IA de confiance : condition nécessaire pour le déploiement de l'IA dans les systèmes critiques

J Mattioli, F Terrier, L Cantat, R Gelin, J Chiaroni, Y Bonhomme, H Amadou-Boubacar, E Escorihuela, S Picard, C Alix

IA de confiance : condition nécessaire pour le déploiement de l'IA dans les systèmes critiques

J. MATTIOLI[8], F. TERRIER[3], L. CANTAT[4], R. GELIN[5], J. CHIARONI[7], Y. BONHOMME[4], H. AMADOU-BOUBACAR[1], E. ESCORIHUELA[2], S. PICARD[6], C. ALIX[8]

[1] Air Liquide, [2] Airbus, [3] CEA, [4]IRT SystemX, [5] Renault, [6] SAFRAN, [7] Secrétariat général pour l'investissement, [8]Thales

juliette.mattioli@thalesgroup.com

Résumé

Avec le renouveau de l'IA, on assiste aujourd'hui à une croissance de ses usages, sans précédent. Ce qui a changé ces dernières années, c'est que la recherche est passée de connaissances théoriques à de nombreuses applications pratiques. Malgré ces résultats très prometteurs trop peu de preuves de concept (PoC) atteignent un déploiement au niveau de la production des systèmes critiques. L'une des causes est que le déploiement dans des industries telles que l'aéronautique, l'énergie, l'automobile, la défense, la santé, la fabrication, etc. nécessite la conformité à des objectifs de qualité, de sûreté, de sécurité et de fiabilité qui ne sont pas complétés par des systèmes d'IA à l'état de l'art. Ainsi, un système critique à base d'IA doit reposer sur des méthodes de développement bien définies, de sa conception à son déploiement et sa qualification. Cela nécessite une chaîne d'outils de bout en bout garantissant la confiance à toutes les étapes : (1) la spécification, les connaissances et la gestion des données ; (2) conception d'algorithmes et d'architecture de système avec la préoccupation de la relation à l'humain ; (3) caractérisation, vérification et validation des fonctions de l'IA ; (4) déploiement, en particulier sur l'architecture embarquée ; (5) qualification, certification d'un point de vue système.

Mots-clés

IA, confiance, méthodologie, ingénierie algorithmique, ingénierie des données, ingénierie de la connaissance, ingénierie système, méthodes formelles, système critique.

Abstract

With the renewal of AI, we observe an unprecedented growth of its usage. What has changed is that research in recent years turns from theoretical insights into various practical applications. Despite these very promising results, too few Proof of Concept (PoC) are reaching production level deployment within critical systems. One of the causes is that deployment in industries as aeronautics, energy, automotive, defense, health, manufacturing, etc. requires conformity to quality, safety, security, reliability objectives that are from being completed by state-of-the-art AI systems. Thus, an AI based critical system needs to have well defined development methods from its design to its deployment and qualification. This requires a complete tool chain ensuring trust at all stages, as : (1) specification, knowledge and data management ; (2) algorithm and system architecture design taking into account human in the loop ; (3) characterization, verification and validation of AI functions ; (4) deployment, particularly on embedded architecture ; (5) qualification, certification from a system point of view.

Keywords

AI, Trust, Methodology, Algorithm Engineering, Data Engineering, Knowledge Engineering, System Engineering, Formal Methods, Critical Systems

1 Les enjeux de l'IA de confiance

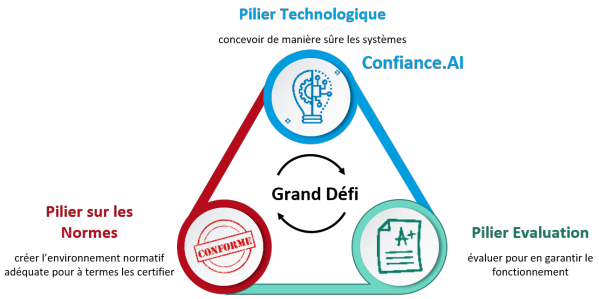

FIGURE 1 – Le Grand Défi de l'IA de confiance repose sur 3 piliers : le pilier technologique avec le programme Confiance.AI, le pilier évaluation et le pilier normalisation

Si l'Intelligence Artificielle (IA) semble promise à un fort développement, de nombreux verrous freinent son adoption, en particulier pour un déploiement dans les systèmes critiques. Ceux-ci doivent par construction garantir des propriétés de sécurité et de sûreté mais aussi suivre des principes de confiance et de responsabilité. En effet, par système critique, nous entendons, tout système pouvant directement ou indirectement engendrer des impacts sur les personnes physiques – impacts de nature corporelle, psychologique, sur la vie privée etc. – mais également sur les personnes morales – typiquement les entreprises avec des impacts de nature industrielle, financière, ou sur l'image au-

près du public, la propriété intellectuelle etc.

