Maîtriser l'IA de confiance : Un cycle de vie complet pour les systèmes critiques

THALES
Building a future we can all trust

THALES
Building a future we can all trust

Maîtriser l'IA de confiance : Un cycle de vie complet pour les systèmes critiques

Juliette MATTIOLI, Patricia BESSON, Fateh KAAKAI,

Michel BARRETEAU, Rémi BLANCHETTE, Fabien FLACHET

Résumé

Le cycle de vie de l'ingénierie de l'Intelligence Artificielle (IA) de confiance, présenté dans ce livre blanc, met en évidence la nécessité d'une approche holistique et rigoureuse pour le développement, le déploiement et la maintenance des systèmes basés sur l'intelligence artificielle (IA) dans des domaines critiques tels que l'aéronautique et l'espace, la défense, ainsi que la cybersécurité et le numérique. Des principes fondamentaux de l'ingénierie des systèmes - qui englobent la sûreté, la sécurité, l'éthique et la caractérisation du domaine opérationnel du système - aux considérations spécifiques de l'ingénierie et de la mise en œuvre des algorithmes d'IA, ce cycle de vie offre un cadre complet pour faire face aux complexités et aux défis inhérents aux technologies de l'IA. L'intégration des pratiques MLO/AIOps et DevOps dans le cycle de vie de l'ingénierie renforce encore l'adaptabilité et la fiabilité des systèmes d'IA, en favorisant une collaboration transparente entre les scientifiques des données, les ingénieurs et les équipes opérationnelles, et en permettant l'intégration, le déploiement et la surveillance continus des modèles d'IA. De même, l'utilisation d'outils basés sur l'IA dans des activités d'ingénierie améliore l'efficacité de l'élaboration des exigences, des processus de test et de la génération de code. Cependant, la nécessité d'un cadre de qualification normalisé pour ces outils est cruciale pour s'assurer qu'ils contribuent à la fiabilité, à la sûreté et à la sécurité des systèmes critiques, plutôt que de les compromettre.

A l'avenir, les défis posés par la complexité croissante de l'IA et son intégration dans les systèmes critiques nécessiteront une collaboration soutenue entre l'industrie, les autorités réglementaires et le monde universitaire. Pour rester adaptables dans des environnements dynamiques tout en garantissant la confiance, les besoins attachés aux systèmes d'IA évoluent pour aller vers une notion de service et être à même de faire face à des enjeux tels que la garantie des performances temps réel ou le développement et la qualification incrémentaux Ces stratégies sont particulièrement importantes dans les domaines où les contraintes opérationnelles exigent des mises à jour rapides, en mettant l'accent sur l'équilibre entre la sécurité, la performance et d'autres objectifs opérationnels essentiels de la mission.

À travers le cycle de vie de bout en bout de l'ingénierie de l'IA de confiance, Thales réaffirme son engagement à développer des systèmes basés sur l'IA qui soient non seulement efficaces, sûrs et sécurisés, mais aussi responsables, transparents et éthiques. CortAIx, l'accélérateur d'IA de Thales, joue un rôle central dans les activités de recherche pour inventer et faire progresser de nouveaux concepts, algorithmes, méthodologies d'assurance et outils d'ingénierie pour l'industrialisation des produits et services d'IA. En repoussant constamment les limites de la technologie de l'IA, Thales établit de nouvelles références d'excellence dans les applications de systèmes critiques, dessinant l'avenir de l'IA dans les environnements les plus exigeants et les plus critiques.

1. Introduction et contexte

L'adoption rapide de l'intelligence artificielle (IA) révolutionne les capacités opérationnelles dans des domaines critiques tels que l'aéronautique et l'espace, la défense, le cyberespace et le numérique [12]. Toutefois, cette transformation s'accompagne de défis importants pour garantir que les systèmes basés sur l'IA sont dignes de confiance, c'est-à-dire valides, explicables, robustes et responsables[1]. Les applications d'IA de pointe, tirant parti de l'apprentissage automatique, de l'IA symbolique et de l'IA hybride, doivent en effet fournir des solutions correctes et robustes, fondées sur des données pertinentes pour les domaines critiques. Ces avancées nécessitent de définir un cadre structuré pour garantir la confiance dans les systèmes d'IA par des pratiques d'ingénierie rigoureuses, cadre méthodologique qui fait l'objet du présent livre blanc.

Au niveau international, l'Agence Européenne de la Sécurité Aérienne (AESA), acteur mondialement reconnu pour la certification des systèmes critiques, a défini une vision globale de l'IA de confiance dans sa feuille de route [14] et son « Concept Paper » [2] publiés, respectivement en 2023 et 2024. Ces documents mettent l'accent sur la sûreté, la cyber-sécurité, la transparence et la responsabilité. En complément, Thales est fortement impliqué dans les initiatives internationales de normalisation telles que celles menées par le WG-114[2] de l'EUROCAE en collaboration avec le SAE G34, qui fournissent des lignes directrices pour la certification de l'IA dans les systèmes critiques. L'ISO et le CENELEC contribuent également au paysage mondial de la normalisation avec leurs normes intersectorielles pour les applications éthiques et fiables de l'IA.

Au niveau national français, Thales est l'un des membres fondateurs, avec l'IRT SystemX et d'autres partenaires industriels, du projet français Confiance.ai[3], pilier de l'initiative « Grand Défi de l'IA de confiance pour l'industrie ».

Ainsi, ce livre blanc présente le cycle de vie de l'ingénierie de l'IA de confiance de bout en bout : une approche globale de la conception, du développement et du déploiement de systèmes basés sur l'IA dans des environnements critiques.

En intégrant les principes énoncés dans les initiatives internationales et nationales aux pratiques internes d'ingénierie avancée, ce cycle de vie garantit que les systèmes d'IA remplissent d'une part les fonctions prévues - et uniquement les fonctions prévues - avec le niveau de performance souhaité, et d'autre part, que les solutions basées sur l'IA soient transparentes, responsables et éthiques.

Ce document examine en outre comment ces cadres d'ingénierie de l'IA peuvent être mis en œuvre dans l'écosystème de normalisation existant, en fournissant des informations exploitables pour les clients et les parties prenantes, notamment les ingénieurs, les chercheurs et les décideurs politiques. La couche d'ingénierie algorithmique de confiance présentée dans ce Livre Blanc agit comme un pont, traduisant les exigences de systèmes complexes en modèles d'IA sophistiqués, tout en assurant la compatibilité avec les processus d'ingénierie de systèmes existants. Elle introduit des méthodes et des outils adaptés à la nature itérative du développement de l'IA, permettant des mises à jour fréquentes à mesure que de nouvelles données ou connaissances deviennent disponibles.

Par ailleurs, cette couche d'ingénierie algorithmique vise à normaliser les interfaces entre les algorithmes d'IA et les couches logicielles/matérielles, afin de garantir une intégration fluide et de réduire le risque d'erreurs lors de la mise en œuvre et du déploiement [1].

L'introduction de cette nouvelle couche a des implications significatives pour les processus d'ingénierie existants, qui sont souvent très normalisés. En s'alignant sur les normes établies, telles que celles qui régissent l'ingénierie des systèmes, des logiciels et du matériel, la couche d'ingénierie algorithmique garantit la cohérence et la traçabilité du développement de l'IA. De plus, elle favorise la collaboration entre les équipes pluridisciplinaires, permettant aux ingénieurs système, aux scientifiques des données et aux développeurs de logiciels de travailler en synergie pour atteindre des objectifs communs. Cette approche globale améliore la fiabilité et la robustesse des systèmes d'IA, garantissant qu'ils répondent aux exigences strictes des applications critiques.

