées sur ce sujet au sein de l'enceinte de la Convention sur certaines armes classiques (CCAC). En effet, depuis douze ans, un groupe d'experts gouvernementaux travaille sur la mise en place d'un cadre normatif et opérationnel applicable aux SALA^[3].

Au début, l'approche était très technico centrée, braquée sur le besoin de trouver des « paramètres techniques de rupture » dont l'activation engendrerait une déshumanisation complète des conflits et donc une interdiction de la technologie. Or aucune des trois faces de la technologie n'en comportent per ipse qu'il s'agisse de la première face « IA » qui permet l'adaptation d'un système à son environnement, de la deuxième face « vecteur » qui permet l'évolution du système dans un environnement et de la troisième face « létalité » qui permet la rupture de et avec son environnement^[4].

Les débats ont ainsi évolué vers une approche humano-centrée fondée sur les capacités humaines à maîtriser cette technologie tout le long de son cycle de vie. Cela a construit une dernière approche dite à double entrée ("Two tier approach") qui consiste en un renoncement aux SALA opérant en dehors de toute forme de contrôle humain (« pleinement

autonomes ») et en une mise en œuvre des mesures nécessaires afin de garantir que les systèmes d'armes létaux dotés d'autonomie dans leurs fonctions critiques (« partiellement autonomes ») soient développés et utilisés en conformité avec le droit international humanitaire^[5]. Nous retrouvons, en substance, l'approche française fondée sur la place de l'homme dans la boucle décisionnelle homme-machine comme élément classificateur. Ce qui est à la fois très ambitieux et particulièrement ardue et ce, pour deux raisons.

La première des raisons est que le contrôle humain n'étant pas une seule chose indivisible^[6], les travaux au sein de l'article 36⁷ devront être précis quant à la panoplie d'actions de contrôle spécifiques à opérer, des premiers temps de développement jusqu'aux derniers sorts jetés sur terrain. Ce n'est pas tant la technologie qu'il faudra ainsi éprouver mais l'efficacité du contrôle humain à travers des critères de performance préalablement définis. **Ce changement de paradigme exprime bien le défi sous-jacent de ce qu'on appelle la confiance en IA qui vise en réalité à opérer une réduction de l'imprévisibilité de cette technologie par une prévisibilité accrue de ceux qui vont l'opérer.**

La seconde raison est que les caractéristiques de ces nouvelles procédures de contrôle devront être idéalement partagées par chaque pays les mettant en œuvre afin de tendre vers une compréhension, voire une adoption commune de celles-ci. **Un exercice de transparence qui s'inscrira utilement dans l'écriture d'une vision commune, non moins du visage de la guerre pour demain, mais bien plus pragmatiquement, des intérêts communs à protéger pour garantir une forme de stabilité internationale.** Il y va de la confiance dans nos institutions internationales et in fine, dans l'État et sa capacité à savoir et pouvoir répondre à cette promesse technologique.





Conclusion



Karl NEUBERGER

VP Capgemini Invent
France



Olivier DENTI

Directeur Data / IA
Capgemini Invent
France



L'initiative de ce Livre Blanc sur l'IA de confiance dans la défense est née d'une conviction profonde : l'intelligence artificielle dans un domaine aussi stratégique et sensible que celui de la défense, doit être pensée, développée et déployée dans un cadre de confiance. Cette démarche ne saurait être limitée aux seuls impératifs technologiques ou opérationnels ; elle exige une prise en compte globale des enjeux éthiques, juridiques, industriels et stratégiques. La diversité des contributeurs ayant participé à ce Livre Blanc — chercheurs, universitaires, ingénieurs, industriels, juristes — reflète cette approche multidisciplinaire. Leur collaboration a permis d'apporter des perspectives variées, complémentaires voire parfois contradictoires sur les défis et opportunités que représente l'IA dans le secteur de la défense.