En parallèle de ce constat industriel, le conseil de l'innovation [1], a lancé fin 2018 le **Grand Défi National de l'IA de confiance**, avec l'objectif de sortir de l'ère des PoC (preuves de concept) par une réponse appropriée à la question de la qualification (homologation voire certification) des systèmes critiques à base d'IA.

En effet, que les systèmes critiques à base d'IA reposent sur des techniques d'apprentissage ou sur des approches plus symboliques, leur conception n'est pas neutre. Ces systèmes doivent suivre des principes de confiance et de responsabilité, garantir par construction (*by design*) des propriétés de sécurité, de sûreté et de fiabilité, qu'il faut pouvoir démontrer. Ainsi, pour permettre un déploiement de l'IA dans ces systèmes critiques, les pratiques d'ingénierie actuelles sont en défaut de par leur démarche basée sur la constitution de bases de données et de connaissances et leur exploitation via des algorithmes génériques qui masquent, voire abstraient, la logique fine des calculs. Il est alors extrêmement difficile de définir et comprendre leur enchaînement et donc d'établir la conformité des fonctions implantées. Il devient donc nécessaire de repenser ces ingénieries classiques (ingénierie algorithmique, ingénierie logicielle, ingénierie des données, ingénierie des connaissances, ingénierie système et ingénierie des facteurs humains) pour garantir la conformité du système vis à vis des concepts d'emploi, des besoins et des contraintes du client et des utilisateurs, de définir des méthodes et outils pour sécuriser l'ensemble des phases de conception mais aussi pour garantir des propriétés de type fiabilité, sécurité et cybersécurité, et de maintenabilité du système et cela tout au long de son cycle de vie. L'enjeu industriel est alors d'outiller de bout en bout toute cette démarche de « génie de l'IA » prenant en compte les dimensions algorithmiques, logicielles et systèmes. Pour cela, la construction des méthodologies de développement et la mise en place d'un environnement les outillant permettra de répondre aux défis posés par l'intégration de composants ou sous-systèmes d'IA dans des systèmes critiques mais aussi d'en accélérer le déploiement. C'est l'ambition du programme Confiance.ai.

L'objectif de cet article est d'identifier les enjeux induits par l'industrialisation de l'IA pour les systèmes critiques et de présenter les principales composantes d'un environnement de confiance pour les développeurs et intégrateurs, les auditeurs internes ou externes. Cet environnement est un atelier d'ingénieries support à la conception, à la validation et au test en vue d'intégrer de l'IA dans les systèmes critiques tout en répondant aux exigences réglementaires des régulateurs.

2 Un environnement de confiance

Comme mentionné précédemment, cet environnement de confiance est constitué d'un ensemble d'outils opérables et fédérés et de méthodologies qui peuvent être interfacés, notamment au sein d'ateliers d'ingénierie industriels.

Concernant l'état de l'art actuel et à venir à court terme, il existe :

- D'une part, une kyrielle d'initiatives développant des briques technologiques pour l'IA de confiance. Ces briques n'implantent que partiellement les fonctionnalités précédemment identifiées. Elles doivent donc être enrichies (voir fig. 2).
- D'autre part, il existe aujourd'hui un certain nombre d'environnements de développement d'IA existants. Ces environnements sont des environnements outillés mais ne traitant pas des problématiques particulières de la confiance.

Par exemple, le développement de logiciels à base d'IA dans de nombreux secteurs d'activité (automobile, aéronautique, ferroviaire, défense, énergie, naval, santé, etc.) pose des questions de garantie de bon fonctionnement dès lors que ces logiciels prennent des décisions de manière autonome dans un contexte critique. Il faut donc être capable d'assurer la transparence et l'auditabilité de ces systèmes pour pouvoir les comprendre et les corriger le cas échéant, mais aussi parfois aller jusqu'à la preuve formelle de leur bon fonctionnement et ceci de manière industrielle comme mentionné précédemment. Pour cela, il est nécessaire de disposer d'outils d'**ingénierie de la donnée et de la connaissance** afin de pouvoir respecter certains principes comme la loyauté et/ou l'équité lors des étapes de collecte, d'acquisition, d'analyse, de manipulation, de qualification des jeux de données d'apprentissage mais aussi de bases de connaissances. Puis, l'**ingénierie algorithmique** [15] doit être enrichie afin de prendre en compte les spécificités de l'IA de confiance, démontrant que les fonctions implantées sont correctes, prévisibles, stables, reproductibles, explicables, fiables et robustes. Cette ingénierie doit prendre en compte l'incertitude induite par la dynamique de l'environnement dans lequel le système évolue. Enfin, il faut être capable de détecter les erreurs sur un domaine d'emploi défini, et in fine si se pose la question de sa spécification et de sa validation. C'est pourquoi, l'intégralité du processus d'**ingénierie système** doit être outillée.

