Intelligence Artificielle
des systèmes critiques

THALES
Building a future we can all trust

THALES
Building a future we can all trust

Intelligence Artificielle des systèmes critiques

Juliette Mattioli et Christophe Meyer
Novembre 2024

Résumé

La conviction de Thales est que l'IA, sous réserve d'être parfaitement maîtrisée, permet à l'humain de prendre des décisions à la fois meilleures et plus rapides.

Les géants de l'Internet et les grands centres de recherche internationaux ont une contribution majeure au développement de l'IA pour le monde civil. L'application de l'IA au monde de la Défense et plus généralement aux systèmes critiques implique la prise en compte de contraintes fortes : la complexité des systèmes concernés (capteurs, effecteurs…); l'embarquabilité dans des composants limités en taille, poids et énergie; la connectivité entre les sous-systèmes qui peut être intermittente, à faible bande pas- sante voire contestée (brouillage).

Par ailleurs, il est impératif pour des systèmes critiques de définir à tout moment, en fonction des circonstances et de l'environnement, quel degré d'autonomie (en cours de mission), quelles capacités d'adaptation (en général entre chaque mission), quels moyens d'échanges avec les opérateurs humains (dialogue homme-machine) sont acceptables et nécessaires pour les IA déployées.

Enfin, si les technologies d'IA en apprentissage permettent d'atteindre des niveaux de performances inatteignables par les algorithmes classiques et rendent possible de nouvelles capacités, elles s'accompagnent de vulnérabilités qui leur sont propres. Cela a conduit Thales à compléter – pour ces besoins spécifiques à l'IA – son expertise et ses moyens en cybersécurité.

Au sein de Thales, la conception et la qualification des systèmes exploitant de l'IA prennent en compte les contraintes de sécurité et de sûreté adaptées à leurs niveaux de criticité, d'autonomie et de capa- cités d'adaptation. Cela permet de tirer profit des technologies les plus modernes tout en garantissant leur parfaite maîtrise.

1. Introduction

Le 28 mars 2024, Thales présente son dispositif d'accélération de l'intégration de l'intelligence artificielle (IA) appliquée aux systèmes critiques, en particulier dans le secteur de la défense tout en capitalisant sur ses atouts :

- Expertise technique et connaissance opérationnelle, maîtrise des architectures, gestion des environnements contraints (embarquabilité, frugalité), sécurisation et fiabilisation.
- Maîtrise de l'ensemble des technologies clés de l'IA (hybride, embarquée, frugale, autonome, distribuée, explicable, interactive, générative) et de leurs applications.

Cette accélération se fait de manière maîtrisée, pour répondre aux enjeux de fiabilité, de transparence, de sécurité et d'éthique, conformé- ment à la stratégie Thales d'IA de confiance pour les systèmes critiques, tout en s'inscrivant dans le cadre réglementaire européen de l'AI Act.

Au sein de Thales, l'IA fédère 600 spécialistes IA, œuvrant de la RT&I (recherche, technologie et innovation) à l'implémentation, dans les senseurs et les systèmes, répartis au sein du triptyque « Sensors »/« Factory »/« Labs » :

- **cortAIx Sensors** permet d'accélérer le développement de cas d'usage IA utilisant les données senseurs. Les développements IA liés aux senseurs resteront hébergés dans les entités opérationnelles, qui en maîtrisent la conception et notamment les composantes IA.
- **cortAIx Factory** industrialise des cas d'usage IA utilisant des don- nées systèmes et ainsi que des méthodes et outils d'ingénierie de l'IA. Cette entité va, dans un premier temps, s'adresser aux clients défense, en France comme à l'étranger. Les thématiques clés de cortAIx Factory sont : C4I (Command, Control, Communications, Computers, Intelligence), autonomie, services de soutien et combat collaboratif.
- **cortAIx Labs** fédère les ressources RT&I du Groupe en IA et valorise les synergies, afin de mettre à disposition des unités opérationnelles et de cortAIx Factory des briques technologiques et des démonstrateurs d'applications, ainsi que des outils et méthodes d'ingénierie de l'IA de confiance.

2. Les systèmes critiques

2.1 LA DÉFINITION DES SYSTÈMES CRITIQUES

Un système critique est un système dont une défaillance peut entraîner des conséquences graves telles qu'un impact négatif sévère sur l'environnement, des dommages matériels significatifs, des blessures graves voire des pertes en vies humaines. Ils sont généralement associés à des domaines pour lesquels la sécurité est primordiale, comme l'automation totale ou partielle dans le transport (aérien, maritime, terrestre – les véhicules autonomes…), les dispositifs médicaux et l'industrie pharmaceutique, la production d'énergie (particulièrement nucléaire), les systèmes de paiement bancaires et l'industrie de sécurité et de défense.

La criticité d'un système est évaluée en fonction de l'impact potentiel de ses éventuels dysfonctionnements. Les systèmes critiques sont donc conçus pour être hautement fiables et sécurisés, répondant à des normes précises et sont assujettis à des certifications rigoureuses. L'introduction des technologies d'IA dans les systèmes critiques enjoint de repenser ou compléter les normes et les certifications. C'est dans cet esprit qu'a été édicté, au niveau européen, l'AI Act visant à encadrer le développement de l'IA pour en réduire les risques (pas uniquement en termes de sécurité, de fiabilité et de sûreté mais également de protection des données privées, de transparence et d'éthique).

2.2 LES ENJEUX/CONTRAINTES

Thales, en tant que systémier, appréhende les IA qu'il développe ou intègre comme un complément de l'intelligence humaine (l'expertise métier) au service de systèmes critiques intelligents. Certaines contraintes spécifiques à ces systèmes critiques doivent être prises en compte dès la conception : contraintes d'embarquabilité (taille, poids, énergie consommée…), de connectivité (intermittente, à faible bande passante voire contestée – brouillage…), contraintes d'exploitation (complexité masquée, usage intuitif…) ainsi que des contraintes de sûreté et de (cy- ber) sécurité. La conviction de Thales est que l'IA, sous réserve d'être parfaitement maîtrisée, permet à l'humain de prendre des décisions à la fois meilleures et plus rapides et de réduire sa charge mentale en déléguant certaines fonctions non critiques, voire, partiellement, des fonctions critiques dans certaines conditions contrôlées. La forte dynamicité des contraintes et des ressources disponibles qui varient en continu dans les contextes de déploiement de la plupart des systèmes critiques (énergie, puissance de calcul et mémoire disponibles, temps imposé pour obtenir une réponse, connectivité entre les différents composants du système…) enjoignent à Thales de concevoir des IA « réactives » et « polymorphes ».
« Réactives » parce qu'en fonction du temps alloué, elles doivent être en mesure de produire un résultat de bonne qualité mais éventuellement non optimal. « Polymorphe » parce qu'en fonction des ressources disponibles, la même fonction doit pouvoir être remplie, éventuellement de manière dégradée, par des traitements adaptés.

Par exemple, Thales conçoit des caméras capables de :

- produire différents niveaux d'analyse de la scène observée en fonction des ressources dont elles disposent en local grâce à la sélection intelligente des algorithmes dont elles sont équipées (IA ou non).
- transmettre, en plus de leur analyse, des données qu'elles compressent plus ou moins (courts textes, images, flux vidéo HD…) en fonction de la bande passante disponible vers d'autres composants du systèmes qui pourront compléter l'analyse.

Dans le contexte des systèmes d'aide à la décision basés sur de l'IA, Thales conçoit des solutions de recommandations en temps contraint pouvant être affinées ou enrichies de façon continue.

3. Les différents paradigmes d'IA

L'expression « Intelligence Artificielle » fut adoptée au Congrès de Dartmouth en 1956 pour désigner ce nouveau domaine de recherche. Sa définition la plus simple est donnée par la mission Villani (2018) : « Une intelligence artificielle est un programme informatique visant à effectuer, au moins aussi bien que des humains, des tâches nécessitant un certain niveau d'intelligence ».