L'importance de l'IA de confiance dans la défense ne peut être sous-estimée. La défense, en tant que domaine critique, doit être exemplaire dans le développement et l'utilisation de systèmes d'IA. Une IA militaire non fiable ou biaisée pourrait engendrer des conséquences graves sur le terrain, allant de décisions erronées à des violations du droit international humanitaire.

Parmi les pistes évoquées dans ce Livre Blanc, plusieurs recommandations concrètes émergent pour renforcer la confiance en l'IA dans la défense comme :

- La formation et la sensibilisation : la maîtrise des technologies IA par les forces armées nécessite un effort massif en formation. Cela inclut non seulement les aspects techniques mais aussi les dimensions éthiques et opérationnelles.
- L'investissement dans les infrastructures critiques : la mise en place de supercalculateurs dédiés à l'IA militaire, comme celui du Mont-Valérien, illustre la volonté française de garantir une autonomie stratégique. Ces infrastructures doivent être accompagnées par un effort accru dans le développement de solutions résilientes capables de fonctionner dans des environnements dégradés.
- La réflexion sur un cadre normatif équilibré : cette démarche s'inscrit dans un équilibre délicat entre innovation technologique et valeurs

fondamentales. Ce cadre devra tenir compte des réalités du terrain militaire moderne, où certaines décisions doivent être prises à une vitesse dépassant les capacités humaines, tout en maintenant un équilibre avec nos principes éthiques fondamentaux.

- Une coopération européenne renforcée : bien que la souveraineté implique une certaine indépendance technologique, elle ne doit pas conduire à un isolement. La France doit privilégier des partenariats stratégiques pour mutualiser les coûts de R&D sur des technologies non critiques et collaborer sur des enjeux communs comme la cybersécurité et le partage d'informations. En adoptant un équilibre entre autonomie nationale, maîtrise industrielle et coopération internationale, elle préservera sa souveraineté technologique et renforcera son leadership dans l'innovation de défense du XXI^e siècle.

Face aux nouveaux rapports de force internationaux, la France affirme son leadership en matière d'innovation responsable dans la défense. Son engagement historique pour une IA éthique lui confère une légitimité unique pour promouvoir un usage maîtrisé et aligné sur les valeurs démocratiques.

Enfin, il est essentiel de reconnaître que cette révolution technologique s'accompagne de nouveaux rapports de force internationaux. La maîtrise de l'IA devient un facteur déterminant pour garantir non seulement la supériorité opérationnelle mais aussi l'autonomie stratégique des nations. En conciliant innovation technologique et principes éthiques, la France peut non seulement répondre aux défis actuels mais également poser les bases d'une IA militaire qui inspire confiance, tant au niveau national qu'international.

Ce Livre Blanc ne prétend pas apporter toutes les réponses aux défis complexes posés par l'intégration croissante de l'IA dans la défense. Il constitue cependant une base solide pour engager une réflexion collective et structurée. L'objectif est clair : faire de l'IA un levier stratégique au service de la sécurité nationale tout en préservant son acceptabilité sociale et sa conformité aux valeurs défendues par la France.

Notes / Références / Annexes



L'IA dans les cyberattaques : un multiplicateur de risques (p.10)

[1] - <https://cdn.openai.com/threat-intelligence-reports/influence-and-cyber-operations-an-update-October-2024.pdf>

[2] - https://www.defense.gouv.fr/sites/default/files/ministere-armees/20210429_Comite_d'ethique_de_la_defense_-_Avis_integrer_autonomie_systemes_armes_letaux.pdf

[3] - <https://www.state.gov/wp-content/uploads/2023/10/Latest-Version-Political-Declaration-on-Responsible-Military-Use-of-AI-and-Autonomy.pdf>

[4] - <https://www.elysee.fr/emmanuel-macron/2025/02/11/paris-declaration-on-maintaining-human-control-in-ai-enabled-weapon-systems>

La souveraineté de l'IA pour la Défense : un impératif stratégique pour une IA de confiance (p.12)