Les fonctionnalités de l'environnement doivent supporter l'ensemble des usages pressentis (cf. fig. 2) regroupant :

- Un coeur de fonctionnalités pour aider à la conception, à la qualification au déploiement et à la maintenance des composants et des systèmes à base d'IA.
- Un ensemble de fonctionnalités proposant des méthodes outillées.
- Un ensemble de fonctionnalités transverses.
- Un ensemble de fonctionnalités permettant de s'interfacer.

Le coeur de ces fonctionnalités se décompose alors en deux sous-ensembles :

- Le premier pour aider à la conception, à la qualification, au déploiement et à la maintenance des composants d'IA : point de vue de l'ingénierie algorithmique de l'IA.
- Le second pour aider à la conception, à la qualifi-

1. Composé de 6 ministres, des administrations concernées (SGPI, DGE, DGRI), de deux opérateurs (ANR et Bpifrance) ainsi que de 6 personnalités reconnues, ce Conseil fixe les priorités stratégiques de la politique d'innovation française.

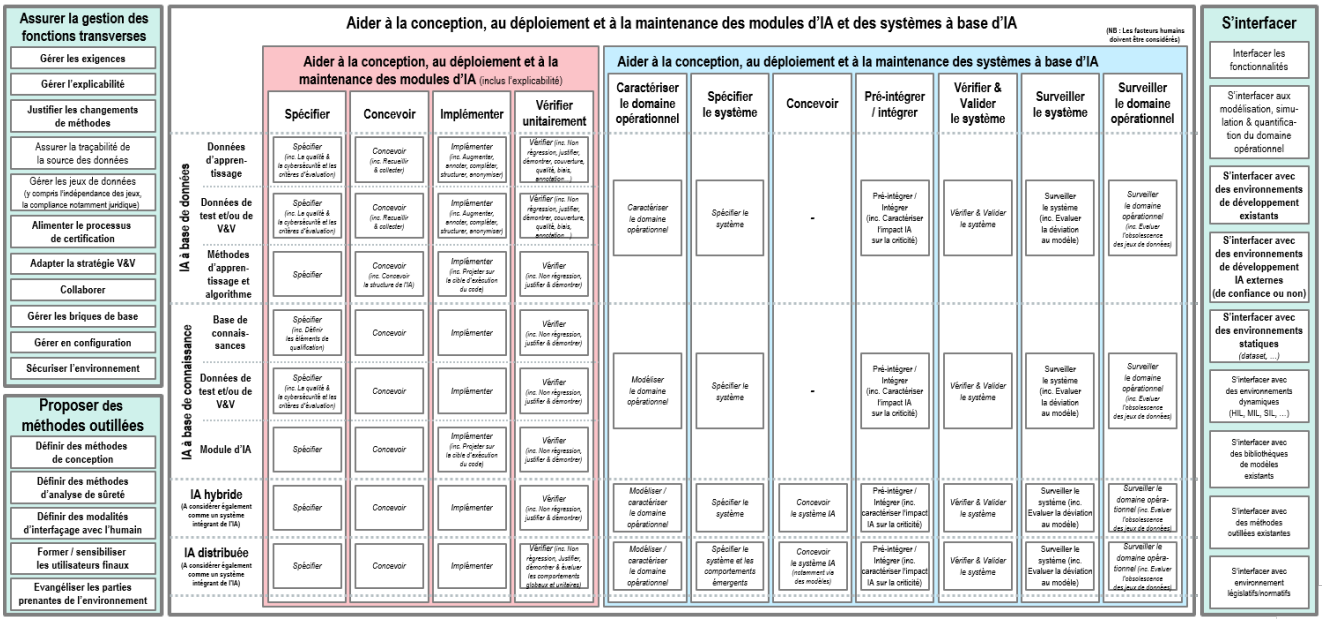

FIGURE 2 – La matrice de fonctionnalités de l'ingénierie de l'IA de confiance pour un déploiement opérationnel [Source Confiance.ai]

cation, au déploiement et à la maintenance des systèmes à base d'IA : point de vue de l'ingénierie système de l'IA.

Chacun de ces sous-ensembles est subdivisé en fonction du cycle de développement, soit :

- Pour la sous partie traitant des composants d'IA, nous trouvons les étapes suivantes : Spécifier, Définir le domaine opérationnel, Concevoir, Implémenter, Vérifier de manière unitaire, incluant bien sûr, en fonction des paradigmes de l'IA choisis, l'ingénierie des données et l'ingénierie des connaissances.
- Pour la sous partie traitant des systèmes à base d'IA (vue ingénierie système) nous trouvons les étapes suivantes : Caractériser le domaine opérationnel, Spécifier le système, Concevoir, Pré-intégrer et intégrer, Vérifier et Valider le système, Surveiller le système, Surveiller le domaine opérationnel, et Garantir son maintien en condition opérationnel.