3.1 CONNEXIONNISTE, SYMBOLIQUE, HYBRIDE

Ces dernières années, l'IA connexionniste via les techniques d'apprentissage (Machine Learning) est à l'honneur, reléguant l'IA symbolique au second plan. Dans ce cadre, l'IA connexionniste et statistique est construite sur un paradigme biologique qui s'inspire du modèle du cerveau humain, cherchant à inférer un modèle conceptuel à partir d'exemples. Cette approche est bien adaptée à la perception mais peu à la résolution de problèmes complexes. L'apprentissage profond (Deep Learning) et les IA génératives (LLM - Large Language Model - ou LVM - Large Vision Model) relèvent de cette catégorie. L'IA symbolique utilise le raisonnement formel et la logique ; c'est une approche cartésienne de l'intelligence, où les connaissances sont encodées à partir d'axiomes desquels on déduit des conséquences.

La différence entre ces deux principaux paradigmes réside dans le fait que, dans un système à base d'IA symbolique, les connaissances sont explicitement définies par des experts, alors que dans une approche connexionniste et statistique, les connaissances sont déduites automatiquement à partir des données. David Sadek, VP Recherche Technologies et Innovation de Thales pour l'IA, explique que « l'IA connexionniste était jusqu'à récemment l'IA des sens, et l'IA symbolique celle du sens ».

Ainsi, dans de nombreux contextes opérationnels, comme celui des systèmes critiques évoluant dans un contexte dynamique et incertain, il convient de prendre en compte d'autres types d'informations que celles portées par les données brutes issues des capteurs, comme celles déjà encapsulées dans des modèles physiques (outils de simulation, équations aux dérivées partielles...) ou des connaissances métiers capturées par des ontologies, des règles logiques, des modèles sémantiques... Plus qu'une simple combinaison de ces paradigmes, le concept d'IA hybride (§6.1.1) peut-être défini dans son sens le plus large, comme incluant toute méthode qui intègre deux approches distinctes y compris une IA enrichie par les mathématiques ou la physique.

Typiquement, l'apprentissage par renforcement (RL : Reinforcement Learning) qui consiste à optimiser des stratégies en se basant sur une approche essai-erreur avec un éventuel feedback humain et des outils de simulation spécifiques, est aussi un exemple d'IA hybride.

De la même façon, des systèmes utilisant des LLM (Large Language Models) ou LVM (Large Vision Models) couplés avec des systèmes de recherche d'information et éventuellement des agents artificiels auto- nomes, tels que ceux basés sur des architectures RAG (Retrieval-Augmented Generation) ou ReACT (Reasoning and Acting - §6.1.5) sont des systèmes qu'on peut qualifier d'hybrides.

4. Autonomie et Adaptabilité

Afin d'établir les conditions suivant lesquelles une IA peut élaborer une décision (ou une stratégie) et surtout si elle la suggère à un décisionnaire humain (aide ou support à la décision) ou la met elle-même directe- ment en œuvre, il faut qualifier son degré d'autonomie et sa capacité d'adaptation et ce, en fonction des circonstances et de façon dynamique. (Fig.1)

Par exemple, pour qualifier le degré d'autonomie d'un véhicule, c'est-à-dire sa capacité à mettre en œuvre des décisions liées à son analyse de l'environnement (freinage d'urgence, maintien d'une distance de

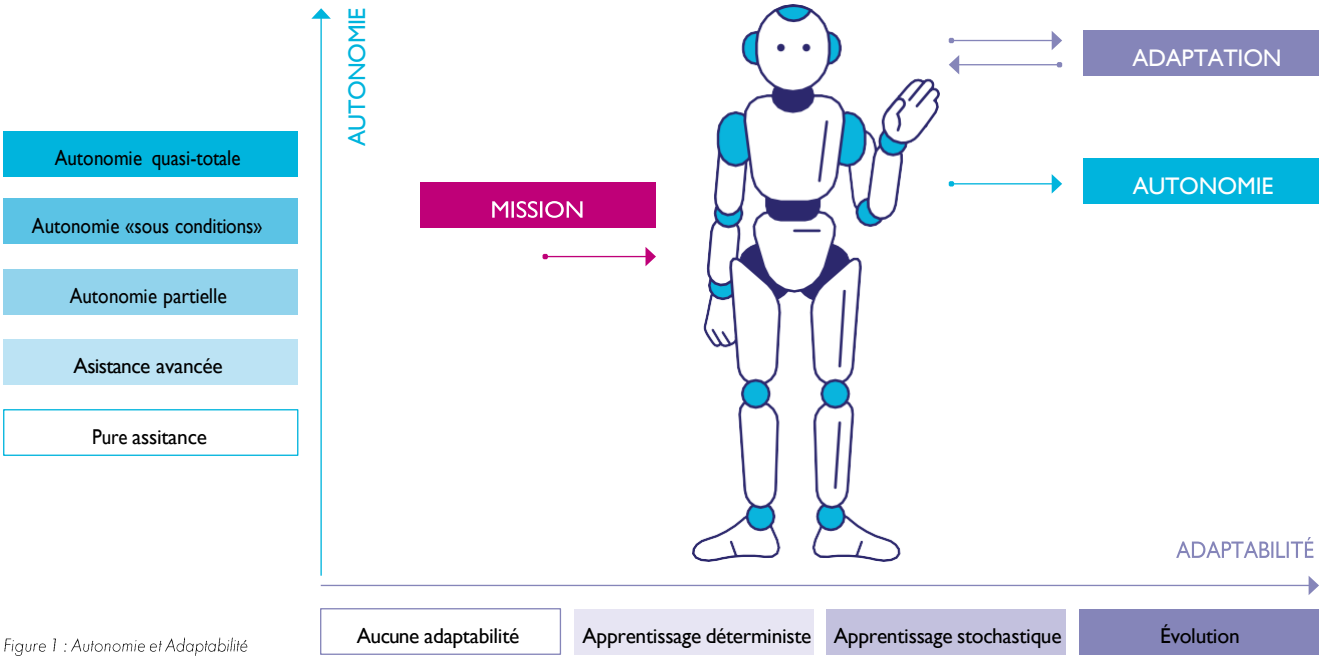

Figure 1 : Autonomie et Adaptabilité

sécurité avec les autres véhicules, changement de voie...), l'industrie automobile a convergé sur une échelle liée à la classification NHTSA (National Highway Traffic Safety Administration) qui comporte cinq ni- veaux de 0 à 4 :

- Niveau 0 : Pure assistance, le véhicule peut éventuellement avertir le conducteur de certaines conditions ou dangers mais ne modifie en rien la conduite.
- Niveau 1 : Assistance avancée, le véhicule peut effectuer certaines actions « simples » comme réguler la vitesse mais toute action contraire du conducteur prend le pas. Il n'effectue pas d'actions plus « complexes » (tel qu'un changement de voie).
- Niveau 2 : Autonomie partielle, à la demande explicite du conducteur, le véhicule peut effectuer certaines manœuvres (comme le stationnement). Toutefois, le conducteur doit rester en capacité permanente de reprendre la main.

Les niveaux 0 à 2 d'autonomie de l'industrie automobile ne peuvent pas rendre l'IA responsable d'éventuels incidents, il s'agit simplement d'aides à la conduite.

- Niveau 3 : Autonomie « sous conditions », à la demande explicite du conducteur, et sous certaines conditions que le conducteur ne peut pas changer (par exemple en maintenant la vitesse sous un certain seuil), le véhicule peut effectuer la plupart des manœuvres.
- Niveau 4 : Autonomie quasi-totale. Le véhicule peut effectuer toutes les manœuvres sans requérir la supervision du conducteur.

Notons que dans tous les cas, le conducteur est libre d'enclencher ou pas les mécanismes d'autonomie de son véhicule. À ce titre, il a le contrôle de son degré d'autonomie. La même échelle se transpose à l'ensemble des systèmes critiques et Thales s'y réfère donc.