[1] **En cohérence avec l'approche Thales TrUE AI (Transparent, Understable, Ethical AI)**

[2] **Thales est co-chair avec Airbus de ce groupe de travail EUROCAE**

[3] **Pour plus de détails sur le programme voir** Confiance.ai

2. Ingénierie de l'IA de confiance

2.1 INGÉNIERIE SYSTÈME

Le développement de systèmes basés sur l'IA pour des applications critiques [12], quelle que soit la technologie d'IA utilisée (IA connexionniste et statistique, IA symbolique ou hybride), nécessite des pratiques d'ingénierie système rigoureuses. En effet, la fiabilité, la disponibilité, la maintenabilité, la sécurité, la cybersécurité, l'utilisabilité et l'éthique sont des propriétés clés qui ne peuvent être évaluées et spécifiées qu'au niveau « système ».

A cet effet, dès 2019, Thales a mis en place l'approche "Thales TrUE AI" fondée sur trois piliers essentiels (Figure 2)

– Transparence : Les données pertinentes utilisées pour parvenir à un résultat sont présentées aux utilisateurs, permettant ainsi une meilleure compréhension des processus décisionnels de l'IA.
– Compréhensibilité : L'IA doit pouvoir être expliquée et ses résultats justifiés, rendant ses décisions plus acceptables et fiables.
– Éthique : L'IA respecte les protocoles de normes objectives, les lois (comme la loi européenne sur l'IA) et les droits de l'homme, assurant ainsi une utilisation responsable et éthique de la technologie.

Domaine d'exploitation du système

Cette étape cruciale bouscule les habitudes des chercheurs ou des ingénieurs en IA, car si le domaine d'exploitation du système fait déjà l'objet d'une description aujourd'hui, le développement d'une solution basée sur l'IA exige une caractérisation encore plus précise et approfondie de l'ensemble des conditions d'exploitation prévisibles, appelées « environnement d'exploitation du système ». Cette description nécessite une collecte de données et une représentation des connaissances adéquates. La fiabilité des algorithmes d'IA dépend directement de l'exactitude et de l'exhaustivité de cette description, en particulier des événements nécessitant une considération spécifique, tels que les événements rares, les combinaisons ou les séquences de conditions d'exploitation présentant un risque potentiel pour la sécurité.

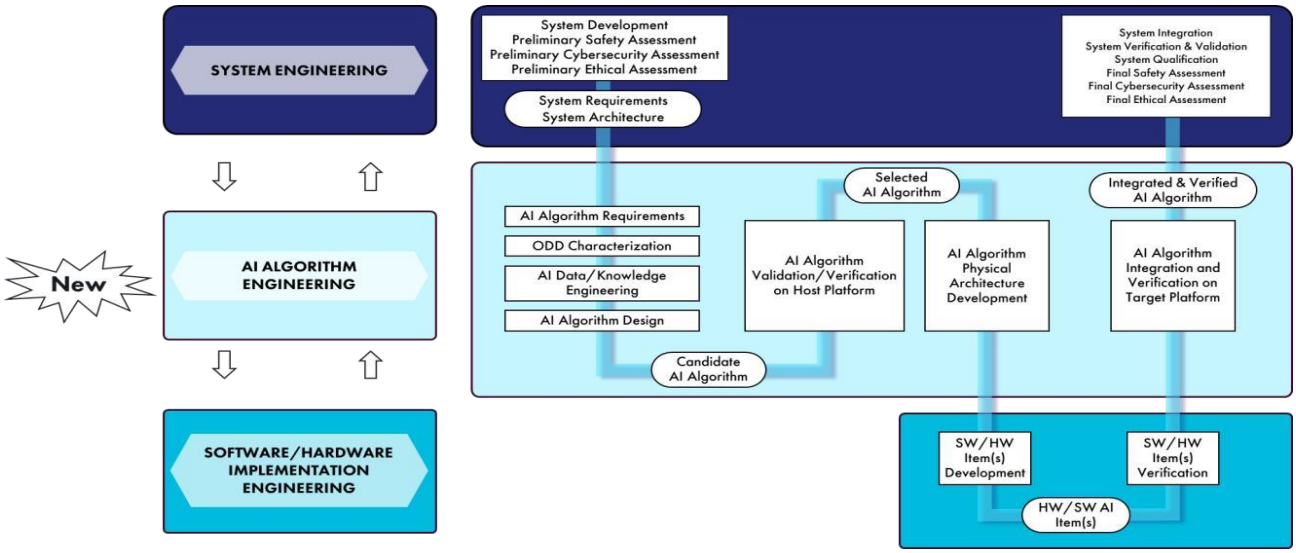

Figure 1 : Couche d'ingénierie algorithmique spécifique aux systèmes basés sur l'IA

Sûreté

Garantir la sécurité, et au-delà la fiabilité, la disponibilité et la maintenabilité, signifie que les systèmes d'IA doivent fonctionner et continuer à fonctionner dans le temps comme prévu, dans toutes les conditions opérationnelles prévisibles, y compris les cas limites/coins et les scénarios de défaillance. Cela permet d'éviter des conséquences imprévues qui pourraient mettre en péril des vies humaines ou des fonctions essentielles à la mission.

L'analyse des dangers et l'évaluation des risques réalisées au niveau « système » tiennent compte des particularités de l'IA. Elles intègrent par exemple les risques d'erreurs

critiques, la présence de biais dans les données d'apprentissage ou la représentation des connaissances. La capacité du modèle à généraliser à des données non vues lors de la conception est également prise en compte.

Les exigences de performance fixées à l'algorithme d'IA sont alors guidées par des objectifs de sécurité. Cela permet de limiter la pire erreur d'approximation crédible à un seuil acceptable.

Figure 2 – IA de Confiance : « Thales TrUE AI – Transparent, Understandable, Ethical »

La validité, l'explicabilité, la sécurité et la responsabilité sont des conditions préalables à la qualification, à l'approbation, voire à la certification d'un système critique basé sur l'IA. Pour atteindre ces objectifs, l'ingénierie des systèmes pour les applications basées sur l'IA combine les pratiques d'ingénierie traditionnelles avec des considérations spécifiques à l'IA, garantissant ainsi que les systèmes développés sont non seulement performants, mais aussi dignes de confiance et éthiques.

Cybersécurité

La cybersécurité met en œuvre des mesures exigeantes pour protéger les systèmes d'IA contre les attaques adverses, les violations de données et les manipulations non autorisées. Ces systèmes intègrent des stratégies avancées de détection et d'atténuation des menaces, ainsi que des mécanismes de résilience qui leur permettent de fonctionner en toute sécurité dans des environnements dynamiques et hostiles. Les mesures de cybersécurité doivent être intégrées non seulement au niveau du système, mais aussi dans les pipelines de données sous-jacents, afin de garantir l'intégrité et la confidentialité des données tout au long du cycle de vie de l'IA.