[1] - Intelligence Artificielle | Comment la Chine et les États-Unis entendent asseoir leur leadership et souveraineté ? - Forbes France (<https://www.forbes.fr/technologie/intelligence-artificielle-comment-la-chine-et-les-etats-unis-entendent-asseoir-leur-leadership-et-souverainete/>)

[2] - <https://www.defense.gouv.fr/sites/default/files/ministere-armees/esprit-defense-numero-13-automne-2024.pdf>

Les enjeux et perspectives de l'IA de défense à l'échelle (p.14)

[1] - Une structure centrale (hub) gère la stratégie et les grandes priorités, les ressources communes (incluant la plateforme technologique) et la gouvernance des données, tandis que des équipes locales ou BU (spokes) développent des projets spécifiques. Ce modèle permet de sécuriser les ressources critiques tout en offrant de la flexibilité aux unités opérationnelles.

[2] - ITAR (International Traffic in Arms Regulations) réglemente la fabrication, la vente et la distribution de matériel, de données et de documentation liés à l'armée, à la défense et à l'espace.

Coopération et confiance dans les projets complexes (p.16)

[1] - <https://www.nist.gov/itl/ai-risk-management-framework>

L'enjeu n'est pas l'automatisation de la guerre mais la distanciation des combattants (p.18)

[1] - « Drones papers », The Intercept, 2015, URL : <https://theintercept.com/drone-papers/the-assassination-complex/>.

[2] - <https://www.youtube.com/watch?v=9CO6M2HsolA&t=0s>.

[3] - François Lecointre, interview du 9 janvier 2025 au Figaro, URL : <https://video.lefigaro.fr/figaro/video/les-soldats-sont-des-sur-citoyens-qui-sengagent-pour-servir-une-cause-qui-les-depassera-celle-de-la-france-francois-lecointre-est-lininvite-de-libre-a-vous/>

[4] - « Drone wars: the gamers recruited to kill », Guardian Docs, URL: <https://www.youtube.com/watch?v=bGA8RFB0VSw>

L'IA de Confiance dans la Défense : Opportunités et Défis (p.20)

[1] - Ministère des Armées article « IA de Défense : le défi de la souveraineté », Article extrait Esprit Défense n°13 p.30 dossier « Comment l'IA transforme le champ de bataille » (magazine téléchargeable : <https://www.defense.gouv.fr/actualites/esprit-defense-ndeg-13-comment-lia-transforme-champ-bataille/>)

[2] - <https://www.defense.gouv.fr/actualites/comite-dethique-defense-usage-lia-ethique-proportionne>

[3] - <https://www.defense.gouv.fr/actualites/comite-dethique-defense-usage-lia-ethique-proportionne>

[4] - <https://www.rfi.fr/fr/podcasts/lignes-de-d%C3%A9fense/20250209-intelligence-artificielle-de-d%C3%A9fense-le-d%C3%A9fi-de-la-masse-et-de-la-souverainet%C3%A9>

[5] - <https://www.defense.gouv.fr/actualites/amiad-bilan-6-mois-son-lancement>

[6] - <https://www.defense.gouv.fr/actualites/bertrand-rondepierre-notre-souverainete-matiere-dia-implique-maitrise-technologies>

[7] - <https://www.ljeunesolution.gouv.fr/articles/formation-intelligence-artificielle-ecoles-france-2025>

Industrialisation de l'IA, un pilier de la confiance (p.24)

- [1] - <https://inside-machinelearning.com/recall-precision-f1-score/>
- [2] - https://fr.wikipedia.org/wiki/M%C3%A9moire_vive
- [3] - https://fr.wikipedia.org/wiki/M%C3%A9moire_vid%C3%A9o

Construire la confiance en l'IA : la nécessité de la certification. Le point de vue d'un CESTI (p.22)

- [1] - Voir https://fr.wikipedia.org/wiki/Alignement_des_intelligences_artificielles.
- [2] - Voir le blog : <https://aleph.se/andart2/ethics/ethics-for-neural-networks/>
- [3] - Exemple : [1802.01933] A Survey Of Methods For Explaining Black Box Models - <https://arxiv.org/abs/1802.01933>

Transformations de la guerre : les enjeux de la confiance dans les applications militaires de l'IA (p.28)

[1] - Un systémier-intégrateur est un maître d'œuvre industriel qui conçoit, développe, produit et intègre des systèmes complexes ou « systèmes de systèmes » afin de répondre aux besoins capacitaires des forces armées.