Le monde moderne est dynamique voire, parfois, chaotique. Une solution pertinente établie à un certain moment peut s'avérer inadaptée lorsque l'environnement change, voire dangereuse dans le cas des systèmes critiques. Les solutions doivent donc évoluer au cours du temps et, bien évidemment, les capacités d'apprentissage des IA qui en sont pourvues leur permet d'adapter leurs réponses, leurs comportements au gré de l'évolution du contexte dans lequel elles sont déployées. Encore faut-il maîtriser ces adaptations.

Thales propose une échelle d'« adaptabilité » (faculté de s'adapter) des IA déployées (cette échelle ne concerne pas les phases de conception ou d'entraînement des IA mais les phases d'exploitation) qui comporte quatre niveaux de 0 à 3 :

- Niveau 0 - Aucune adaptabilité : les sorties de l'IA sont calculées de façon déterministe ou stochastique en fonction des entrées mais sans variation dans le temps (pour les systèmes stochastiques ou reposant sur de la logique floue, cela signifie que les répartitions probabilistes ne varient pas). Aucune capacité d'adaptation en « opération ». Même si l'IA a évolué durant sa phase de conception ou d'apprentissage elle n'évolue plus en exploitation.
- Niveau 1 - Apprentissage déterministe : les sorties « possibles » de l'IA sont connues (et en nombre fini) mais le choix de la décision unique et déterministe (en fonction des entrées) peut évoluer par apprentissage.
- Niveau 2 - Apprentissage stochastique : les sorties « possibles » de l'IA sont connues (et en nombre fini); toutefois le choix de la décision n'est pas unique mais lié à une répartition probabiliste (en fonction des entrées) qui peut évoluer par apprentissage.
- Niveau 3 - Evolution : l'IA est susceptible d'évoluer en générant de nouvelles sorties voire en modifiant sa structure. La programmation génétique « continue » relève de cette catégorie.

Les modalités d'adaptabilité des systèmes IA requièrent une attention particulière. Il faut distinguer des IA susceptibles de s'adapter « en continu » pendant leur opération d'IA en capacité d'évoluer entre deux opérations ce qui permet de les tester, valider, qualifier avant « remise » en opération et surtout de mesurer le contrôle qu'ont les opérateurs humains sur ce processus d'évolution.

5. Les défis en IA des systèmes critiques

De nombreux verrous freinent l'adoption de l'IA, en particulier pour un déploiement dans systèmes critiques qui, par construction, doivent garantir des propriétés de sécurité (cybersecurity) et de sûreté (safety) mais aussi suivre des principes de confiance et de responsabilité. En effet, intégrer des solutions d'IA basées sur des techniques d'apprentis- sage ou sur des approches plus symboliques entraîne une série de difficultés vis-à-vis des pratiques actuelles d'ingénierie : comment spécifier un domaine d'usage à grande dimension (applications basées sur la vision ou le langage naturel), comment garantir la robustesse des systèmes à base d'IA au regard de leurs vulnérabilités de type cyber, com- ment garantir la réplicabilité des algorithmes intrinsèquement non déterministes. L'enjeu pour Thales est alors de définir et outiller cette démarche de « ingénierie de l'IA » (AI Engineering) de bout en bout, en prenant en compte les dimensions algorithmiques, logicielles et systèmes.

5.1 EXPERTISE TECHNIQUE, EXPERTISE MÉTIER ET EXPERTISE IA

Un défi majeur de l'introduction de l'IA dans les systèmes critiques est certainement l'équilibre entre l'exploitation de l'expertise métier (typiquement concernant la physique des senseurs), de la connaissance du domaine (trafic aérien, opérations militaires...) et de la maîtrise des technologies avancées d'intelligence artificielle.

Ce qui caractérise Thales et amène à des différentiateurs majeurs ce sont ses compétences dans ces trois domaines. Les équipes intégrées qui développent les solutions sont constituées notamment d'experts techniques, d'experts «métier» et d'experts en IA. Dans l'état actuel des systèmes informatiques et des capacités de gestion des données, il reste indispensable, pour garantir les performances attendues de certaines applications, d'effectuer des prétraitements sur les données issues des capteurs. Par exemple, un radar génère des téraoctets de données par heure et un système de guerre électronique des petaoctets de données par heure... Seule l'expertise dans ces domaines permet de concevoir des prétraitements qui transforment des données brutes en « pistes » qui, elles, peuvent être appréhendées par des systèmes d'IA. De la même façon, c'est l'expérience et la connaissance du métier qui permettent de définir quels doivent être le rôle, les sorties pertinentes et les limites d'un système d'assistance tels que ceux que l'IA rend possibles aujourd'hui.

Thales constitue des équipes multidisciplinaires composées d'ingénieurs ayant des doubles compétences permettant d'appréhender efficace- ment l'introduction de l'IA dans certaines solutions. Typiquement ce sont des experts à la fois en IA et en cybersécurité qui définissent les architectures de référence pour la cybersécurité de l'IA (§6.3.2) et ce sont des experts à la fois en IA et en traitement du signal qui définissent celles pour la phonie « augmentée » (compression, débruitage, intelligibilité améliorés par des technologies de Deep Learning) ou pour les nouvelles solutions en guerre électronique.

6. La réponse de Thales

6.1 LES TECHNOLOGIES CLÉS

La validité, l'explicabilité, la sécurité et la responsabilité sont des pré- requis pour aller vers la qualification, l'homologation voire la certification d'un système critique à base d'IA. Il est donc nécessaire d'apporter des garanties permettant une utilisation dans les domaines d'usage de Thales. C'est pourquoi, le Groupe se focalise plus particulièrement sur huit technologies différenciantes décrites ci-dessous. (Fig.2)

6.1.1 IA Hybride

Malgré les récentes avancées, les défis induits par le déploiement de l'IA dans les système critiques restent encore difficiles. En effet, tout d'abord, l'IA dirigée par les données manque de transparence, d'interprétabilité, de robustesse et s'avère peu frugale (en données et en énergie); ensuite, l'IA symbolique n'est pas toujours robuste face aux incertitudes. Pour pallier les faiblesses de chacunes de ces approches, l'IA hybride s'appuie sur les forces des méthodes connexionnistes en y intégrant au- tant que possible toutes connaissances disponibles (expertise métier, connaissances physiques, mathématiques, etc...) lors de l'élaboration des modèles, et cela pour en faciliter l'interprétabilité, garantir la robustesse et permettre ainsi leur validation en vue d'une homologation, voire une certification. La DARPA a d'ailleurs lancé en 2022 le programme « Assured Neuro Symbolic Learning and Reasoning » (ANSR) pour relever ces défis sous la forme de nouveaux algorithmes d'IA hybrides (neuro-symboliques) qui intègrent profondément le raisonnement symbolique à l'apprentissage afin de créer des systèmes robustes, sûrs et donc dignes de confiance.

Dans cette mouvance, la nouvelle classe de réseaux de neurones « Physics-Informed Neural Networks » (PINNs) repose sur une hybridation avec des modèles physiques. Ces réseaux neuronaux sont entrainés pour résoudre des tâches d'apprentissage supervisé tout en respectant toutes les lois de la physique décrites par des équations différentielles qui limitent l'espace des solutions admissibles par le réseau de neurones lors de la phase d'apprentissage. Les lois physiques pouvant être intégrées dans ces PINNs sont très diverses, et vont de la mécanique des fluides (équations de Navier-Stokes) à l'électromagnétisme (équations de Maxwell) en passant par la thermique (équations de Fourrier). Plus généralement, toute loi physique sous forme d'équations différentielles peut être capturée par un PINN. Cette technologie va avoir dans les prochaines années des applications importantes dans de nombreux domaines comme l'aéronautique ou la défense au travers, par exemple, des jumeaux numériques.