Il est également important de noter que la frontière entre la sécurité et la sûreté n'est pas étanche dans le contexte de l'IA. Des changements dans les entrées du modèle conduisant à des résultats erronés peuvent résulter d'actions malveillantes intentionnelles. A l'opposé, des mesures de cybersécurité peuvent avoir un impact sur le fonctionnement nominal d'un modèle IA. Une approche globale est donc indispensable pour traiter à la fois les attaques malveillantes (telles que l'altération des données ou les manipulations adversariales) et les risques inhérents (tels que les défaillances techniques ou les aléas environnementaux), afin de garantir la robustesse et la fiabilité des systèmes d'IA.

2.2 INGÉNIERIE ALGORITHMIQUE D'IA

Apprentissage automatique

Pour l'IA connexionniste et statistique, en particulier pour l'apprentissage automatique (ML), l'algorithme d'IA est composé de différents éléments ML qui sont : un ou plusieurs modèles ML et les traitements de données associés (concept de constituent ML dans [2]). Ces éléments ML sont ensuite intégrés dans le système auquel ils appartiennent.

Le développement de systèmes basés sur le ML peut être visualisé comme un cycle de vie en forme de W [2]. La forme en W peut être divisée en deux parties : le 1er V comprend les processus d'ingénierie ML effectués sur la plateforme hôte (par exemple sur un cluster de calcul privé ou sur une partition sécurisée dans le cloud), et le 2nd V comprend les processus d'ingénierie de mise en œuvre effectués sur la plateforme cible (ex. matériel spécifique intégré dans un véhicule terrestre ou aérien).

Dans le 1er V, le cycle de vie de l'ingénierie du ML [1][2][4] commence par la définition des exigences de l'algorithme d'IA/ML affinées à partir de la spécification système. Cette étape de spécification est complétée par la caractérisation

du domaine opérationnel de conception (ODD). Cette description est élaborée, d'une part, comme un raffinement de la partie du domaine opérationnel du système allouée à l'algorithme d'IA/ML (approche descendante) et, d'autre part, comme une analyse centrée sur les données (approche ascendante). L'ODD est un moyen de concilier les données et l'intention fonctionnelle, c'est-à-dire d'aligner les données utilisées pour l'entraînement et le ou les modèles ML résultants avec l'utilisation opérationnelle prévue, en capturant un large éventail de conditions de fonctionnement qu'ils peuvent rencontrer [5].

L'ingénierie des données joue un rôle fondamental, impliquant l'identification, la collecte, le prétraitement et l'extraction des caractéristiques de vastes ensembles de données essentiels à la conception des modèles ML et à leur vérification. Cette phase intègre souvent des techniques avancées telles que l'augmentation des données, la génération de données synthétiques et la détection d'anomalies afin d'améliorer la représentativité, l'exhaustivité et la pertinence de l'ensemble de données (minimisation de l'écart entre la simulation et la réalité). Des processus de contrôle qualité rigoureux, basés sur des exigences spécifiant les critères de qualité des données (*Data Quality Requirements* ou DQRs), garantissent la qualité des données d'entrée (par exemple, les critères de représentativité et de complétude des données vis-à-vis de l'ODD, d'atténuation des biais, de confidentialité et de souveraineté, etc.), permettant d'obtenir des résultats d'entraînement précis et cohérents. Lors de la conception des modèles ML, les ingénieurs se concentrent sur la sélection des algorithmes d'apprentissage les plus appropriés, l'élaboration d'architectures de modèles et leur maturation à travers des cycles itératifs de d'entrainement et d'évaluation. En outre, des stratégies d'optimisation telles que l'élagage (*pruning*), la quantification (*quantization*) et la distillation de modèles sont appliquées afin d'équilibrer l'efficacité computationnelle et les performances. Les activités de validation et de vérification (V&V) sont guidées par des propriétés clés telles que la stabilité, la généralisation, la robustesse et la reproductibilité des modèles ML. Ces « bonnes propriétés » sont spécifiées dans des exigences ML de bas niveau avec la définition de critères de satisfaction, de métriques et de scores de confiance afin de gérer l'incertitude liée aux données, à l'approximation des modèles et aux hypothèses statistiques. Les activités de validation sont principalement menées pour garantir l'exactitude et l'exhaustivité des exigences produites tout au long du cycle de développement ML en réalisant des revues et des analyses complétées par de la traçabilité vis-à-vis des exigences amont. Les activités de vérification comprennent des simulations approfondies, des tests nominaux et de robustesse des cas limites/extrêmes, des tests basés sur des scénarios, une analyse de l'explicabilité des modèles ML et une analyse de la couverture de l'ODD. Le 1er V se termine par la sélection d'un algorithme d'IA/ML qui répond à toutes les exigences dans l'environnement de développement (apprentissage) et sert de spécification de conception (détaillée), prête à être implémentée dans des logiciels et/ou du matériel électronique complexe (par exemple, FPGA) dans le 2ème V du cycle en W.

IA symbolique

Pour l'IA symbolique, il est essentiel de classer ses techniques en deux groupes distincts, chacun ayant des implications différentes pour le développement, la qualification et la certification. Le premier groupe comprend les techniques d'IA symbolique simples déjà intégrées dans des systèmes existants qui ont été qualifiées et certifiées selon les normes existantes. Il s'agit par exemple des systèmes décisionnels basés sur des règles, des cadres logiques déterministes et des systèmes experts simples tels que ceux utilisés dans la maintenance prédictive et la surveillance de l'état de santé des systèmes. Ces systèmes fonctionnent dans des limites bien définies et s'appuient sur des règles simples et interprétables, ce qui les rend adaptés aux méthodes de vérification et de validation conventionnelles.

Le deuxième groupe englobe des techniques d'IA symbolique plus complexes, conçues pour mettre en œuvre des fonctions complexes du système. Il s'agit par exemple des moteurs de raisonnement avancés, des ontologies complexes (multicouches) et de la programmation logique dynamique utilisés pour la planification adaptative, la prise de décision stratégique ou l'intégration de connaissances à grande échelle. Ces systèmes présentent souvent un comportement non linéaire et nécessitent un degré d'abstraction et d'interaction plus élevé, ce qui pose des défis aux cadres traditionnels de qualification et de certification. Par conséquent, ce second groupe de techniques d'IA symbolique nécessite un nouveau cadre de qualification et la certification inspiré de l'ingénierie des algorithmes d'IA du cycle en W présenté précédemment, afin de répondre à leur complexité et de garantir la sécurité, la fiabilité et la conformité dans le contexte particulier des systèmes critiques. Ainsi, pour ce deuxième groupe, le processus d'ingénierie est ancré dans un paradigme axé sur les connaissances, qui commence par la traduction des exigences système de haut niveau en exigences structurées pour les algorithmes d'IA. Cela implique l'élaboration d'architectures logiques et de descriptions détaillées de scénarios qui capturent de manière exhaustive le domaine du problème et les nuances opérationnelles. Les ingénieurs se concentrent sur la

sélection de la représentation des connaissances et de l'algorithme de raisonnement les plus appropriés pendant la phase de conception du modèle. Ainsi, pendant la phase de gestion des données/connaissances de l'IA, les efforts se concentrent sur l'organisation systématique des connaissances du domaine dans des bases de connaissances structurées, en utilisant une conception sophistiquée des schémas, des ontologies et des cadres hiérarchiques basés sur des concepts, des relations, des attributs, et des règles. Ces structures garantissent la modularité, l'accessibilité et l'évolutivité des tâches de raisonnement. La phase de conception du modèle d'IA symbolique implique le développement de règles d'inférence détaillées, la construction de structures logiques robustes et l'intégration de mécanismes avancés de représentation des connaissances adaptés aux besoins de l'application. Des outils tels que des solveurs symboliques, des arbres de décision et des algorithmes d'optimisation peuvent être déployés pour affiner ces modèles en termes de précision et d'efficacité. La validation et la vérification de l'IA symbolique s'appuient sur des simulations basées sur des scénarios de complexité gérable, combinées à des méthodes formelles et à la vérification des modèles afin de garantir la cohérence logique, l'exactitude et l'alignement avec les exigences prédéfinies.