Vers la sécurisation et la frugalité des fonctions IA embarquées (p.30)

- [1] - www.skyld.io
- [2] - H. Khaleghi, S. Paquelet, Unleashing Timing Advance Precision: CRLB Limit and Neural Implementation, WiMob 2024.
- [3] - J-P. Kahane, Sur le théorème de superposition de Kolmogorov, Journal of Approximation Theory, vol. 13, issue 3, pp. 229-234, mars 1975.
- [4] - Z. Liu et al., KAN: Kolmogorov-Arnold Networks, arXiv:2404.1975, juin 2024.
- [5] - broadcastprome.com/...synaptics-vs680, septembre 2024.

Face à une IA source de crises, la confiance comme solution ? (p.32)

- [1] - <https://www.sciencedirect.com/science/article/pii/S2215016122001273>
- [2] - https://huggingface.co/learn/cookbook/fr/llm_judge
- [3] - https://www.wto.org/library/events/event_resources/sps_1411202310/329_1130.pdf
- [4] - <https://www.riskinsight-wavestone.com/2023/10/quand-les-mots-deviennent-des-armes-prompt-injection-et-intelligence-artificielle/>
- [5] - <https://www.cnil.fr/fr/definition/red-teaming>
- [6] - <https://www.cyber-management-school.com/ecole/les-fondamentaux-de-la-cybersecurite/que-signifie-la-blue-team-en-cybersecurite/>

L'IA de confiance est une illusion : construisons des produits de confiance

- [1] - <https://www.ibm.com/fr-fr/think/topics/embedding>
- [2] - Shrivu Shankar (2024). How to Backdoor Large Language Models. [online] Available at: <https://blog.sshh.io/p/how-to-backdoor-large-language-models>
- [3] - Royal United Services Institute (2023). Tactical Lessons from IDF Operations in Gaza. Available at: <https://static.rusi.org/tactical-lessons-from-idf-gaza-2023.pdf>
- [4] - John, D. (2024). Why Boeing's problems with 737 MAX began more than 25 years ago. [online] Harvard Business School Working Knowledge. Available at: <https://www.library.hbs.edu/working-knowledge/why-boeings-problems-with-737-max-began-more-than-25-years-ago>

L'IA de confiance pour la défense : de la conception à l'évaluation (p.40)

[1] - <https://www.iso.org/fr/standard/81230.html>

[2] - <https://www.iso.org/fr/standard/77304.html>

[3] - Maintien en Condition Opérationnelle

Vers des IA vérifiables (p.42)

[1] - <https://arxiv.org/abs/1706.10268>

[2] - <https://dl.acm.org/doi/abs/10.1109/TPDS.2021.3068195>

[3] - <https://www.cnil.fr/fr/definition/chiffrement-homomorphe>

[4] - <https://www.ibm.com/fr-fr/think/topics/gradient-descent>

L'influence de l'IA sur la cybersécurité : défis et opportunités du point de vue de la recherche (p.44)

[1] - Les attaques typiques modifient les modèles d'IA via l'injection de fausses données d'entraînement ou l'inversion des étiquettes ou des algorithmes d'inférence. Grâce à la complexité de l'IA, cela peut avoir des effets importants qui sont difficiles à détecter.

[2] - La recherche de la chaire cybersécurité des infrastructures critiques (Cyber CNI), <https://cyberCNI.fr>, accédé 17/02/2025

[3] - Le spear-phishing est une forme très ciblée d'attaque par phishing où les cybercriminels adaptent leurs messages frauduleux à une personne ou entité spécifique.