L'ensemble des fonctionnalités proposant des méthodes outillées est alors constitué des éléments suivants :

- Définir des méthodes de conception,
- Définir des méthodes d'analyse de sureté,
- Définir des modalités d'interfaçage avec l'humain,
- Former / sensibiliser les utilisateurs finaux,
- Evangéliser les parties prenantes de l'environnement.

Il convient de préciser que les 3 premières fonctionnalités sont transverses et seront donc utilisées dans l'ensemble des autres fonctionnalités de l'environnement de confiance, puisqu'elles constituent le sous-jacent méthodologique.

Des fonctionnalités transverses permettent de :

- Gérer les exigences,
- Gérer l'explicabilité,
- Justifier les changements de méthode – Passage à une méthode basée sur de l'IA par opposition à une méthode traditionnelle,
- Assurer la traçabilité de la source de donnée,
- Gérer les jeux de données,
- Alimenter le processus de certification,
- Adapter la stratégie V& V,
- Collaborer – entre plusieurs acteurs via l'environnement,
- Gérer les briques de base – briques de composant d'IA dite de confiance
- Gérer en configuration – toute ou partie de l'environnement et des objets manipulés,
- Sécuriser l'environnement.

Enfin, cet environnement doit permettre de garantir le respect des normes et de la réglementation, comme par exemple, la conformité au RGPD via des approches respectant la vie privée (*Privacy by design*) ou avec la loi pour une République numérique qui induit transparence et loyauté. De plus, dans un contexte de montée de l'autonomie de certaines fonctions, ces ingénieries doivent être repensées pour prendre en compte les contraintes d'**embarquabilité** et la **relation humain-système**.

3 Ingénierie algorithmique de l'IA de confiance

Historiquement, la conception d'algorithmes d'IA émerge dans les années 1950 au travers de deux courants. L'**IA à base de connaissances**, qualifiée aujourd'hui de GOFAI (*Good Old Fashioned AI*) ou d'IA symbolique, se base quasi exclusivement sur le raisonnement symbolique et la logique. Elle se distingue de l'**IA dirigée par les données**, appelée aussi IA statistique et connexionniste, sous les feux de la rampe ces dernières d'années avec la collecte massive des données et l'arrivée de l'IA subsymbolique (et du deep learning), bien qu'aussi ancienne. Ainsi, l'IA symbolique

utilise des connaissances transmises à la machine pour résoudre des problèmes et l'IA dirigée par les données part d'exemples de solutions qu'elle essaie d'extrapoler par des méthodes statistiques. Leurs domaines d'emploi diffèrent. Alors que l'IA connexioniste est l'IA des sens, l'IA symbolique est celle du sens.
Plusieurs travaux cherchent à hybrider ces deux paradigmes, comme le souligne N. Asher [2] : "*L'addition de ces deux courants, IA symbolique et IA connexionniste, constitue le défi d'aujourd'hui*". Par exemple, l'apprentissage par renforcement consiste à récompenser les comportements souhaités et/ou à sanctionner les comportements non désirés avec des stratégies de récompense ou de sanction basées sur des connaissances métiers ou heuristiques issues de l'IA symbolique.