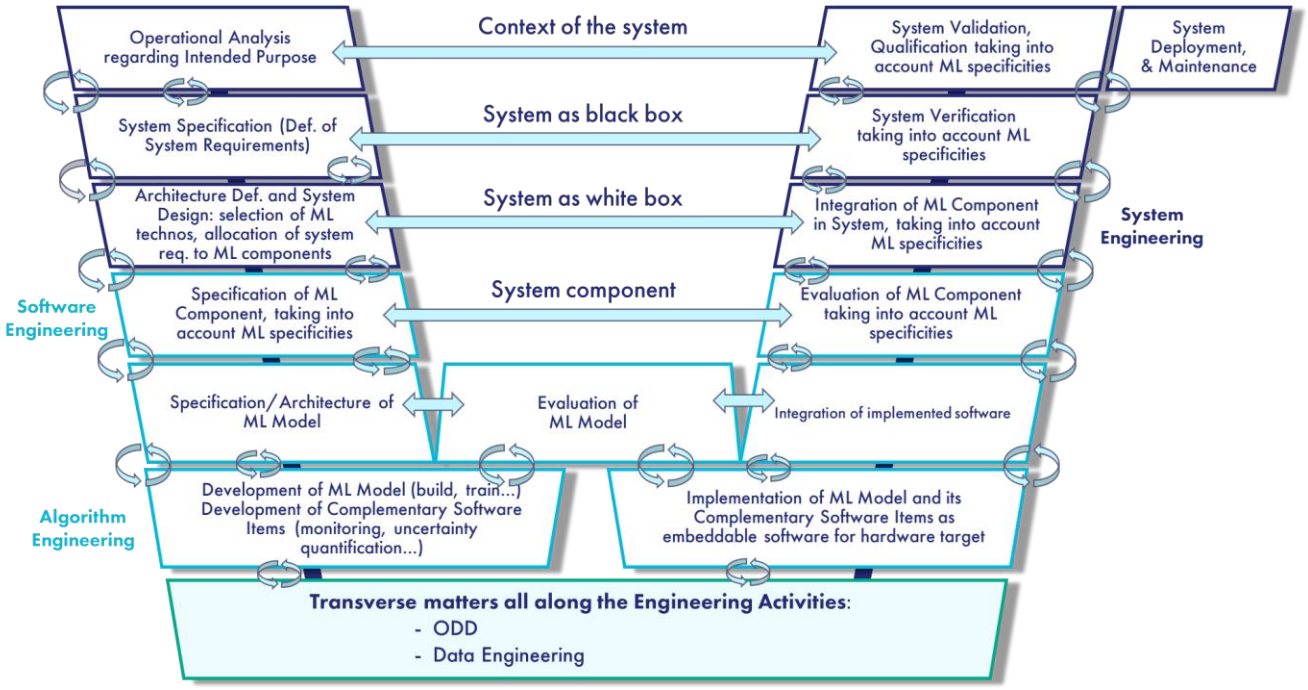

Figure 3 - Cycle de vie IA en forme de « W » pour l'apprentissage automatique (source: https://bok.confiance.ai)

IA hybride

L'IA hybride combine l'apprentissage automatique (ML) et l'IA symbolique pour créer des systèmes robustes, efficaces et explicables en tirant parti de leurs atouts complémentaires. Le ML pour le traitement des données non structurées, la reconnaissance de modèles et l'adaptation par l'apprentissage, et l'IA symbolique pour ses atouts en matière de représentation structurée des connaissances, de raisonnement et d'interprétabilité. Ainsi, les efforts déployés dans la phase de gestion des connaissances pour l'IA symbolique sont exploités pour améliorer les performances du modèle ML. Pour exploiter ces capacités complémentaires, un cycle de vie unifié de l'ingénierie de l'IA hybride intègre les méthodologies des deux paradigmes dans un cadre cohérent.

Ce cycle de vie pour l'IA hybride commence par le développement des exigences et la caractérisation du domaine opérationnel, en équilibrant les composants basés sur les données et ceux basés sur les connaissances (e.g., règles). Le ML s'appuie sur de grands ensembles de données pour définir le comportement dans le domaine opérationnel de conception (ODD), tandis que l'IA symbolique utilise des structures logiques et des règles d'inférence pour le raisonnement. Cette phase de spécification des systèmes d'IA hybride nécessite l'intégration de bases de connaissances avec des ensembles de données, permettant l'interaction entre le raisonnement symbolique et les processus d'apprentissage du ML.

Dans la phase de conception, les architectures des modèles ML s'alignent sur les mécanismes d'inférence symbolique. Les ingénieurs définissent des points d'intégration où les bases de connaissances (par exemple, les équations physiques ou les propriétés géométriques) ou le raisonnement symbolique (par exemple, les ontologies) informent le ML ou vice versa. Les flux de travail hybrides sont optimisés en affinant les modèles ML à l'aide de retours symboliques ou en utilisant le ML pour compléter les bases de connaissances symboliques. La validation et la vérification peuvent combiner la simulation et les tests de jeux de données (*test dataset*) avec des méthodes formelles afin de garantir la cohérence de l'IA symbolique, ce qui nécessite des stratégies innovantes pour tester les interactions hybrides dans divers scénarios.

2.3 MISE EN ŒUVRE LOGICIELLE/ MATÉRIELLE

La mise en œuvre d'algorithmes d'IA dans des systèmes critiques doit respecter les normes existantes (et à venir)[4], en tenant compte des spécificités des technologies d'IA.

Apprentissage automatique

Pour l'apprentissage automatique, la phase de mise en œuvre de l'algorithme d'IA implique le développement d'une architecture physique (éléments logiciels/matériels) de l'algorithme d'IA. En fonction de sa complexité, les ingénieurs peuvent être amenés à décomposer l'algorithme d'IA en plusieurs éléments, appelés « items », afin de respecter les contraintes de mise en œuvre imposées par les plateformes matérielles cibles, telles que les CPU, GPU, FPGA ou les plateformes embarquées spécifiques utilisant des accélérateurs comme le Thales Neural Processor (TNP). Des pipelines de tests automatisés sont utilisés pour vérifier la mise en œuvre de l'algorithme d'IA optimisé dans un ou plusieurs logiciels et/ou matériels électroniques complexes (CEH : *complex electronic hardware*). L'approche consiste à exécuter le ou les ensembles de données de test utilisés pendant la phase de conception (1er V), complétés par des cas de test spécifiques supplémentaires afin d'évaluer les propriétés de performance telles que le temps d'exécution dans le pire des cas (WCET : *Worst-Case Execution Time*), la latence, l'empreinte mémoire et la consommation d'énergie sur la ou les plateformes cibles. Cette phase de vérification supplémentaire, réalisée à la fin du 2e V (voir Figure 3), garantit qu'aucune erreur de mise en œuvre ou d'intégration (régressions), ni aucune incompatibilité avec la ou les plateformes cibles ne dégradent les performances souhaitées de l'algorithme d'IA.