[4] - JudgeGPT - une application web pour explorer le potentiel de génération de désinformation de différents outils d'IA générative, <https://judgegpt.streamlit.app/>, accédé 17/02/2025

Recommandations de sécurité pour la mise-en œuvre de plateformes d'IA générative dédiées à l'ingénierie des systèmes critiques (p.46)

[1] - ENISA, Multilayer Framework for good cybersecurity practices for AI (<https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai>)

[2] - ANSSI Guide de recommandations de sécurité pour un système d'IA générative ([https://cyber.gouv.fr/publications/Recommandations de sécurité pour un système d'IA générative](https://cyber.gouv.fr/publications/Recommandations_de_sécurité_pour_un_système_d'IA_générative))

[3] - MITRE ATLAS Matrix (<https://atlas.mitre.org/matrices/ATLAS>)

[4] - NIST AI 600-1 Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence profile (<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf>)

[5] - OWASP LLM AI Cybersecurity & Governance Checklist (https://owasp.org/www-project-top-10-for-large-language-model-applications/llm-top-10-governance-doc/LLM_AI_Security_and_Governance_Checklist-v1.pdf)

[6] - CWE AI Working Group (<https://github.com/CWE-CAPEC/AI-Working-Group/tree/main>)

IA raisonnée pour les systèmes décisionnels critiques (p.48)

[1] - Zalila, Z., Intellitech & Xtractis (2024) XTRACTIS® the General Reasoning AI for Trusted Decisions. Use Case #27 | Human Resources : Discovery of Discriminatory Biases in the Professional Evaluation of Employees – Benchmark vs. Logistic Regression, Random Forest, Boosted Tree & Neural Network. INTELLITECH [intelligent technologies], October, v1.0, Compiègne, France, 6p. <https://xtractis.ai/use-cases/>

[2] - <https://github.com/google/magika?tab=readme-ov-file>

[3] - Zalila, Z. (2019-2024) Livre Blanc | Spécificités et Avantages de l'IA Cognitive Floue Augmentée XTRACTIS® : IA Raisonnée Générale, Robuste, Intelligible & Auditable. INTELLITECH [intelligent technologies], Octobre 2024, v6.6, Compiègne, France, 25 p, <https://xtractis.ai/white-papers/>

De la sécurité du Machine Learning (p.50)

[1] - <https://www.cnil.fr/fr/definition/calcul-multipartite-secureise>

[2] - <https://www.cnil.fr/fr/definition/chiffrement-homomorphe>

[3] - <https://www.cnil.fr/fr/definition/confidentialite-differentielle>

Intelligence Artificielle responsable ou de confiance pour la Défense : quelle nouvelle ingénierie ? (p.52)

[1] - [HLEG. Ethics guidelines for trustworthy AI". Avril 2019.](#)

[2] - https://www.nato.int/cps/en/natohq/official_texts_187617.htm

[3] - [RISC-V International](#)

[4] - <https://www.confiance.ai/>

[5] - [EUROCAE WG 114](#)

[6] - [Cycle en V](#)

Les principales vulnérabilités des LLMs face aux attaques malveillantes (p.56)

[1] - <https://www.cnil.fr/fr/definition/execution-en-bac-sable-ou-sandboxing>

[2] - <https://www.ibm.com/fr-fr/topics/ai-hallucinations>

La cybersécurité de l'IA au cœur du champ de bataille (p.58)

[1] - Étape où le modèle applique les connaissances acquises lors de l'entraînement à de nouvelles informations pour prendre des décisions ou faire des prédictions.

[2] - Les recommandations listées dans cet article ne sont pas exhaustives mais visent à proposer un premier niveau de réponse.