3.1 Conception algorithmique

Pour garantir une conception algorithmique de confiance (robuste, fiable...), l'ingénierie algorithmique (*Algorithm Engineering*) définit par P. Sanders [12, 16], doit intégrer les paradigmes induits par l'IA ainsi que les dimensions de (cyber)-sécurité et l'humain dans la boucle.
De plus, la sûreté et la sécurité des systèmes critiques à base d'IA nécessitent, de démontrer que les algorithmes sont corrects, c'est-à-dire qu'ils font ce qu'on attend d'eux. Il est donc nécessaire de vérifier la conformité entre leurs spécifications et leur comportement, autrement dit l'écart entre ce qu'il est supposé faire et ce qu'il fait réellement. Certaines approches en IA symbolique comme la programmation par contraintes offrent, par construction, cette propriété de correction, mais il reste nécessaire de la démontrer dans les autres cas comme pour l'IA connexionniste.
La robustesse d'un algorithme caractérise son aptitude à fournir des réponses correctes face à des situations inconnues ou à des malveillances. Cependant, cette propriété est plus dure que la précision. En effet, un système non précis ne peut être robuste. Mais surtout, un système précis peut ne pas être robuste. C'est le cas d'un système à base d'apprentissage ayant appris par coeur les données d'apprentissage qui se trompera dans ses décisions futures basées sur de nouvelles données. Ce phénomène est appelé *overfitting* (sur-apprentissage). De plus, l'IA reste vulnérable, et si l'on n'y prend pas garde, particulièrement sensible aux attaques dites "adversarial" (contradictoire), attaques qui tirent parti du fonctionnement des algorithmes sous-jacents pour générer des perturbations de faible amplitude dans les données analysées et et force l'IA à renvoyer un résultat incorrect. Heureusement, l'existence d'attaques 'adversarial' induit l'existence de défenses. De nombreuses défenses ont été proposées ces dernières années par la communauté scientifique [2] mais qui sont parfois réfutées avec de nouvelles attaques les rendant obsolètes. C'est pourquoi, il faut de se doter de méthodes et outils pour concevoir des algorithmes robustes et a minima caractériser leur robustesse.
Il est aussi nécessaire de prouver que les systèmes critiques sont contrôlables, c'est-à-dire qu'ils sont bien-fondés ou cohérents (on emploie aussi l'anglicisme consistant), si l'on peut prouver qu'ils ne font que ce qu'on l'attend d'eux. Les questions relatives aux problèmes de robustesse et de consistance commencent à faire l'objet de travaux liés aux preuves formelles. Ces dernières visent à apporter des garanties a priori sur la sûreté de fonctionnement d'un programme, contrairement aux méthodologies de validation par expérimentations directes qui visent à apporter des garanties a posteriori. Enfin, la compréhension de l'IA et de son raisonnement est nécessaire pour déterminer à quel point nous pouvons lui faire confiance. Un avantage des approches symboliques est de permettre de tracer le raisonnement. Mais même dans ce cas, il est nécessaire pour les usagers d'avoir une explication intelligible (explicabilité) plus que la traçabilité du raisonnement. Par contre, les approches connexionnistes s'apparentent aujourd'hui à des boîtes noires dont la complexité et l'abstraction qui sous-tendent ses décisions nous éloignent davantage de cette compréhension. Il devient alors nécessaire d'offrir des méthodes et outils pour rendre l'IA plus transparente, ouvrir les boîtes noires afin de comprendre comment un résultat a été atteint.

3.2 Vérification, validation et qualification

Lors de la vérification, la validation et la qualification du bon fonctionnement d'un algorithme d'IA, les situations suivantes doivent être abordées [11] :

- Le cas des composants livrés en boîte noire sur lesquels on cherchera principalement à en évaluer la robustesse. Par exemple, des approches ont été proposées dans la littérature [21, 20] présentant des méthodes pour étudier la robustesse de réseaux de neurones sur des problèmes de classification.
- Lorsque le composant est en boîte blanche (accès aux détails de sa structure, configuration, code source), il est alors possible de réaliser une analyse fine à l'aide de méthodes formelles (interprétation abstraite [5], Satisfiabilité modulo théories [8], programmation linéaire, etc.), mathématiques de ses comportements possibles. Un enjeu majeur restant la capacité de formaliser les propriétés de sûreté attendues afin que donner un sens fort aux preuves développées [6]. Cela permet, par exemple, de mettre en place des stratégies de test de robustesse face aux attaques adverses dans le cas d'approches à base d'apprentissage [13]. Il est également possible d'aller plus loin dans la caractérisation en définissant des domaines de stabilité.

Une voie prometteuse d'évaluation de la robustesse consiste à utiliser des approches de randomisation, à l'aide de bruits ajoutés de manière contrôlés à l'entrée du processus de décision, permettant de conduire à des notions de certificats statistiques de robustesse [3].

2. Nicolas Asher chercheur CNRS à l'Institut de recherche en informatique de Toulouse (IRIT) est le directeur scientifique du 3IA ANITI

4 Ingénierie des données et des connaissances

Pour les approches statistiques et connexionnistes, les données sont donc cruciales pour l'apprentissage, le test et la validation. Il ne suffit pas d'avoir beaucoup de données, il faut qu'elles soient de "bonne qualité" et représentatives du domaine d'emploi du système concerné, sans quoi ces approches donnent de mauvais résultats. De même, en IA symbolique, l'exploitation de connaissances de mauvaise qualité conduit à des résultats médiocres voire des erreurs. Il est nécessaire de repenser l'ingénierie des données et l'ingénierie des connaissances au regard de ces exigences.

De nouvelles méthodologies sont à définir pour une meilleure maîtrise des étapes d'acquisition, d'exploration, d'enrichissement, d'annotation et de préparation des données. Par exemple, la décomposition du jeu de données en plusieurs sous-ensembles dédiés aux différentes phase d'apprentissage, de validation et de test, doit respecter la représentativité du jeu de données pour permettre une bonne inférence en adéquation avec le domaine d'emploi.