IA symbolique

Pour l'IA symbolique, la mise en œuvre transforme les architectures logiques en architectures physiques, garantissant la compatibilité entre les ressources informatiques pour le raisonnement symbolique de l'IA et l'architecture et les ressources plus larges du système. Les processus de vérification doivent confirmer que les éléments d'IA symbolique répondent à leurs exigences, notamment en termes de performances, de robustesse, de sûreté et de sécurité. Ces processus comprennent la vérification de la précision des inférences, la garantie de la cohérence au sein des représentations des connaissances mises en œuvre et entre celles-ci, et le test de l'évolutivité dans des conditions opérationnelles dynamiques. En intégrant ces principes, les systèmes d'IA symbolique peuvent offrir des capacités de raisonnement fiables, même dans des environnements très complexes ou en évolution.

IA hybride

Dans les systèmes d'IA hybride, la mise en œuvre intègre des modèles d'apprentissage automatique et des composants de raisonnement symbolique, ce qui nécessite souvent du matériel spécialisé pour le traitement logique en plus des accélérateurs pour les modèles d'apprentissage automatique. Cette double approche nécessite une conception minutieuse des interfaces afin de garantir une intégration et une compatibilité parfaites entre ces deux technologies complémentaires. Les pratiques de programmation modulaire permettent à ces systèmes de fonctionner efficacement dans divers scénarios. Les activités de vérification comprennent celles mentionnées dans les sections précédentes consacrées à l'ingénierie des algorithmes dédiée au ML et à l'IA symbolique, complétées par des tests d'interface, des simulations basées sur des scénarios et des tests de résistance (*stress tests)*, garantissant que les systèmes d'IA hybride atteignent les objectifs combinés de performance, de robustesse, d'évolutivité et de conformité aux normes de sécurité et réglementaires.

2.4 POUR ALLER PLUS LOIN : LES DÉFIS INTERDISCIPLINAIRES

Cette section décrit brièvement certains des principaux défis interdisciplinaires, qui requièrent une attention

[4] **Normes existantes telles que ED-12C [9] et ses suppléments ou ED-80 [10] dans le domaine aéronautique ; EUROCAE WG114 / SAE G34 Intelligence artificielle dans l'aviation ED-324/ARP6983 Directives relatives au développement et à l'assurance des systèmes et équipements aéronautiques mis en œuvre avec l'IA**

particulière dans le contexte des systèmes critiques développés avec de l'IA.

Évaluer la Confiance envers l'IA

L'évaluation de la confiance envers l'IA est essentielle pour le développement et l'exploitation des systèmes critiques afin d'améliorer leur cycle de vie global, de la spécification des exigences à la maintenance en passant par la conception, l'implémentation, l'intégration, vérification, validation, qualification (IVVQ) et le déploiement. Pour rendre le système basé sur l'IA digne de confiance, il doit être évalué par des processus rigoureux. Cela implique d'utiliser des indicateurs appropriés basés sur des attributs de confiance et d'attribuer des scores à ces attributs de confiance.

En général, elle est déterminée par la quantification d'indicateurs élémentaires (par exemple, pour la fiabilité : score Fleiss Kappa, tests d'adéquation, ou pour la précision : précision, rappel, score F...). D'un point de vue métrologique strict, la fiabilité ne peut pas être mesurée, car il ne s'agit pas d'une propriété physique pouvant être comparée à une quantité de référence du même type. La fiabilité n'a pas d'unité. Le programme Confiance.ai [13] décrit différents attributs qui contribuent au concept de fiabilité, en examinant de près chaque attribut afin d'identifier les indicateurs clés de performance (KPI), les méthodes d'évaluation ou les points de contrôle, et en établissant une méthodologie d'agrégation pour ces attributs. Une autre question est celle de l'utilisation du système. Le contexte est important en matière de fiabilité. Il est façonné par une multitude de facteurs. Par exemple, un système critique basé sur l'IA dans le domaine de l'aéronautique peut avoir des exigences de fiabilité différentes de celles d'un système utilisé dans l'automobile ou les soins de santé. De plus, certains attributs peuvent être quantitatifs, comprenant généralement des valeurs numériques dérivées de mesures ou fournissant un aperçu statistique complet d'un phénomène. Il n'est donc pas simple de sélectionner les attributs pertinents pour évaluer la fiabilité de l'IA, car ce choix dépend de la facilité d'utilisation et du contexte d'application. Ce contexte est modélisé en fonction de plusieurs éléments, notamment le domaine de conception opérationnelle (ODD), le domaine d'utilisation prévu, la nature et les rôles des parties prenantes, etc.

L'assurance d'exécution pour compléter l'assurance de développement

En ingénierie des systèmes, l'assurance désigne les actions systématiques visant à garantir qu'un système répond à des exigences spécifiées. Ce processus structuré minimise les erreurs de développement contribuant à des conditions de défaillance potentielles, telles que des problèmes de sécurité ou de sûreté, grâce à une évaluation rigoureuse effectuée au moment de la conception, connue sous le nom de processus d'assurance du développement [8]. Après le déploiement, l'assurance d'exécution garantit que le système se comporte correctement pendant son exécution. Elle implique une surveillance en temps réel du comportement du système et des mesures correctives prédéfinies en cas d'écart par rapport au comportement souhaité. Cela est particulièrement crucial pour les systèmes critiques, où les défaillances peuvent avoir des conséquences graves.

Les éléments clés de l'assurance d'exécution pour les solutions basées sur l'IA comprennent, entre autres, la surveillance du comportement opérationnel du système afin de détecter des anomalies[5] potentielles et la vérification des assertions afin de vérifier les conditions du système pendant l'exécution. Les stratégies de récupération fournissent des réponses automatisées aux erreurs, allant de simples alertes au basculement vers des sauvegardes ou à la réduction des fonctionnalités à des fins de sécurité (paradigme de sécurité intégrée ou d'arrêt en cas de défaillance selon le domaine). Pour les systèmes fonctionnant dans des environnements distribués ou décentralisés, tels que les essaims de drones ou les réseaux de véhicules autonomes, l'assurance d'exécution doit s'étendre au-delà des composants individuels pour englober l'ensemble du collectif. Cela nécessite des mécanismes permettant de coordonner les mécanismes d'assurance d'exécution entre plusieurs entités, afin de garantir le maintien de la sécurité, de la sûreté et des performances au niveau de l'essaim ou du réseau. Ces approches impliquent une surveillance collaborative, une prise de décision synchronisée et une tolérance aux pannes distribuée afin de relever les défis uniques posés par ces systèmes interconnectés. Une autre considération importante est la conception de mécanismes d'assurance de l'exécution qui tiennent compte des contraintes du système en termes de ressources, telles que la puissance de calcul, la mémoire, l'énergie, la taille et le poids. Cela est essentiel pour garantir que les processus d'assurance ne surchargent pas la capacité du système à remplir ses fonctions critiques principales.