La Géométrie de l'Information est devenue un outil très populaire en IA utilisant le gradient associé à la métrique de Fisher pour prendre en compte la structure géométrique de l'espace des paramètres des ré- seaux multicouches. Des premières preuves de concepts de « Geometric-Informed Neural Networks » (GINNs) ont été mises en œuvre pour des fonctions comme l'ATDR (Automatic Target Detection & Recognition) sur les signatures micro-Doppler, les cinématiques des cibles ou sur de la reconnaissance d'images à partir de caméras 360° fisheyes.

Figure 2 : Les technologies différenciantes en IA

6.1.2 IA Frugale

L'apprentissage d'un système d'IA est gourmand en données annotées, alors que dans le domaine de la défense, les données sont souvent rares ou sensibles. L'enjeu est de concevoir des solutions d'apprentis- sage frugal et de création de bases de données représentatives à partir de peu de données réelles annotées. Pour cela, Thales s'appuie sur :

- **La compression de modèles** qui vise à réduire la taille des modèles d'IA en utilisant des techniques de compression telles que la quantification, la troncature ou la distillation de modèle. Ainsi, en réduisant la taille des modèles, on peut les exécuter efficacement sur des appareils moins puissants.
- **L'apprentissage fédéré** qui permet de former des modèles d'IA en distribuant le processus d'apprentissage sur plusieurs appareils ou serveurs locaux réduisant ainsi la dépendance vis-à-vis d'infrastructures centralisées coûteuses tout en respectant la confidentialité des données en gardant les données sur les appareils locaux.
- **L'utilisation de techniques d'apprentissage par transfert** en utilisant un modèle d'IA pré-entraîné sur une tâche spécifique et en l'adaptant pour une tâche similaire, ce qui permet de réduire le temps et les ressources nécessaires pour former de nouveaux modèles spécifiques.

Enfin, Thales conçoit des systèmes d'IA moins énergivores grâce à des matériaux nanoélectroniques innovants et des approches neuromor- phiques pour stocker l'information et réaliser des calculs.

6.1.3 Apprentissage par renforcement et simulation

L'apprentissage par renforcement consiste à optimiser des stratégies en se basant sur une approche essai-erreur avec un éventuel feedback hu- main et des outils spécifiques de simulation. Sous réserve de disposer d'une fonction d'évaluation permettant de juger de la qualité d'une réponse au problème ou de la qualité d'une stratégie, un système basé sur de l'apprentissage par renforcement va pouvoir parcourir l'espace des solutions possibles – par différentes méthodes – afin d'améliorer, essai après essai, sa proposition de solution.

La fonction d'évaluation peut être un simple feedback humain indiquant si une solution est meilleure qu'une autre (évaluation relative) ou un jugement humain étalonné (évaluation absolue). La fonction d'évaluation peut également être le résultat d'un calcul spécifique, par exemple le nombre de points obtenus pour un jeu. La puissance de l'apprentissage par renforcement réside dans sa gestion des évaluations sporadiques (parfois, on ne peut évaluer la qualité d'un coup dans un jeu qu'à la fin de la partie !) et sa capacité à distinguer - lors de l'évaluation - ses actions qui ont eu un effet positif de celles qui ont eu un effet négatif.

La plupart du temps, l'apprentissage par renforcement s'appuie sur des simulations. Le système « apprend » grâce à un simulateur et il est déployé dans le monde réel une fois son apprentissage terminé. Cela permet, en théorie, d'effectuer de nombreuses itérations d'essai-erreur. La qualité (finesse des modèles, représentativité…) de la simulation ainsi que le traitement du « reality gap » (différence entre la simulation et le monde réel) conditionnent la qualité de l'apprentissage. Thales possède une expertise considérable en simulation (qu'il s'agisse de simulations de capteurs, de phénomènes physiques, d'équipements réels – simulateur d'avions, de chars, d'hélicoptères dédiés à l'entraînement, etc…). Cette expertise lui confère un avantage certain pour développer des systèmes d'apprentissage par renforcement basés sur des simulations.

Début 2020, EDF, Thales et TotalEnergies ont ouvert le laboratoire SINCLAIR (Saclay INdustrial Collaborative Laboratory for Artificial Intelligence Research) sur le site d'EDF Saclay. Son programme de recherche est dédié à l'élaboration de méthodes et outils d'IA répondant à des nécessités partagées de ces trois entreprises dont un des trois axes de R&T est celui de l'apprentissage par renforcement avec l'explicabilité et la simulation.

6.1.4 Explicabilité

Depuis 2012, Thales travaille sur l'explicabilité des systèmes d'aide à la décision multicritère. Puis dans le cadre du laboratoire SINCLAIR, les travaux sur l'explicabilité se concentrent sur le fait qu'un système automatique puisse rendre compte du raisonnement qui l'a mené à la réponse proposée. Sans une réelle explicabilité, l'IA restera une boîte noire. Cependant, les explications doivent être adaptées au niveau de compréhension de la personne auxquelles elles sont destinées. Cette capacité peut aussi aider les développeurs à vérifier que le système fonctionne comme prévu. L'IA explicable contribue ainsi à promouvoir la confiance de l'utilisateur final ou l'auditabilité des modèles. Elle permet également d'atténuer les risques de conformité, de droit, de sécurité et de réputation liés aux résultats produits par l'IA.

Les IA génératives permettent de générer des textes extrêmement bien rédigés, même lorsqu'ils énoncent des contrevérités. Typiquement, à sa sortie, lorsqu'on demandait à ChatGPT quel était le plus lourd entre un œuf d'éléphant et un œuf de baleine, il répondait que c'était celui de l'éléphant avec une explication très bien écrite et qui pouvait sembler parfaitement logique pour un enfant.

Il faut donc distinguer avec la plus grande vigilance les IA convain- cantes des IA explicables qui doivent être en mesure de présenter à des experts d'un domaine ou à des utilisateurs, les éléments pertinents à leur portée.

6.1.5 IA Générative

En s'appuyant sur une séquence textuelle d'entrée, un LLM (Large Language Model) va générer une suite (déterministe ou pseudo-aléatoire suivant le choix de l'utilisateur) qui la poursuit ou la complète de façon consistante en se basant sur l'apprentissage (Deep Learning) effectué à partir d'un corpus éventuellement gigantesque (typiquement tous les textes accessibles sur internet) et de règles d'acceptabilité (par exemple, des règles éthiques supposées interdire de générer un texte sexiste ou raciste). Ce principe même de fonctionnement implique qu'un LLM génère des séquences « logiques » (et parfaitement acceptables au sens de la grammaire, de l'orthographe et de la conjugaison) mais dont le sens peut-être totalement erroné. On parle alors d'hallucinations.

De la même façon, un LVM (Large Vision Model) est conçu pour ap- prendre automatiquement les structures et connexions sous-jacentes de vastes ensembles d'images (ou de vidéos) et permet d'en générer de nouvelles.

L'enjeu majeur de l'usage de ces technologies est la capacité à exploiter les données opérationnelles, éventuellement confidentielles sans les exposer, ainsi qu'à en maîtriser le contenu généré (vérification, validation…).

