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

Juliette Mattioli, François Terrier, Loic Cantat, Julien Chiaroni, Michel Barreteau, Yannick Bonhomme, Christophe Guettier, Christophe Alix

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

J. MATTIOLI[5], F. TERRIER[1], L. CANTAT[2], J. CHIARONI[4], M. BARRETEAU[5], Y. BONHOMME[2], C. GUETTIER[3], C. ALIX[5]

[1] CEA
[2] IRT SystemX
[3] SAFRAN
[4] Secrétariat général pour l'investissement
[5] Thales
juliette.mattioli@thalesgroup.com

Résumé. La conception de systèmes critiques comme dans le domaine militaire, aéronautique ou spatial à base d'IA n'est pas neutre. Celle-ci peut démarrer par un PoC (Proof of Concept - preuve de principe) destiné à démontrer la faisabilité d'une solution. Mais ce PoC n'est pas en soi une finalité pour la simple raison que les systèmes militaires doivent être par construction fiables, sûrs et (cyber)-robustes. Il devient alors nécessaire de repenser et d'outiller les ingénieries algorithmique, logicielle et système pour industrialiser le PoC en produit ou solution, conforme aux concepts d'emploi et doctrines tout en respectant des propriétés comme la sûreté.

Mots clés: Ingénierie algorithmique de l'IA· Sûreté · Fiabilité · Vérification et validation · Qualité · Evaluation.

1 Les enjeux de l'intelligence artificielle de confiance pour la défense

Les avancées en intelligence artificielle (IA) de ces dernières années ont montré des résultats prometteurs dans le domaine de la défense, contribuant à la supériorité informationnelle et décisionnelle [1]. Cette discipline, définie [2] comme le "*champ, interdisciplinaire et pratique ayant pour objet la compréhension de mécanismes de la cognition et de la réflexion et leur imitation par un dispositif matériel et logiciel à des fins d'assistance ou de substitution à des activités humaines*" permet de comprendre, de prédire, d'anticiper des situations mais aussi d'optimiser des actions, grâce à des capacités d'apprentissage, de raisonnement et de décision. Cependant, le développement de systèmes militaires à base d'IA pose des questions de garantie de fiabilité et de sûreté. En parallèle de ce constat, le conseil de l'innovation[6], a lancé en 2018 le **Grand Défi** de l'IA de confiance,

[6] 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.

avec l'objectif de sortir de l'heure des PoC (preuves de concept) par une réponse à la question de la qualification et de la sûreté des systèmes critiques à base d'IA. Il faut cependant noter que la France n'est pas seule à se lancer dans cette course. La DARPA a lancé en 2018 [3], un programme de 2 M$, **Next AI** visant à améliorer la robustesse et la fiabilité des outils d'IA.

1.1 Le Grand Défi national de l'IA de confiance

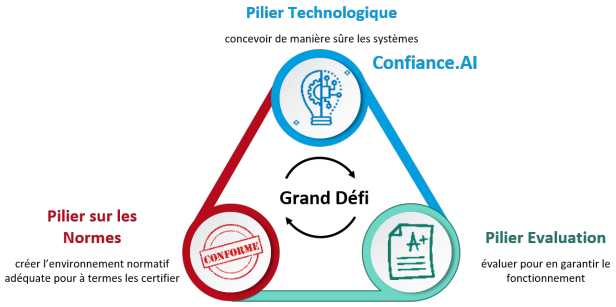

Fig. 1. Le grand défi national de l'IA de confiance repose sur 3 piliers : le pilier technologique avec le programme Confiance.AI, le pilier évaluation et le pilier normalisation

Un certain nombre de verrous freine aujourd'hui, le déploiement de l'IA dans les systèmes militaires. Qu'ils reposent sur des techniques d'apprentissage ou sur des approches plus symboliques, leur conception n'est pas neutre. En effet, ces systèmes doivent suivre des principes de confiance et de responsabilité, garantir par construction des propriétés de sécurité, de sûreté et de fiabilité, qu'il faut pouvoir démontrer. Or les pratiques d'ingénierie de l'IA sont fortement en rupture vis-à-vis des pratiques plus classiques de part 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 devient alors extrêmement difficile de définir et comprendre leur enchaînement et donc d'établir la conformité des fonctions implantées. Il est alors nécessaire de repenser et outiller les ingénieries classiques pour garantir la conformité des produits réalisés vis à vis des concepts d'emploi et des doctrines. **Confiance.AI**[7] (fig. 1), le programme du Grand Défi comble ces attentes, en définissant des méthodes et des outils pour sécuriser l'ensemble des phases de la conception au déploiement tout en garantissant les propriétés de fiabilité, de (cyber)-sécurité du système tout au long de son cycle de vie.