L'assurance d'exécution complète la vérification lors de la conception en traitant les problèmes imprévus lors du fonctionnement réel. Alors que les méthodes de conception permettent d'éviter les erreurs de

[5] Anomalie signifie « hors de la norme ».

développement telles que les exigences incorrectes ou les bogues de codage, l'assurance d'exécution ajoute une protection dynamique pour gérer les scénarios imprévus. Avec l'essor de l'IA et de l'autonomie, le recours à l'assurance d'exécution a posteriori joue déjà aujourd'hui, et continuera à jouer à l'avenir, un rôle plus important pour compenser les limites des activités d'assurance de développement a priori. À mesure que l'IA et les capacités autonomes se développent, l'importance de l'assurance d'exécution augmentera également, comme l'illustre la Figure 4.

Du DevOps et du MLOps à l'AIOps

Pour accompagner cette évolution de l'assurance, les pratiques d'ingénierie qui se limitaient jusqu'à présent à la conception devront s'étendre à l'exécution, à mesure que les systèmes évoluent au cours de leur cycle de vie. L'intégration des pratiques DevOps (*Development Operations*) et MLOps (*Machine Learning Operations*) est essentielle pour soutenir l'industrialisation de l'ingénierie des algorithmes d'IA dans ce contexte en évolution. DevOps fournit un cadre pour automatiser les workflows de développement et de déploiement de logiciels, garantissant des itérations plus rapides et une qualité constante. MLOps étend les principes de DevOps en répondant aux défis uniques des technologies ML, tels que la gestion des pipelines de données, l'orchestration de la formation des modèles et la surveillance des performances des modèles après leur déploiement. Ensemble, ces pratiques rationalisent le cycle de vie des algorithmes d'IA en permettant l'intégration continue et le déploiement continu (CI/CD), le contrôle des versions des modèles et des données, et les tests automatisés pour détecter rapidement les problèmes potentiels. Le MLOps garantit que les performances des modèles d'IA restent optimales en intégrant des workflows de réentraînement déclenchés par de nouvelles données ou des conditions opérationnelles changeantes, réduisant ainsi le risque de dérive des modèles. Parallèlement, DevOps complète cette approche en gérant l'infrastructure logicielle au sens large, garantissant ainsi que les mises à jour des modèles ML s'alignent parfaitement avec les autres éléments du système. La synergie entre DevOps et MLOps renforce la collaboration entre les data scientists, les ingénieurs ML, les ingénieurs logiciels et les équipes opérationnelles, ce qui améliore l'efficacité et réduit les délais de mise sur le marché des solutions d'IA. Thales s'efforce d'étendre cette approche holistique à l'IA symbolique et à l'IA hybride afin de construire un cadre général, appelé AIOps[6], qui est essentiel pour construire et maintenir des systèmes complexes impliquant différentes technologies d'IA, en particulier dans les applications critiques où une adaptation rapide et des performances robustes sont indispensables. Afin de fournir des systèmes d'IA robustes et à la pointe de la technologie, ce processus multidisciplinaire doit être outillé, dans les environnements de conception et de déploiement, afin de rationaliser la collaboration entre les différentes équipes et disciplines.

Approche incrémentale du développement et de la qualification

En adoptant le développement et la qualification incrémentaux, les domaines civil et militaire peuvent s'adapter aux changements opérationnels fréquents et rester sûrs, tout en maintenant leurs performances dans les applications critiques.

Les applications militaires, telles que la mise en œuvre d'une fonction de détection et de reconnaissance automatique des cibles (ATDR), exigent des mises à jour fréquentes et rapides des modèles d'IA afin de contrer les manœuvres de dissimulation en constante évolution de l'ennemi. Par exemple, du jour au lendemain, les ennemis peuvent camoufler leurs chars avec de la végétation, ce qui affecte les capacités de détection de l'IA et augmente le nombre de faux négatifs. Les dommages causés par les attaques militaires modifient également en permanence l'environnement du champ de bataille (par exemple, cratères, débris, poussière qui change la couleur des infrastructures et de la végétation).

[6] MLOps signifie « Machine Learning Operations » (opérations d'apprentissage automatique). Même si l'AIOps est souvent qualifié d'« intelligence artificielle pour les opérations informatiques », dans ce document, l'AIOps est l'extension et l'adaptation des principes DevOps à l'écosystème de l'IA, y compris l'IA statistique et connexionniste, l'IA symbolique, l'IA hybride, l'IA générative et l'IA agentielle.

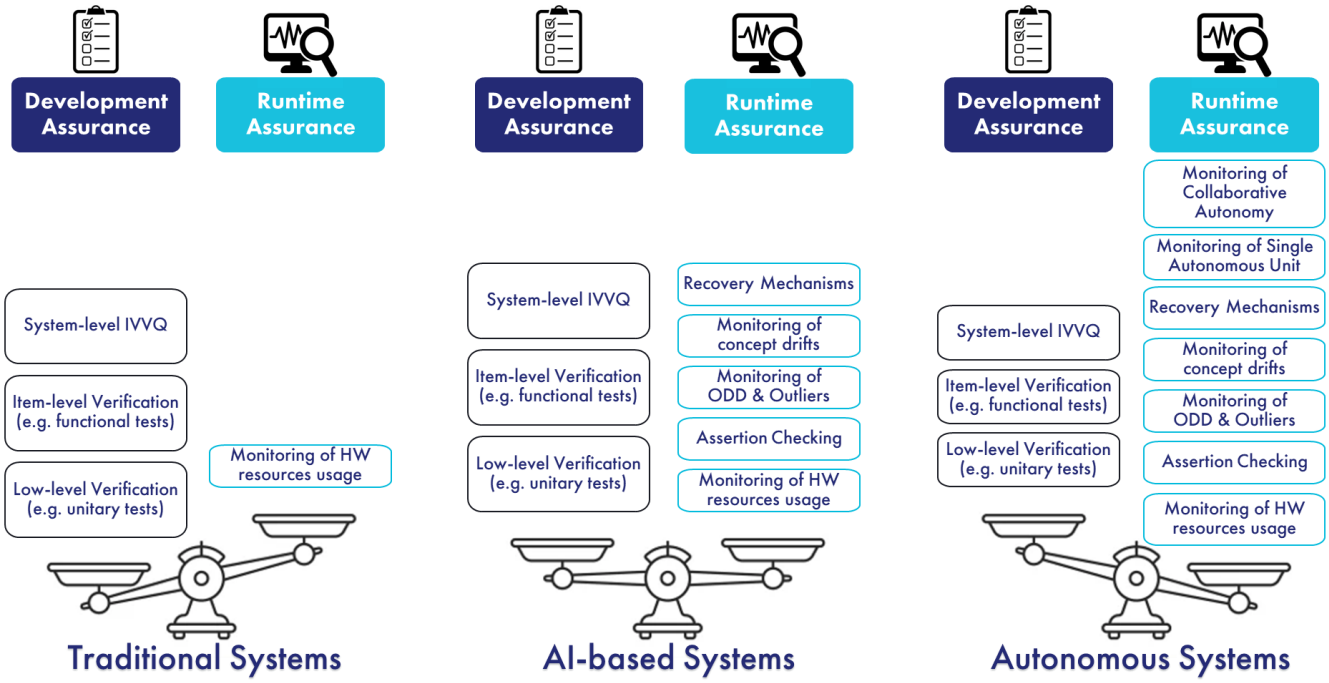

Figure 4 – Assurance du développement versus assurance d'exécution

Dans les applications civiles, telles que la mobilité urbaine et la gestion du trafic aérien, les modèles d'IA doivent s'adapter à des conditions dynamiques telles que les changements météorologiques ou les nouveaux obstacles, et doivent s'aligner sur la validité temporelle des données d'entraînement. Les drones de livraison, par exemple, peuvent être amenés à naviguer dans des zones où des travaux imprévus de construction ont démarré. Dans le domaine de la gestion du trafic aérien, les systèmes doivent intégrer des mises à jour en temps réel pour faire face aux retards de vol ou aux changements météorologiques imprévus, et donc maintenir des performances élevées sans interruption opérationnelle, en équilibrant les enjeux de sécurité et d'efficacité.