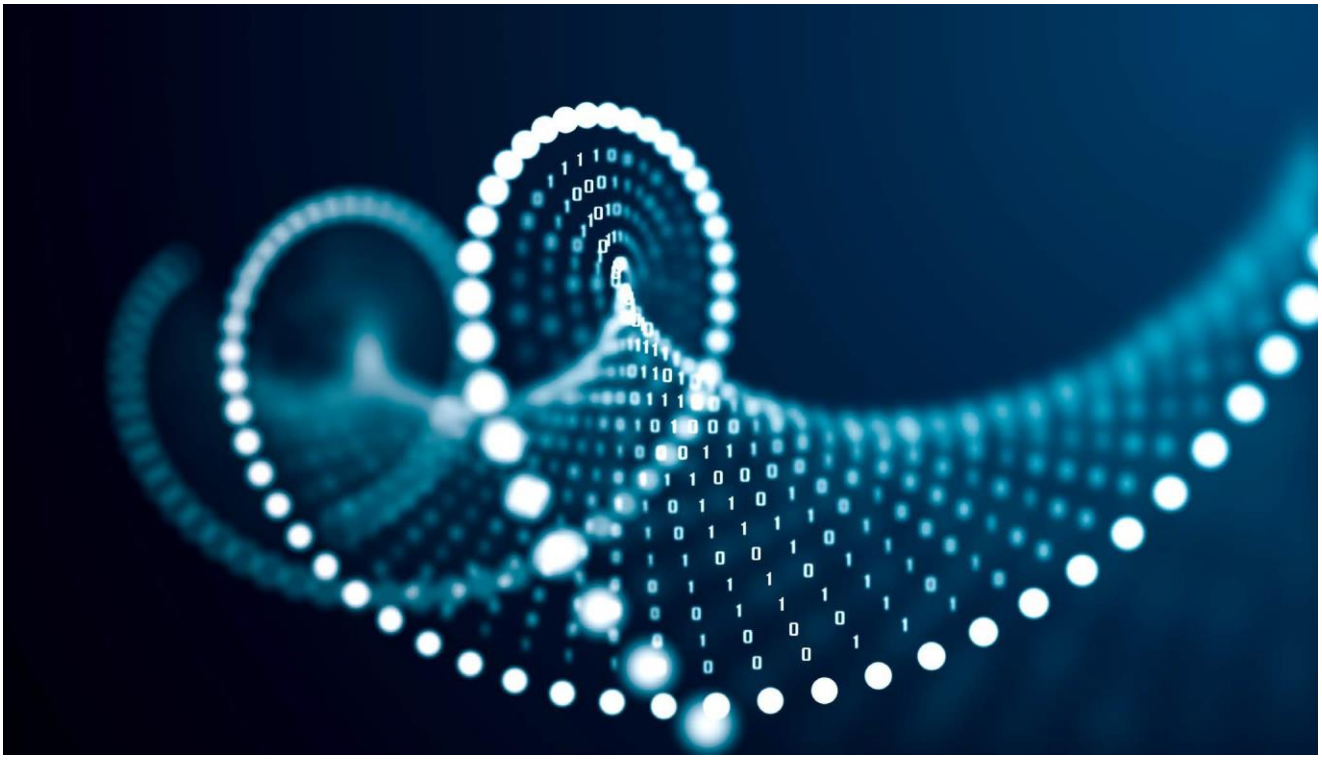

Figure 3 : IA générative @Thales

La maîtrise des technologies d'IA générative permet d'envisager, pour Thales, de nombreuses applications : (Fig.3)

- Aide aux fonctions transverses telles que les ressources humaines, le commerce, le marketing, le juridique (génération de résumés automatiques, génération de visuels originaux, exploitation des archives pour adaptation à un nouveau contexte, etc...).
- Génération automatique de code ou transcodage.
- Exploitation avancée de fonds documentaires opérationnels (documentation technique par exemple) avec agent conversationnel.

Plus d'une centaines d'expérimentations sont en cours à Thales sur ces applications. Les progrès récents et exponentiels de la mise en système des LLM/LVM telle que l'approche RAG (Retrieval Augmented Gene- ration) basée sur un LLM qui exploite un système moderne d'extraction d'information connecté aux données opérationnelles d'une entreprise afin de répondre à toutes sortes de questions ou l'approche ReACT (Reasoning and Acting) basée sur un LLM qui génère des formes de raisonnements interprétables par les utilisateurs humains et potentielle- ment activables offrent des perspectives disruptives à Thales. Quelques initiatives Thales ont démarré et exploitent ces approches :

- GenAI4SOC (Generative AI for Security Operating Centers) qui a pour objectif de générer automatiquement les règles de détection de menaces cyber.
- GenAI4MCS (Generative AI for Mission Critical Systems) qui com- bine aux LLM des technologies d'agents, d'appels à des services métier et de cyber-sécurisation pour développer le futur des assistants conversationnels autonomes et adaptatifs des systèmes de commande et contrôle.

Enfin, l'étude et la maîtrise des avancées en matière de multimodalité des IA génératives pour lesquelles l'espace de représentation interne vectoriel des données, quels que soient leurs types, est consistant (typiquement, le chiffre 3, les mots « trois » ou « three », une image représentant trois objets ont des représentations internes équivalentes ou très proches). Cela permettra de proposer des systèmes de génération de données atypiques réalistes et consistantes entre elles comme, par exemple, des données radar et des trajectoires de cibles réalistes et cohérentes entre elles vis-à-vis d'une situation tactique simulée. (Fig.4)

Il s'agit, en quelque sorte, de proposer pour les systèmes critiques des solutions comparables aux assistants conversationnels intégrés aux applications de certaines suites bureautiques, telle que celle de Microsoft, capables d'exploiter l'intégralité des données de chaque utilisateur (ses mails, son agenda, ses documents, ses tableurs, ses présentations...) afin d'améliorer de façon significative sa productivité en réalisant à sa place un certain nombre de tâches.