1.2 Une conception systémique pour un déploiement sûr

Il est nécessaire de disposer d'outils d'**ingénierie de la donnée et de la connaissance** permettant la collecte, l'acquisition, l'analyse, la manipulation, la

[7] Les partenaires de Confiance.AI sont : Airbus, Air Liquide, ATOS, CEA, EDF, Inria, IRT SystematiX, IRT St Exupéry, Renault, Safran, SopraSteria, Thales et Valéo

qualification des jeux de données d'apprentissage mais aussi des bases de connaissances au regard de certains principes comme la loyauté et/ou l'équité[8]. Ensuite, l'**ingénierie algorithmique** [4] doit être enrichie afin de prendre en compte les spécificités de l'IA de confiance, démontrant que les fonctions implémentées sont corrects, prévisibles, stables, reproductibles, explicables, fiables, robustes. Celle-ci doit prendre en compte l'incertitude induite par la dynamique de l'environement dans lequel le système évolue. Enfin, il faut être capable de détecter les erreurs sur un domaine d'emploi défini, et in fine si nécessaire être certifiable. C'est pourquoi, l'intégralité du processus d'**ingénierie système** doit être outillée. De plus, dans un contexte de montée de l'autonomie de certaines fonctions, ces ingénieries doivent être revisitées pour prendre en compte les contraintes d'**embarquabilité** et la **relation homme-système**.
La cartographie des fonctionnalités nécessaires à la conception et au maintien en condition opérationnelle d'un système critique à base d'IA est présentée figure 2

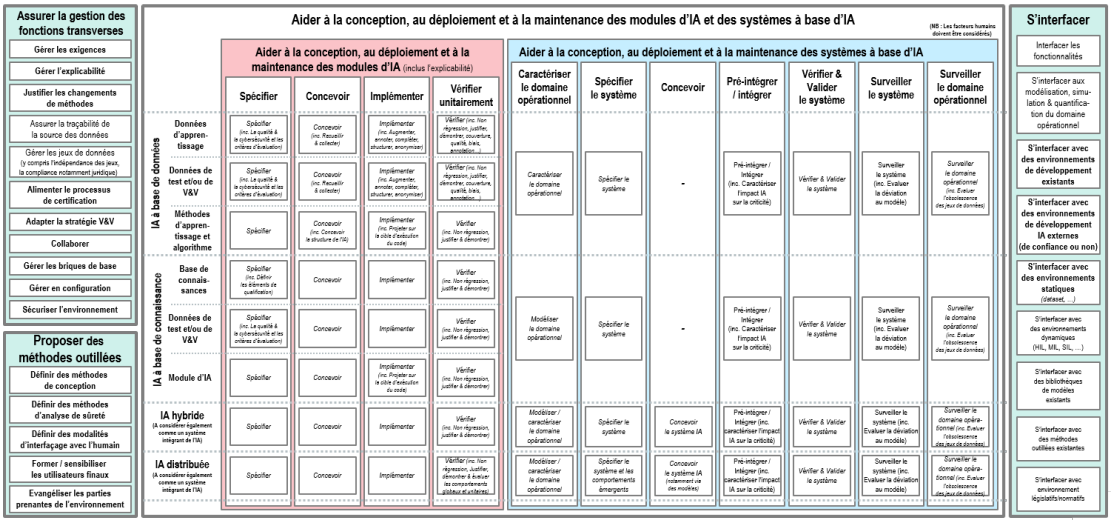

Fig. 2. La matrice de fonctionnalités de l'ingénierie de l'IA de confiance pour un déploiement opérationnel [Source Confiance.AI]

L'objectif de cet article est de faire un zoom sur les enjeux induits et leur déclinaison sur l'ingénierie algorithmique d'IA de confiance.

2 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**

[8] La loyauté implique que les résultats soient conformes aux attentes des utilisateurs. L'équité va plus loin. Les résultats ne doivent pas opérer de distinction entre les personnes en fonction d'attributs protégés par la loi telle que l'ethnie, le genre...

dirigée par les données, appelée aussi IA statistique et connexionniste, sous les feux de la rampe ces dernières d'années avec 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. Comme le montre la figure 3, les fonctionnalités d'observation comme la détection d'anomalies, de menaces et de reconnaissance de cibles permettant une fine compréhension de la situation tactique reposent aujourd'hui sur des algorithmes à base d'apprentissage [5, 6] alors que celles de décision comme la préparation de mission nécessitent des outils d'IA symbolique comme la planification sous contraintes [7, 8].