Cependant, d'un point de vue industriel, il n'est pas envisageable de mettre à jour rapidement les modèles d'IA par des modifications incrémentales efficaces avec une gestion du risque résiduel maitrisé en suivant, pour chaque incrément, l'ensemble du cycle de vie de développement (par exemple, le cycle de vie en W présenté plus haut) pour des raisons évidentes de temps et de coût de développement incompatibles avec les besoins et les contraintes opérationnels. À cet égard, les industriels doivent se donner pour priorité de mener des recherches sur de nouvelles méthodologies en collaboration avec les autorités responsables de l'homologation et de la certification des systèmes. Ces méthodologies visent à formaliser les compromis nécessaires entre le maintien d'un niveau suffisant de sûreté et de sécurité et la réalisation des objectifs stratégiques de la mission, et à évaluer le risque inhérent généré par ces compromis afin d'en rendre compte aux autorités compétentes, telles que le commandement militaire.

Quand l'IA accélère la transition des produits vers les services

Le développement de services IA destinés à compléter ou à remplacer les produits IA est une nécessité industrielle stratégique, car les modèles IA doivent être mis à jour fréquemment et rapidement afin de maintenir des niveaux élevés de performance et de sécurité. Les algorithmes d'IA, tels que les modèles d'apprentissage automatique, sont des systèmes intrinsèquement dynamiques qui dépendent fortement de la qualité, du volume et des caractéristiques des données à partir desquelles ils sont entraînés. Au fil du temps, des facteurs tels que les changements de comportement des utilisateurs ou la variabilité de l'environnement opérationnel (phénomène souvent appelé « dérive conceptuelle ou *concept drift* ») peuvent impacter fortement les performances des produits d'IA « figés », i.e., non mis à jour de façon régulière, ce qui entraîne inéluctablement une diminution de leur pertinence, de leur sûreté et de leur sécurité au fil du temps.

Les services d'IA, en revanche, adoptent une approche nativement fluide et itérative pour s'adapter aux évolutions des besoins de l'utilisateur final et de son environnement, ce qui consiste à surveiller, ré-entraîner et redéployer les modèles d'IA via des incréments plus ou moins grands, selon les besoins. Ce paradigme orienté services reconnaît que les systèmes basés sur l'IA ne sont pas des solutions qui peuvent être mises en place puis

oubliés (*set-and-forget*). Au contraire elles nécessitent une infrastructure et un environnement de production qui permettent de développer et qualifier de manière incrémentale, flexible et rapide. En tirant parti d'une architecture modulaire (par exemple sous forme de micro-services) et de pipelines de déploiement basés sur le cloud, y compris un cloud souverain pour les applications de défense, les services d'IA permettent des mises à jour transparentes sans perturber les opérations des utilisateurs finaux. Cela garantit que le système reste réactif et robuste aux changements du contexte d'entrée, qu'ils soient dus à des à facteurs internes tels que les changements de priorités opérationnelles ou à des facteurs externes tels que la dérive des données.

De plus, le paradigme commercial des services d'IA s'aligne sur la demande croissante de capacité de passage à l'échelle (scalabilité) et d'interopérabilité dans les applications d'IA. Les systèmes modernes basés sur l'IA fonctionnent souvent au sein d'écosystèmes où l'intégration avec d'autres services et plateformes est essentielle. La flexibilité des services d'IA facilite l'adaptation à de nouvelles exigences, telles que la conformité à des réglementations mises à jour ou l'intégration de nouvelles sources de données. De plus, le paradigme orienté services prend en charge une boucle de rétroaction, dans laquelle les mesures de performance et les retours d'expérience des utilisateurs alimentent directement un processus d'amélioration continu, favorisant ainsi l'évolution du système pour répondre aux besoins émergents.

D'un point de vue opérationnel, le paradigme de service favorise également la rentabilité et l'optimisation des ressources. En déployant les mises à jour de manière incrémentale et en se concentrant sur des modules spécifiques plutôt que sur l'ensemble du système, des organisations telles que Thales et ses principaux clients peuvent minimiser les temps d'arrêt et réduire les risques associés aux mises à niveau à grande échelle. Cette approche est particulièrement avantageuse dans les secteurs qui traitent des systèmes critiques, où les défaillances et les interruptions de service ne sont tout simplement pas tolérées.

L'IA pour l'ingénierie

Les technologies d'IA sont également de plus en plus reconnues pour leur potentiel à améliorer les activités d'ingénierie dans les différentes phases du cycle de développement, telles que l'ingénierie des exigences, le codage, et les tests. Les outils basés sur l'IA peuvent, par exemple, intervenir lors du développement et de la validation des exigences, mais il n'est pas recommandé d'utiliser l'IA dans ces deux activités simultanément. En effet, les outils d'IA peuvent analyser les sources d'information disponibles[7] et générer automatiquement des exigences structurées spécifiant les fonctions à développer. Ces exigences générées par l'outil d'IA doivent être validées par des humains ou tout autre moyen différent de l'outil d'IA utilisé afin d'éviter tout mode commun de défaillance. Pour la validation par l'IA des exigences développées par les ingénieurs, les outils d'IA peuvent recouper ces exigences avec les sources utilisées ou les données de projets précédents afin d'identifier des incohérences ou des lacunes de complétude. De même, les outils d'IA peuvent accélérer le développement du code source grâce à la génération automatisée de code qui traduit les spécifications de haut niveau en code source, à condition que ce code source généré soit revu par des développeurs logiciel. Pour la revue du code source écrit par des développeurs logiciel, les outils d'IA dotés de capacités de traitement du langage naturel et d'analyse statique peuvent identifier les vulnérabilités potentielles, les erreurs ou violations des normes de codage, améliorant ainsi considérablement la qualité et la sécurité des logiciels développés. Dans le domaine des tests, l'IA permet d'automatiser la génération de cas et de scénarios de test en exploitant des algorithmes avancés pour simuler les conditions opérationnelles nominales, identifier les cas aux limites et les cas particuliers, et garantir une couverture suffisante de l'ODD selon des critères à définir en fonction du cas d'étude.