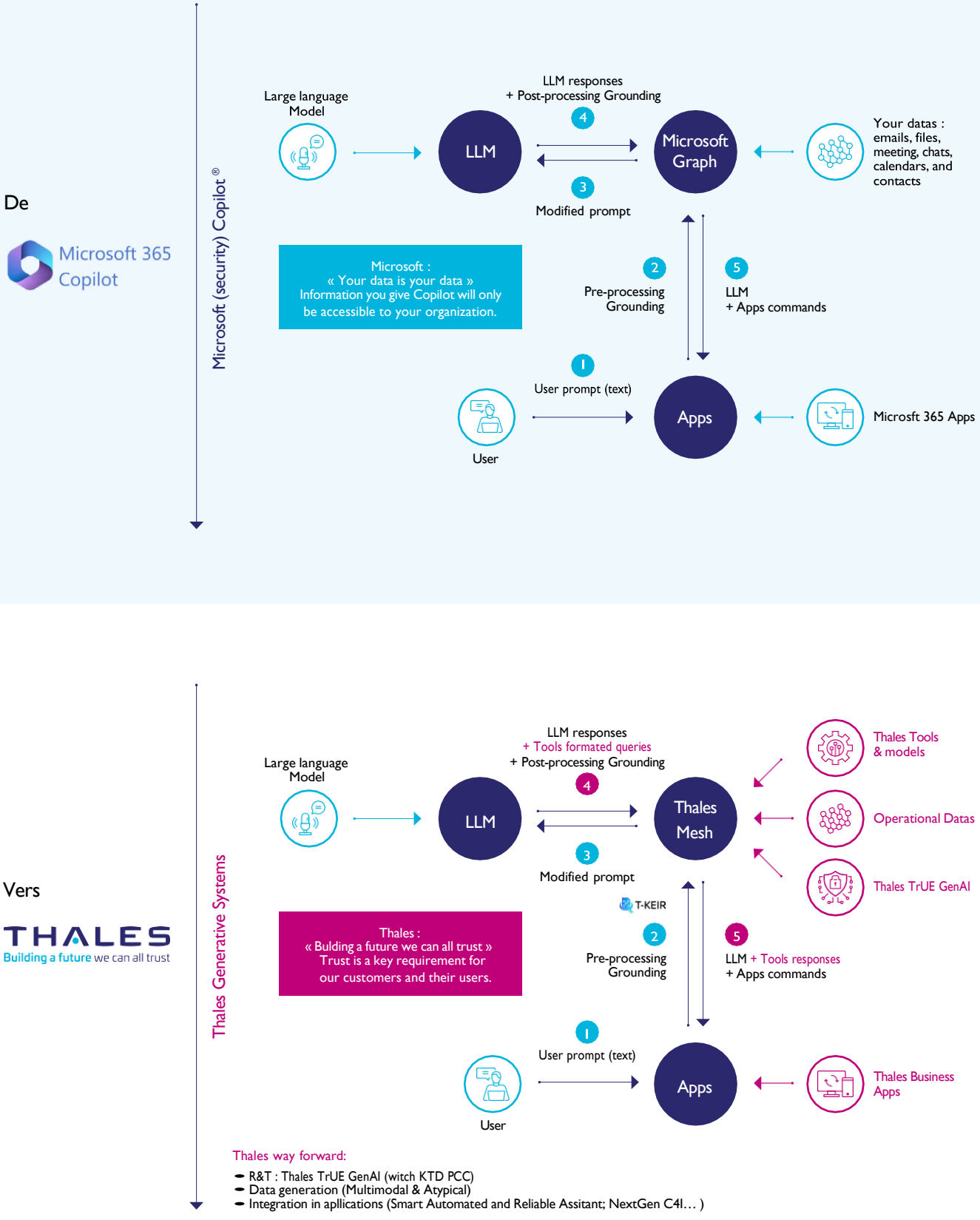

Figure 4 : IA générative pour les systèmes critiques

6.2. LES CONTRAINTES SYSTÉMES

6.2.1 Du cloud à l'extrême Edge

Passer d'une IA centralisée sur le cloud à une IA embarquée, plus proche des utilisateurs finaux est le principal défi de l'« Edge AI ». Ce concept désigne la mise en œuvre de l'IA dans un environnement d'edge computing. Cette technologie permet de traiter les données en temps réel, sans avoir à se connecter à un cloud. Les capteurs peuvent ainsi prendre des décisions plus intelligentes plus rapidement. Avec cette technologie, les capacités de calcul haute performance sont étendues aux sites d'edge computing, là où se trouvent les capteurs et les objets connectés. Les utilisateurs peuvent traiter des données en temps réel sur leurs appareils, car aucune connectivité ni intégration entre les systèmes n'est nécessaire. Les bénéfices pour les applications de Thales sont alors multiples comme :

- **La consommation d'énergie** : la mise en œuvre des algorithmes d'IA au plus près du capteur et de l'utilisateur, en particulier au sein de systèmes embarqués, permet de réduire la consommation énergétique et les délais de traitement de manière drastique, tout en réduisant les coûts et les risques liés à la transmission des données.
- **La réduction de la bande passante** : la diminution de l'utilisation de la bande passante dans les flux de données et la minimisation des coûts grâce au traitement, à l'analyse et au stockage des données localement au lieu de les envoyer dans le cloud.
- **La sécurité et la confidentialité** : la priorité accordée au transfert de données modifie complètement la taille et la forme de la surface d'attaque. De plus, l'Edge AI permet de filtrer les données à faire remonter dans le cloud si nécessaire. Seules les informations souhaitées sont transmises, après avoir été, par exemple, anonymisées.
- **La réduction de la latence** : la réduction de la charge de la plateforme cloud et l'analyse des données en local permettent de libérer la plateforme pour d'autres tâches.
- **Une fiabilité améliorée** : en distribuant les calculs sur plusieurs appareils, l'Edge AI offre une meilleure redondance et une fiabilité accrue, car même si certains appareils échouent, les autres peuvent continuer à fonctionner.
- **Déploiement « sur le terrain »** : l'Edge AI se prête par définition très bien aux usages de l'Internet des objets (IoT), mais également de l'ensemble des appareils mobiles, à commencer par les plateformes autonomes comme les drones.

S'appuyant sur son expertise en HPC (High Performance Computing), Thales travaille depuis plus de dix ans sur l'IA embarquée.

6.2.2 Relation IA-Humain

Pour tout système critique susceptible de faire collaborer des humains et des machines « intelligentes », Thales entend définir le cadre et les éléments régissant les accès aux données sensibles, les actions autorisées, les rôles respectifs et les modes d'interaction en temps réel, en fonction de l'état global du système et du contexte.

Dans une perspective de collaboration efficace entre les humains et les IA, il convient d'établir sans ambiguïtés les prérogatives des uns et des autres en fonction de leurs compétences, performances et fiabilité respectives, de l'état du système avec lequel cette coopération s'exerce, de l'environnement (l'ensemble des paramètres exogènes), des éventuelles réglementations et de l'éthique (nationale, d'entreprise...).

Dans le même temps, il est impératif de préciser les degrés d'autonomie et d'adaptabilité des IA (§4) en fonction de l'état du système et de l'environnement. La notion de dialogue humain-IA n'est pas nouvelle mais les perspectives offertes par les IA génératives concernant les interactions en langage naturel s'appuyant, notamment, sur les LLM permettent d'envisager désormais un dialogue bien plus direct et efficace.

Lorsque les IA sont utilisées comme aide ou support à la décision humaine, l'humain reste maître des actions et en assume la complète responsabilité. Dans certains cas cependant, la décision et sa mise en œuvre ne sont pas compatibles avec le temps d'analyse humain et celui du dialogue entre ce dernier et une IA : typiquement, il n'est pas raisonnable d'imaginer qu'un véhicule autonome demande la validation explicite à un passager humain pour un freinage d'urgence (même si le fait d'enclencher cette possibilité relève de ses prérogatives).

Figure 5 : Relation humain-IA

6.3. MISE EN ŒUVRE

Le déploiement de l'IA dans des applications dites critiques nécessite la conformité à des objectifs de fiabilité, de maintenabilité, de disponibilité de sûreté et de sécurité (RAMS : Reliability, Availability, Maintainability, Safety). Ainsi, un système critique doit reposer sur des méthodes de développement rigoureuses, de sa conception à son déploiement et sa qualification. Les pratiques d'ingénierie doivent être alors enrichies par des méthodes et outils garantissant la confiance à toutes les étapes du cycle de vie d'un tel système : (1) spécification du domaine opérationnel et de sa déclinaison pour la gestion des données et des connaissances;
(2) conception d'algorithmes et d'architecture; (3) caractérisation, vérification et validation; (4) déploiement, en particulier sur une architecture embarquée; (5) qualification, certification et (6) maintien en condition opérationnelle et de cybersécurité.

6.3.1 L'ingénierie de l'IA de confiance

Pour l'aide à la conception d'un composant d'IA de confiance et plus particulièrement à base d'apprentissage (ML : Machine Learning), un processus MLOps (Machine Learning Operations), fortement inspiré de l'approche DevOps[1] a pour objectif primaire de parfaire le proces- sus de développement et de déploiement logiciel, ceci afin d'améliorer et d'unifier le développement et l'exploitation du composant. L'un des principaux avantages est qu'il permet un déploiement plus rapide. Ce- pendant, son application dans le contexte de systèmes critiques doit être repensée pour assurer leur transparence et leur auditabilité afin de corriger, le cas échéant, les défaillances, mais aussi pour aller jusqu'à la preuve de la conformité des propriétés attendues comme la validité, la robustesse, l'explicabilité, l'embarquabilité... Pour cela, Thales contribue depuis 2020, au programme Confiance.ai[2] pour définir des méthodologies et des outils de l'ingénierie de l'IA de confiance couvrant les étapes suivantes :

- La spécification du problème, capturée au travers des différentes exigences (fonctionnelles/non fonctionnelles) et de la couverture opérationnelle (ODD : Operational Design Domain), décrit les conditions dans lesquelles la capacité est conçue pour fonctionner correctement, comme les conditions environnementales et d'autres contraintes du domaine. Ceci impacte la tâche de collecte des don- nées et de modélisation des connaissances.
- L'acquisition des données/connaissances guidée par l'ODD aboutit à une agrégation des données/des connaissances en un ensemble homogène, de taille et de qualité suffisantes (compréhensible, pertinent, fiable, équilibré...). Cependant, pour pouvoir être utilisées, ces données sont en général nettoyées, organisées, voire labellisées. Dans certains cas, un traitement est nécessaire afin de rendre ces informations brutes exploitables. Il s'agit de la tâche de « Data Engineering » pouvant être complétée par du « Knowledge Enginee- ring ».
- Un algorithme d'IA peut être conçu ou sélectionné parmi une bibliothèque d'algorithmes existants. Dans le cadre du ML, une fois l'apprentissage terminé, le modèle est affiné en utilisant l'ensemble des données de validation. Cela peut impliquer la modification ou l'élimination de variables, l'ajustement des paramètres spécifiques du modèle (hyperparamètres) jusqu'à un niveau de précision acceptable. L'implémentation sur la plateforme matérielle et/ou le système cible peut impacter certaines exigences techniques comme la latence, l'espace mémoire ou la consommation énergétique. Puis, après avoir trouvé un ensemble acceptable d'hyperparamètres et optimisé la précision du modèle, ce dernier est testé et caractérisé sur un ensemble de données, voire évalué par une vérification formelle. L'évaluation peut aller au-delà de la performance fonctionnelle (telle que la précision) et englober des métriques relatives à tout autre critère de performance attendu comme la robustesse aux bruits et/ou aux attaques adverses.