De plus, comme les performances sont évaluées statistiquement sur un jeu de test préalablement constitué, la fiabilité de l'indice de performance est étroitement liée à la représentativité de ce jeu. La difficulté de cette décomposition réside dans la contrainte de constituer des ensembles distincts tout en garantissant qu'ils préservent des distributions comparables. De plus, il est nécessaire d'identifier automatiquement les situations qui mettent les systèmes en échec critique, et en retrouver le plus grand nombre possible parmi les données déjà acquises est nécessaire et difficile. Les techniques d'apprentissage actif (aussi appelé *machine teaching* [10]) n'y suffisent pas.

L'enrichissement permet de pallier la rareté des données. Cela consiste à ajouter artificiellement certaines données dans le jeu d'apprentissage ou de validation. Allant au-delà de la simple identification ou sélection intelligente d'outils, les techniques suivantes permettent d'augmenter la robustesse des modèles appris, ou de tester la robustesse lors de phase de validation :

- La génération artificielle de cas limites à base de réseaux neuronaux génératifs, pour créer de façon plausible de telles situations. Il sera par exemple possible de produire (et annoter) des situations rarissimes.
- L'utilisation de données réelles peut s'avérer complexe et le recours à des données synthétiques obtenues avec des simulateurs constitue une alternative intéressante.
- La création de nouvelles données à partir des données existantes, en appliquant par exemple, dans le cas de classification d'images, des transformations géométriques sur les images d'origine.

Mais aujourd'hui, les techniques d'apprentissage les plus efficaces sont supervisées reposant donc sur des annotations. La production d'annotations fiables est donc incontournable, puisque l'algorithme va ajuster ses paramètres afin d'associer une donnée d'entrée avec l'annotation cible. Cette phase a fait l'objet de nombreux travaux comme l'apprentissage actif ou l'automatisation de l'annotation par la création de fonctions d'annotation (supervision faible). De plus, caractériser la qualité d'un jeu de données n'est pas aisé. Il existe une pléthore de dimensions [14] qu'il faut choisir au regard d'un contexte décisionnel particulier. Même s'il existe très peu de normes relatives à la qualité des données [3], la question de la qualité de la donnée (*Data Quality* [7, 19]) n'est pas nouvelle : meilleure sera la qualité de la donnée, plus pertinente sera la décision. Dans son programme "*Total Data Quality Management*" (TDQM), le MIT s'attaque à cette question depuis le début des années 1990 [17], identifiant ainsi de nombreuses dimensions telles que la précision, la pertinence, la couverture, la complétude, la crédibilité, la cohérence, la fiabilité...

Enfin, si les données qui nourrissent les algorithmes à base d'apprentissage sont biaisées, les décisions que ceux-ci prendront le seront également. L'identification de ces biais peut poser question car l'une des difficultés consiste à comprendre comment un modèle généralise l'apprentissage qu'il a effectué à partir des données d'entraînement. C'est pourquoi, l'ingénierie des données doit être outillées pour permettre en particulier d'identifier les biais d'échantillonnage, d'enregistrement, de nettoyage, d'exclusion [4], induits par les transformations d'ingénierie des caractéristiques (*Feature Engineering*), voire de confirmation [5].

Les systèmes à base de connaissances, quant à eux peuvent représenter et traiter des principes et des règles de décisions, des taxonomies, des théories, des processus et des méthodes mémorisées dans un système artificiel. Mais pour concevoir un système à base de connaissances ayant un comportement compréhensible et acceptable par l'utilisateur passe par une modélisation à un niveau d'abstraction pertinent qui fait sens pour les différents acteurs impliqués dans sa conception (experts métiers, utilisateurs, etc.). En phase d'utilisation du système, le modèle est rendu opérationnel de manière à ce que l'utilisateur s'approprie le comportement du système et puisse interagir avec lui. L'ingénierie des connaissances (IC) propose des concepts, méthodes et techniques permettant de modéliser et/ou d'acquérir les connaissances dans des domaines ou la formalisation est difficile ou la compréhension des phénomènes partielle. L'IC peut être schématiquement définie par trois étapes : l'acquisition de connaissances disponibles, leur représentation informatique et l'utilisation de celles-ci à des fins de simulation, de prédiction, de validation, d'optimisation pour aider à la décision [18]. Rappelons que l'extraction des connaissances couvre le processus permettant de transformer les connaissances des experts dans un domaine sous forme d'informations organisées, alors que l'acquisi-

3. norme ISO 8000 relative à la qualité des données de référence – Master data

4. le biais d'exclusion provient de données qui sont retirées de manière inappropriée de la source de données.