Si ces progrès dans le domaine des outils d'IA apportent des avantages considérables, ils introduisent également le risque d'injecter et de propager des erreurs ou des biais inhérents aux outils d'IA eux-mêmes. Sans une surveillance et une vérification appropriée, ces outils pourraient compromettre involontairement la sécurité et la fiabilité des systèmes qu'ils sont censés contribuer à développer. Pour atténuer ces risques, il est essentiel de disposer d'une approche normalisée pour la qualification des outils d'IA, similaire au cadre rigoureux défini dans la norme EUROCAE ED-215 [11] pour les outils utilisés dans le développement de systèmes critiques. Cette nouvelle norme définirait des critères d'évaluation des outils d'IA, notamment leur précision, leur robustesse et leur fiabilité, ainsi que des méthodes de vérification de leurs résultats. En outre, elle établirait des exigences en matière de traçabilité et de transparence afin de garantir que toutes les décisions ou

[7] **Par exemple, au niveau du système, les contributions des parties prenantes peuvent inclure le concept d'opérations / d'utilisation (ConOps / ConUse), les normes réglementaires, les analyses de risques portant sur la sûreté, la sécurité et les facteurs humains.**

résultats générés par les outils d'IA puissent être vérifiés et compris par des ingénieurs humains.

L'adoption d'un cadre normalisé de qualification pour les outils d'IA est essentielle pour garantir que ces technologies améliorent les processus d'ingénierie sans introduire de risques inacceptables dans le développement d'applications critiques.

Gouvernance de l'IA

La gouvernance de l'IA englobe les règles, les processus et les pratiques qui garantissent le développement et le déploiement éthiques, sécurisés et conformes des systèmes d'IA pour ses clients. Cela comprend (sans s'y limiter) la définition des rôles et des responsabilités des différents acteurs impliqués dans la chaîne de valeur de l'IA et la garantie d'un cycle de vie rigoureux des données, depuis l'accès aux sources de données/connaissances appropriées jusqu'à l'archivage, la conservation et, si nécessaire, la suppression et l'élimination des données. Cette gouvernance de l'IA garantit également le respect des exigences éthiques et réglementaires (par exemple, dans l'Union européenne, la réglementation sur l'IA (*AI Act*), et le RGPD). Dans des domaines spécifiques tels que la défense, elle garantit que les processus et les pratiques préservent la confidentialité des données ainsi que les exigences en matière de souveraineté[8].

3. Point de vue de Thales

À travers ses systèmes critiques basés sur l'IA [12], Thales est convaincu que les modèles d'IA vont devenir de plus en plus complexes et seront chargés de mettre en œuvre des fonctions système de plus en plus sophistiquées (par exemple, opérations avec un seul pilote, détection et reconnaissance collaborative de cibles, etc.). Ainsi, l'apparition d'un niveau d'ingénierie spécifique consacré à l'ingénierie des algorithmes complexes est devenue indispensable [2]. Située entre l'ingénierie système traditionnelle et la mise en œuvre logicielle/matérielle, cette couche d'ingénierie algorithmique répond aux défis uniques posés par les applications modernes de l'IA [3][6]. Contrairement aux modèles précédents, qui étaient conçus pour exécuter des tâches techniques petites et bien définies, les systèmes d'IA actuels doivent gérer des fonctions multiples dans des conditions d'exploitation changeantes qui exigent des niveaux plus élevés d'abstraction, de flexibilité et d'intégration. L'initiative cortAIx de Thales est à la pointe de ces préoccupations en matière d'ingénierie IA de confiance de bout en bout.

[8] **Par exemple, au Royaume-Uni, le JSP 936 du ministère de la Défense britannique.**

4. Bibliographie

[1] Confiance.ai, Towards the engineering of trustworthy AI applications for critical systems, Second Edition (2024)

[2] EASA Artificial Intelligence (AI) Concept Paper Issue 2: Guidance for Level 1&2 machine learning applications, March 2024.

[3] EUROCAE WG114, Artificial Intelligence in Aviation Committee. AIR 6988, Artificial Intelligence in Aeronautical Systems: Statement of Concerns. SAE International (2021)

[4] Kaakai, F., Adibhatla, S., et al. Toward a machine learning development lifecycle for product certification and approval in aviation. SAE Int. J. Aerosp. 15 (2022)

[5] Kaakai, F., Adibhatla, S., Pai, G., Escorihuela, E. Data-Centric Operational Design Domain Characterization for Machine Learning-Based Aeronautical Products. SAFECOMP 2023. Springer, Lecture Notes in Computer Science, vol 14181 (2023)

[6] Delseny, H. et al. White Paper Machine Learning in Certified Systems. ArXiv abs/2103.10529 (2021)

[7] Kaakai, F. and Raffi, PM.. Towards Multi-timescale Online Monitoring of AI Models. SafeAI@AAAI (2023)

[8] EUROCAE ED-79B, Guidelines for Development of Civil Aircraft and Systems. SAE International (2023)

[9] EUROCAE ED-12C, Software considerations in airborne systems and equipment certification (2012)

[10] EUROCAE ED-80, Design Assurance Guidance for Airborne Electronic Hardware (2000)

[11] EUROCAE ED-215, Software Tool Qualification Considerations (2012)

[12] Mattioli, J., Meyer, C. (2024) Artificial Intelligence for critical system, Thales Position Paper.

[13] Mattioli, J., Sohier, H., Delaborde, A et al. (2024). An overview of key trustworthiness attributes and KPIs for trusted ML-based systems engineering. AI and Ethics, 4(1), 15-25.

[14] ARTIFICIAL INTELLIGENCE ROADMAP 2.0: Human-centric approach to AI in aviation, May 2023.

5. À propos de Thales

Thales est directement impliqué dans les défis et opportunités décrits dans ce document à travers ses activités variées. Depuis plusieurs années, l'intelligence artificielle est intégrée dans l'ensemble de son portefeuille de produits — en particulier dans les systèmes critiques et les solutions de défense déployées ou destinées à être déployées dans de nombreux pays à travers le monde. Cette expérience opérationnelle a permis aux équipes de recherche et d'ingénierie de Thales de développer des composants, des outils et des méthodologies IA adaptés à la fois aux contextes souverains et internationaux.

Certaines de ces innovations ont été mentionnées dans les sections précédentes pour aider l'écosystème à mieux comprendre les complexités de l'IA utilisée pour la guerre électromagnétique et les solutions pratiques qui peuvent être mises en œuvre.

Aujourd'hui, Thales rassemble plus de 800 experts en IA au sein de son organisation interne dédiée, cortAIx, créée en 2024. Ces experts sont répartis en trois équipes principales :

- Labs : axés sur la recherche fondamentale et le développement de modèles.
- Factory : responsables des chaînes d'outils, de la cybersécurité et des infrastructures critiques.
- Sensors : intégrant l'IA dans les systèmes embarqués et edge.

Ensemble, ces équipes garantissent que les capacités IA sont non seulement à la pointe de la technologie, mais aussi opérationnellement viables, sécurisées et conformes aux exigences de souveraineté. Les équipes cortAIx de Thales sont stratégiquement réparties entre la France, le Royaume-Uni, le Canada, Singapour, l'Allemagne et les Émirats arabes unis, reflétant l'empreinte mondiale de l'entreprise et son engagement à soutenir la souveraineté régionale grâce à une expertise locale.

---

cortAIx
Artificial Intelligence by THALES

THALES

Building a future we can all trust

4, rue de la Verrerie 92190
Meudon FRANCE

Tél. + 33(0)1 57 77 80 00

www.thalesgroup.com