[3] - Il est possible de retrouver une liste de travaux portant sur ce thème ici : <https://srg-research.github.io/mlsec/modelExtDef>

[4] - Les attaques par canaux auxiliaires sur un SIA exploitent des informations indirectes ou "fuites" issues du système, comme la consommation d'énergie, les variations de temps d'exécution ou les émissions électromagnétiques, pour analyser le fonctionnement du SIA et déduire des informations sur son fonctionnement.

[5] - Les travaux du projet PICTURE, piloté par le CEA sont à ce titre éclairants : <https://picture-anr.cea.fr/presentation>

[6] - Les attaques en déni de service sur SIA visent à rendre un SIA indisponible en saturant ses ressources avec un flux massif de requêtes.

[7] - Il est possible d'avoir une vision d'ensemble plus large de ces problématiques dans le guide de l'ANSSI sur l'IA générative : <https://cyber.gouv.fr/publications/recommandations-de-securite-pour-un-systeme-dia-generative>

Systèmes d'Armes Létales Autonomes – aspects juridiques (p.62)

[1] - <https://legrandcontinent.eu/fr/2025/01/07/lia-aux-racines-de-la-guerre/>

[2] - <https://disarmament.unoda.org/fr/le-desarmement-a-geneve/convention-sur-certaines-armes-classiques/>

[3] - https://www.europarl.europa.eu/doceo/document/TA-9-2021-0009_FR.pdf

[4] - https://www.assemblee-nationale.fr/dyn/15/rapports/cion_def/115b3248_rapport-information.pdf_p65

[5] - <https://www.defense.gouv.fr/dems/systemes-darmes-letales-autonomes-sala>

[6] - <https://www.diplomatie.gouv.fr/fr/politique-etrangere-de-la-france/securite-desarmement-et-non-proliferation/desarmement-et-non-proliferation/systemes-d-armes-letales-autonomes-quelle-est-l-action-de-la-france/>

[7] - https://www.defense.gouv.fr/sites/default/files/ministere-armees/20210429_Comit%C3%A9%20d%27%C3%A9thique%20de%20la%20d%C3%A9fense%20-%20Avis%20int%C3%A9gration%20autonomie%20syst%C3%A8mes%20armes%20l%C3%A9taux.pdf

[8] - <https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf>

[9] - <https://www.usine-digitale.fr/article/comment-les-start-up-d-ia-de-defense-veulent-conquerir-le-champ-de-bataille.N2207171>

Trust by Design : comment être éthique et un leader européen de l'IA (p.64)

[1] - <https://www.tortoisemedia.com/intelligence/global-ai>

[2] - <https://www.info.gouv.fr/actualite/25-recommandations-pour-lia-en-france>

Comment garantir qu'une IA respectera les principes de proportionnalité et de distinction (p.66)

[1] - https://www.defense.gouv.fr/sites/default/files/ministere-armees/20210429_Comit%C3%A9%20d%27%C3%A9thique%20de%20la%20d%C3%A9fense%20-%20Avis%20int%C3%A9gration%20autonomie%20syst%C3%A8mes%20armes%20I%C3%A9taux.pdf

[2] - Weizenbaum J. (1976), Computer Power and Human Reason: From Judgment to Calculation, San Francisco, W. H. Freeman.

[3] - <https://europepmc.org/backend/ptpmcrender.fcgi?accid=PMC2545288&blobtype=pdf>

[4] - <https://www.france24.com/fr/afrique/20231205-aunig%C3%A9ria-un-drone-de-l-arm%C3%A9e-tue-par-erreur-85-civils-lors-d-une-f%C3%A9te>

[5] - <https://www.rfi.fr/fr/afrique/20240325-mali-une-dizaine-d-enfants-tu%C3%A9s-par-une-frappe-de-drones-de-l-arm%C3%A9e>

[6] - <https://amp.theguardian.com/us-news/2023/jun/01/us-military-drone-ai-killed-operator-simulated-test>

[7] - <https://standards.ieee.org/ieee/7003/11357/>

[8] - <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf>

[9] - Skitka L. , Mossier K., Burdick M. (1999), Does automation bias decision-making?, International Journal of Human-Computer Studies, 51(5):991-1006.