5. le biais de confirmation est le désir de sélectionner uniquement les informations qui soutiennent ou confirment quelque chose que vous connaissez déjà, plutôt que des données qui pourraient suggérer quelque chose qui va à l'encontre d'idées préconçues.

tion des connaissances est le processus inverse qui consiste à transformer, par l'apprentissage, les informations et les savoirs disponibles en connaissances. Là encore, il est important de se doter d'une démarche méthodologique pour garantir la complétude, à la pertinence et à la qualité des modèles.

5 Evaluation de la qualité de l'algorithme

Évaluer les performances d'une IA dirigée par les données, consiste à évaluer la qualité d'une fonction, apprise selon des principes d'apprentissage statistique, lorsqu'elle sera déployée. Si la théorie donne un cadre clair à l'évaluation du risque théorique, sa mise en pratique implique de définir la notion de risque empirique qui s'appuie sur deux concepts : d'une part la distribution réelle des données n'est pas connue, elle est remplacée par un ensemble de données, ou une distribution approchée ; d'autre part elle repose sur la définition d'une fonction de coût, qui doit au mieux retranscrire l'intention finale. Dans le cadre strict de l'évaluation des performances, les deux problèmes principaux sont donc : 1) comment choisir la bonne métrique d'évaluation ; 2) quelle méthodologie pour l'estimation robuste de cette métrique de performance. Dans ce cadre, un guide d'évaluation a été rédigé par la DGA [11] pour les approches d'apprentissage supervisé. À ces deux problèmes issus de la nature intrinsèque de l'apprentissage statistique, il faut ajouter la question de la reproductibilité des performances rapportées, vis-à-vis de paramètres considérés jusqu'ici comme mineurs. Il faut aussi noter que de nombreux projets s'attellent à la question de l'évaluation ou de l'explication. Citons les travaux issus du programme DEEL (France et Canada) pour le cadre IA des données ou du "GT Explicabilité du GDR IA" pour l'IA symbolique.

Enfin, le changement radical des pratiques de développement des systèmes à base d'IA et la complexité induite pour leur validation, amènent à envisager l'introduction d'approches de qualification et de certification plus souples pour faire face aux différents types d'incertitude que présentent ces systèmes. Outre la définition de référentiels de risques spécifiques liés à l'IA, deux approches de la qualification semblent particulièrement intéressantes : (1) la qualification basée sur des propriétés globales du système [4, 1], offrant plus de souplesse dans la manière de gérer la complexité et l'implantation des pratiques de qualification ("*Assurance Case*" qualification based on "system overarching properties" satisfaction) ; et (2) la qualification modulaire, incrémentale et évolutive, par exemple via des approches par contrat, permettant de prendre en compte l'évolution nécessaire des systèmes liées aux évolutions des données, connaissances et de l'environnement qui risquent d'être beaucoup plus rapides pour l'IA.

6 Conclusion

Sécuriser, certifier et fiabiliser les systèmes qui ont recours à l'intelligence artificielle posent des questions d'ingénierie algorithmique, d'ingénierie des données et des connaissances, d'ingénierie système, de la sûreté et de la (cyber)-sécurité, mais aussi d'ingénierie des facteurs humains, dès lors que ces logiciels prennent des décisions de manière autonome dans un contexte critique. Le programme "Confiance.ai" du Grand Défi national a pour objectif de définir et d'outiller une approche rigoureuse et interdisciplinaire en formalisant l'ensemble du cycle de vie de ces systèmes à base d'IA de confiance [9].

Les besoins principaux auxquels l'environnement devra répondre, in fine, sont les suivantes :

- Disposer de méthodes et d'outils de gestion des données et des connaissances : conception, d'analyse, manipulation, collecte, acquisition, qualification, génération, filtrage des jeux de données d'apprentissage et base de connaissances pour la validation des systèmes cibles.
- Capacité à produire (concevoir, valider, implanter) un algorithme d'intelligence artificielle dit de confiance : correct, prévisible, stable, reproductible, explicable, fiable, robuste, capable de détecter les erreurs sur un domaine d'emploi défini et maîtrisé et donc in fine et si nécessaire certifiable.
- Capacité à définir et outiller l'intégralité du processus de développement, d'intégration et de qualification/certification sur l'ensemble du cycle de vie des systèmes intégrant de l'IA en interopérabilité avec les autres environnements de conception.
- Sortir d'une approche basée uniquement sur les preuves de concepts et passer à l'échelle industrielle en revisitant et repensant la chaîne d'ingénierie de l'algorithme, du logiciel et du système ainsi que la prise en compte du hardware pour le développement de composants à base d'IA.