Après intégration du composant IA/ML dans un système critique, il faut démontrer, en apportant les preuves, que celui-ci possède les « propriétés de confiance » attendues. Il faut donc définir un cadre d'analyse « d'Ingénierie système de l'IA de confiance » permettant d'élaborer les stratégies de développement de systèmes et d'IVVQ (Intégration, Vérification, Validation, Qualification).

6.3.2 La cybersécurité de l'IA

Si les technologies d'IA les plus modernes en apprentissage permettent d'obtenir des niveaux de performances inatteignables par les algorithmes classiques et rendent possibles de nouvelles capacités, elles s'accompagnent de vulnérabilités qui leur sont propres. Cela a conduit Thales à compléter – pour ces besoins spécifiques à l'IA – son expertise et ses moyens en cybersécurité. (Fig. 6)

(1) Le DevOps est une démarche qui consiste à faire collaborer étroitement les équipes de développement & études avec les équipes des opérations et d'exploitation.
(2) L'objectif du programme Confiance.ai (www.confiance.ai), pilier technologique du Grand Défi National « sécuriser, certifier et fiabiliser les systèmes fondés sur l'IA » du plan France 2030 est de proposer un atelier d'ingénierie(s) de l'IA de confiance, bâti sur des composants technologiques et méthodologiques, outillant ainsi de bout en bout le processus de conception d'un système critique à base d'IA.

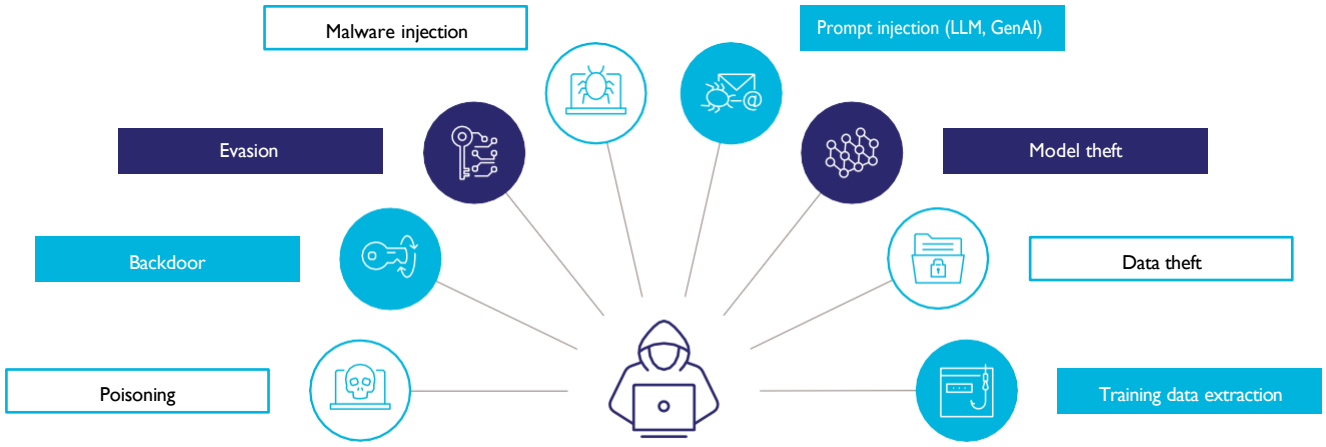

Figure 6 : Cybersécurité de l'IA : les attaques

Voici quelques exemples des attaques étudiées par l'équipe « AI Friendly Hacking » de Thales en charge des travaux de recherche sur la cybersécurité de l'IA :

- Les algorithmes d'apprentissage profond (Deep Learning) étant basés sur l'exploitation de données d'entraînement, il est évident que la qualité de ces données (représentativité et labélisations correctes notamment) influe sur les performances de l'algorithme. Afin de dégrader les performances d'un tel algorithme il est ainsi possible d'« empoisonner » les données d'apprentissage.
- Il est également possible de développer des IA capables de générer, par apprentissage, des données d'entrée susceptibles de tromper les meilleurs algorithmes d'IA de reconnaissance d'images. Ainsi l'équipe « AI Friendly Hacking » a conçu un algorithme d'IA qui modifie légèrement (modifications invisibles à l'œil nu pour un hu- main) les caractéristiques de certains pixels d'une image (leurs composantes RGB) afin de générer une réponse erronée de la part des meilleurs algorithmes au monde de reconnaissance d'images. Cette équipe s'est révélée capable de produire des attaques ciblées, c'est- à-dire choisissant même la réponse erronée donnée par l'algorithme attaqué (par exemple, modification de certains pixels de l'image d'un char pour le faire passer explicitement pour une ambulance).

Elle est également parvenue à produire des attaques génériques, c'est- à-dire capables de tromper des modèles différents de reconnaissance (issus d'éditeurs différents).

- Un autre type d'attaques réalisé par l'équipe « AI Friendly Hacking » consiste à extraire du modèle appris une partie des données d'apprentissage. Cela pose évidemment la question de la protection des données d'apprentissage, typiquement comme celles liées à la vie privée et encadrées par le RGPD (Règlement Général sur la Protection des Données) pour des modèles basés sur des visages humains.

L'équipe « AI Friendly Hacking » de Thales a remporté le challenge 2023 organisé par la Direction générale de l'armement (DGA) sur l'identification et l'extraction de données d'apprentissage d'un modèle censé s'être prémuni contre de telles attaques.

Une partie de l'équipe « AI Friendly Hacking » travaille sur les vulnérabilités intrinsèques des IA génératives. L'équipe a notamment été en mesure de contrecarrer les protections « éthiques » de ChatGPT (la version officielle déployée dans le cloud) en lui faisant rédiger un tutoriel pour fabriquer des bombes artisanales avec les ustensiles et produits que l'on trouve généralement dans une cuisine ou un garage ! Interrogé « normalement » sur un tel sujet, l'apprentissage par renforcement avec retour humain (Reinforcement Learning with Human Feedback ou RLHF) réalisé par des équipes d'opérateurs humains avant la sortie avant la sortie officielle de ChatGPT l'enjoint d'offrir une fin de non-recevoir, mais une IA développée spécifiquement par Thales a su générer un prompt (une question accompagnée de commandes spécifiques) amenant ChatGPT à faire une réponse cohérente et inquiétante...

La vocation de l'équipe « AI Friendly Hacking » n'est pas uniquement d'étudier et forger des attaques, elle s'évertue également à proposer des méthodes de défense des IA développées par Thales contre des attaques potentielles.