[10] - Patt A., Zeckhauser R. (2000), Action bias and environmental decisions, Journal of Risk and Uncertainty, 21(1):45-72.

[11] - Wason P. C. (1960), On the Failure to Eliminate Hypotheses in a Conceptual Task, Quarterly Journal of Experimental Psychology, 12(3):129-140.

[12] - <https://www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets>

[13] - Kahneman D. (2012), Système 1 / Système 2 : Les deux vitesses de la pensée, Flammarion.

Quelle régulation de l'IA de défense ? (p.68)

[1] - Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle (règlement sur l'intelligence artificielle), JO L, 2024/1689, 12.7.2024.

[2] - Article 36 du protocole additionnel I aux Conventions de Genève de 1949 du 8 juin 1977.

[3] - United Nations' Group of Governmental Experts on Lethal Autonomous Weapons Systems - "Possible Guiding Principles", 2018.

[4] - Résolution du Parlement européen du 12 septembre 2018 sur les systèmes d'armes autonomes (2018/2752(RSP)).

[5] - Ministère des Armées, Centre interarmées de concepts, de doctrines et d'expérimentations, Concept exploratoire - le soldat augmenté (2022), CEIA-3.0.3_SOLD-AUGM(2022), N° 92/ARM/CICDE/NP du 30 juin 2022.

[6] - V. aussi en ce sens, l'Avis du comité d'éthique de la défense du 18 septembre 2020 portant sur le soldat augmenté.

[7] - Articles L1121-1 et suivants du code de la santé publique.

[8] - Concept exploratoire - le soldat augmenté, op. cit. p. 56.

[9] - Comité d'éthique de la défense, Avis sur l'usage des technologies d'intelligence artificielle par les forces armées, 14 janvier 2025.

SALA - SALIA, une distinction « de confiance » ? (p.70)

[1] - Florence Parly, discours du 5 avril 2019.

[2] - Comité d'éthique de la défense, Avis sur l'intégration de l'autonomie dans les systèmes d'armes létaux, 29

avril 2021, p.4. « Il existe une différence de nature entre les systèmes d'armes létaux autonomes, les SALA, et les systèmes d'armes létaux intégrant de l'autonomie mais demeurant sous maîtrise de l'humain, ci-après dénommés SALIA ».

[3] - Ministère de l'Europe et des Affaires étrangères, « Systèmes d'armes létales autonomes, quelle est l'action de la France ? », Février 2020.

[4] - Comité d'éthique de la défense, Avis sur l'usage des technologies d'intelligence artificielle par les forces armées, 14 janvier 2025, p.16.

[5] - Assemblée nationale, 17ème législature, Position de la France concernant les systèmes d'armes autonomes, question écrite n°1610, question écrite n°1610 publiée au JO le 5 novembre 2024, Réponse publiée au JO le 10 décembre 2024.

[6] - Winston MAXWELL, Le contrôle humain des systèmes algorithmiques – un regard critique sur l'exigence d'un « humain dans la boucle ». Droit, Université Paris 1 Panthéon-Sorbonne, 2022. Tel-04010389, p. 83. 7 Article 36 du Protocole additionnel I aux Conventions de Genève du 12 août 1949 relatif à la protection des victimes des conflits armés internationaux du 8 juin 1977 (PA I) : « dans l'étude, la mise au point, l'acquisition ou l'adoption d'une nouvelle arme, de nouveaux moyens ou d'une nouvelle méthode de guerre, une Haute Par/e contractante a l'obligation de déterminer si l'emploi en serait interdit, dans certaines circonstances ou en toutes circonstances, par les dispositions du présent Protocole ou par toute autre règle du droit international applicable à cette Haute Partie contractante ».

PÔLE D'EXCELLENCE
CYBER

www.pole-excellence-cyber.org