Pour appuyer la spécification des besoins, valider et caractériser les solutions, le programme s'appuie sur un ensemble de cas d'usages ciblés et partagés par l'ensemble des acteurs. On peut citer notamment :

- Compréhension de scène pour la mobilité autonome à partir d'un capteur caméra 2D ;
- Surveillance et détection de déviation de l'efficacité opérationnelle d'une usine ;
- Contrôle embarqué par réseaux de neurones pour une fonction d'anticollision en vol ;
- Détection de conformité de cordons de soudure par inspection visuelle ;
- Maintenance prédictive des propulseurs de navire.

Des cas d'usage complémentaires seront intégrés au cours du programme afin d'aborder les volets techniques complémentaires, notamment autour de l'IA à base de connaissances et hybride.

Remerciements

Les partenaires de Confiance.ai sont par ordre alphabétique : Airbus, Air Liquide, ATOS, CEA, Inria, IRT SystemX, IRT St Exupéry, Naval Group, Renault, Safran, Sopra Steria, Thales et Valéo

Références

[1] Darpa program assured autonomy, https ://www.darpa.mil/program/assured-autonomy.

[2] A. Araujo, L. Meunier, R. Pinot, and B. Negrevergne. Robust neural networks using randomized adversarial training. *arXiv preprint arXiv :1903.10219*, 2019.

[3] J. Cohen, E. Rosenfeld, and JZ. Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv :1902.02918*, 2019.

[4] E. Denney, G. Pai, and I. Habli. Dynamic safety cases for through-life safety assurance. In *IEEE/ACM 37th IEEE Int. Conf. on Software Engineering*, volume 2, pages 587–590. IEEE, 2015.

[5] T. Gehr, M. Mirman, D. Drachsler-Cohen, et al. Ai2 : Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2018.

[6] J. Girard-Satabin, G. Charpiat, Z. Chihani, and M. Schoenauer. CAMUS : A framework to build formal specifications for deep perception systems using simulators. In *24th European Conference on Artificial Intelligence (ECAI)*, 2020.

[7] Y. Huh, F. Keller, Thomas C. Redman, and A. Watkins. Data quality. *Information and software technology*, 32(8) :559–565, 1990.

[8] G. Katz, D. Huang, D. Ibeling, K. Julian, et al. The marabou framework for verification and analysis of deep neural networks. In *Int. Conf. on Computer Aided Verification*, pages 443–452. Springer, 2019.

[9] J. Mattioli, F. Terrier, L. Cantat, J. Chiaroni, M. Barreteau, Y. Bonhomme, C. Guettier, and C. Alix. Ia de confiance : condition nécessaire pour le déploiement de l'ia dans les systèmes de défense - hal id : hal-02955575, 2020.

[10] S. Mei and X. Zhu. Using machine teaching to identify optimal training-set attacks on machine learners. In *AAAI Conf. on Artificial Intelligence*, 2015.

[11] DGA (French MoD). Guide méthodologique pour la spécification et la qualification des systèmes intégrant des modules d'intelligence artificielle - version 1.0 b, 2019.

[12] M. Muller-Hannemann and S. Schirra. *Algorithm engineering : bridging the gap between algorithm theory and practice*. Springer-Verlag, 2010.

[13] MI. Nicolae, M. Sinn, MN. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, et al. Adversarial robustness toolbox v1. 0.0. *arXiv preprint arXiv :1807.01069*, 2018.

[14] L. Pipino, Y. Lee, and R. Wang. Data quality assessment. *Communications of the ACM*, 45(4) :211–218, 2002.

[15] P. Sanders. Algorithm engineering–an attempt at a definition. In *Efficient Algorithms*, pages 321–340. Springer, 2009.

[16] P. Sanders. Algorithm engineering–an attempt at a definition using sorting as an example. In *12th Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 55–61. SIAM, 2010.

[17] F. Sidi, P. Panahy, L. Affendey, M. Jabar, H. Ibrahim, and A. Mustapha. Data quality : A survey of data quality dimensions. In *2012 Int. Conf. on Information Retrieval & Knowledge Management*, pages 300–304. IEEE, 2012.

[18] R. Studer, VR. Benjamins, and D. Fensel. Knowledge engineering : principles and methods. *Data & knowledge engineering*, 25(1-2) :161–197, 1998.

[19] R.Y Wang and D.M Strong. Beyond accuracy : What data quality means to data consumers. *Journal of management information systems*, 12(4) :5–33, 1996.

[20] L. Weng, PY. Chen, L. Nguyen, M. Squillante, A. Boopathy, I. Oseledets, and L. Daniel. Proven : Verifying robustness of neural networks with a probabilistic approach. In *Int. Conf. on Machine Learning*, pages 6727–6736, 2019.

[21] H. Zhang, TW. Weng, PY. Chen, and others. Efficient neural network robustness certification with general activation functions. In *Advances in neural information processing systems*, pages 4939–4948, 2018.