À titre d'exemple, l'équipe « AI Friendly Hacking » a développé, en partenariat avec le programme confiance.ai, des méthodes permettant d'insérer un watermark (une signature invisible) dans un modèle d'intelligence artificielle afin d'en démontrer la propriété dans le cas où il serait copié ou contrefait. (Fig. 7)

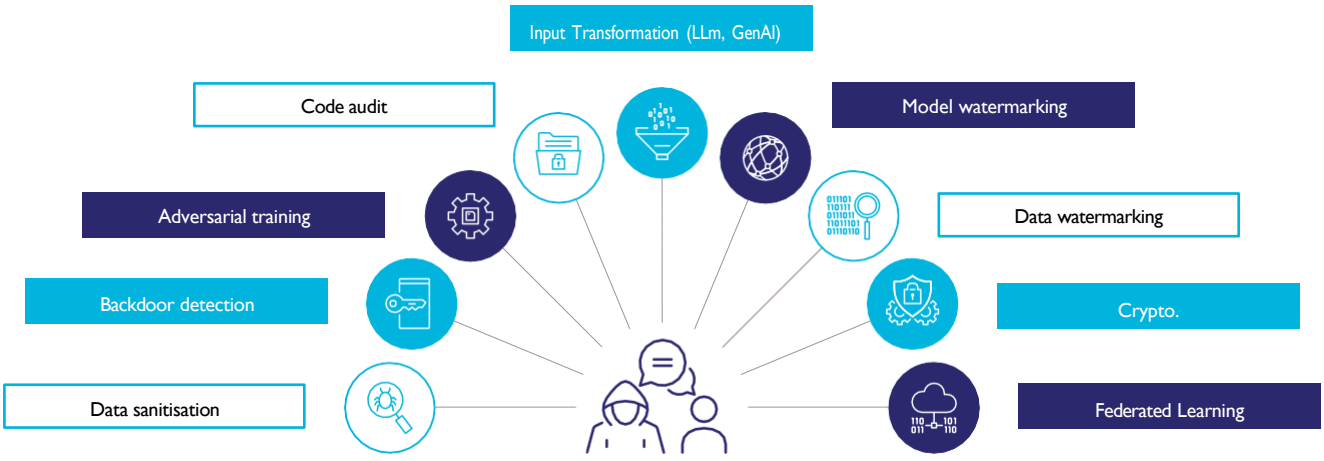

Figure 7 : Cybersécurité de l'IA : les défenses

7. Conclusion

Parce que les clients de Thales gèrent des opérations, des infrastructures et des services vitaux pour la société, leur confiance envers les solutions du Groupe est essentielle. Cette confiance est soutenue par une dé- marche constante d'innovation. Thales se concentre sur le renouvelle- ment de ses solutions à un niveau toujours croissant de performance, de sécurité et de durabilité. Ainsi le Groupe a souhaité se doter de lignes directrices pour un numérique responsable et de confiance au travers d'une charte éthique du numérique. (Fig.8)

Celle-ci insiste sur la nécessité, lors de la conception de système à base d'IA, de définir avec le client les cas d'emploi en garantissant que l'humain puisse conserver la capacité de reprendre le contrôle des systèmes, soit en amont de l'action, soit pendant l'action lorsque cela est pertinent (§ 6.2.2). L'IA est alors utilisée pour accroître les capacités de décision de l'humain et non pour se substituer à lui.

De plus, transmettre les éléments sur lesquels se fondent ces systèmes d'IA pour produire des recommandations est une condition nécessaire à la confiance. Ainsi, la mise à disposition de ces informations concerne aussi bien les règles de fonctionnement des algorithmes que la conception des outils numériques eux-mêmes, dans les limites liées à la confidentialité ou à la sensibilité des données concernées.

Les pratiques de **« privacy and cybersecurity by design »** sont appliquées dans le cadre du développement de ces systèmes. Dans cette optique, les ingénieurs et scientifiques de l'IA à Thales cherchent toujours le meilleur compromis entre la nature et la quantité de données utilisées et le résultat attendu, adoptant une démarche proportionnée d'utilisation et de consommation de données. Les approches dites de «smart data» sont privilégiées aux « big data », favorisant la qualité des données traitées ou transmises plutôt que leur volume. Enfin, la mise en œuvre d'outils et pratiques disponibles permettant d'éviter ou de détecter des biais lors de la conception de systèmes d'IA garantit l'utilisation d'échantillons de données équilibrés.

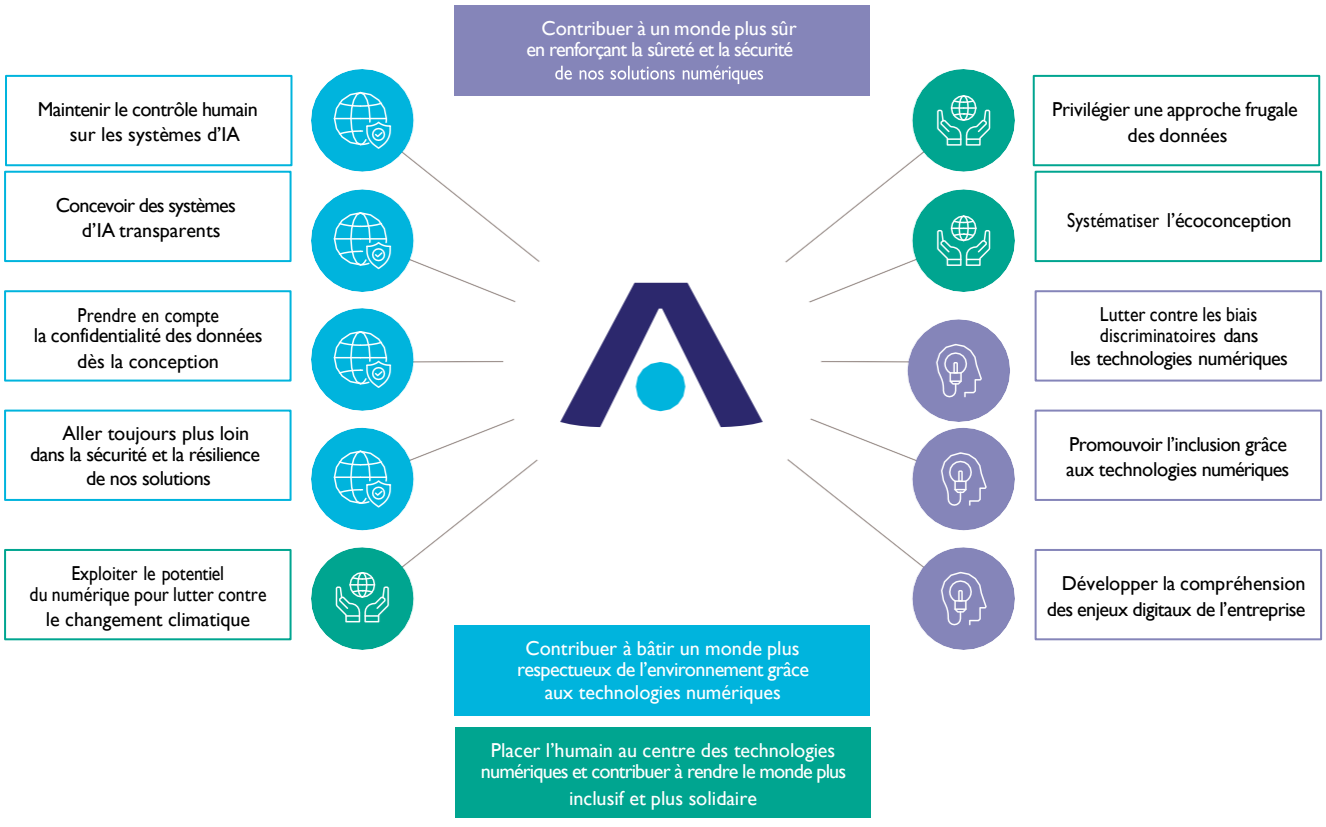

Figure 8 : Les 10 principes de la Charte Ethique du Numérique de Thales

4, rue de la Verrerie
92190 Meudon
FRANCE

Tél. + 33(0)1 57 77 80 00

www.thalesgroup.com