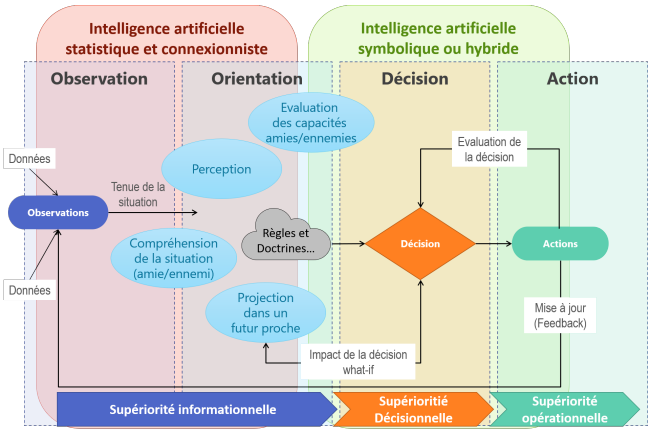

Fig. 3. L'IA statistique et connexioniste, l'IA symbolique et l'IA hybride au service de la boucle OODA

Plusieurs travaux cherchent à hybrider ces deux paradigmes, comme le souligne N. Asher[9]: "*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.

2.1 Conception algorithmique de confiance

Pour garantir une conception algorithmique de confiance (robuste, fiable...), l'ingénierie algorithmique (Algorithm Engineering) définit par P. Sanders [9, 10], doit intégrer les paradigmes induits par l'IA ainsi que les dimensions de (cyber)-sécurité et l'humain dans la boucle. Une attention particulière doit être portée sur :

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

- **La correction** en prouvant que l'algorithme produit au moins un résultat conforme à la spécification. Ainsi, l'algorithme fait bien ce qu'on attend de lui, tout ce qu'on attend de lui.
- **L'explicabilité** : l'algorithme doit être capable d'expliquer, de façon intelligible pour un opérateur/usager, les raisons de ses choix même s'il manipule des notions ou concepts qui échappent à la compréhension humaine.
- **La contrôlabilité** : Il faut valider que l'algorithme fait uniquement ce qu'on attend de lui, tout en restant dans son domaine d'emploi.
- **La robustesse** : Il faut évaluer son aptitude à fournir des réponses correctes face à des situations inconnues ou à des malveillances.

2.2 Vérification et Validation d'un algorithme d'IA

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

- 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 [11, 12] 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 [13], Satisfiabilité modulo théories [14], programmation linéaire, etc.), mathématiques de ses comportements possibles. 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 [15]. 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 [16].

3 Ingénierie des données et Ingénierie des connaissances

Dans sa version data-centrée, les données sont donc cruciales pour l'apprentissage, le test et la validation des IA. 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 qu'il faut éviter. 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'exploration, d'enrichissement, d'annotation et de préparation des données. Par exemple, la **décomposition du jeu de données** en plusieurs sous-ensembles

dédiés à l'apprentissage, la validation et le test, et l'opération du modèle doit respecter la **représentativité** du jeu de données pour la tâche et son domaine d'emploi. 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" [17]) 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 [18] 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[10], la question de la qualité de la donnée (data quality [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.

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. L'ingénierie de la connaissance (IC) fournit quant à elle, une démarche méthodologique de représentation et de résolution de problèmes [20], s'intéressant à la complétude, à la pertinence et à la qualité des modèles.

[10] norme ISO 8000 relative à la qualité des données de référence – Master data

4 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 rédigépar la DGA [21] 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, noté que de nombreux projets s'attellent à la question de l'évaluation. 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 [22, 23], 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 rapide pour l'IA.

5 Conclusion

Pour déployer de l'IA dans les systèmes critiques comme les systèmes militaires, il est nécessaire de revisiter les ingénieries algorithmique, ingénierie logicielle et système. Le programme Confiance.AI du Grand Défi national a pour objectif de définir et d'outiller une approche rigoureuse et interdisciplinaire formalisant la conception et la validation de ces systèmes.

References

1. Ministère des Armées. L'intelligence artificielle au service de la défense, 2019.
2. Journal Officiel texte N°58 du 9/12/2018. Définition de l'intelligence artificielle.
3. DARPA. Ai next campaign., 2018.

4. P. Sanders. Algorithm engineering–an attempt at a definition. In *Efficient Algorithms*, pages 321–340. Springer, 2009.
5. M. Leclerc, R. Tharmarasa, M. Florea, and others. Ship classification using deep learning techniques for maritime target tracking. In *21st Int. Conf. on Information Fusion (FUSION)*, pages 737–744. IEEE, 2018.
6. T. Liu, R. Tharmarasa, S. Halle, M. Florea, and others. Anomaly detection with pattern of life extraction for gmti tracking. In *22th Int. Conf. on Information Fusion (FUSION)*, pages 1–8. IEEE, 2019.
7. F. Aligne, P. Savéant, and V. Vidal. Prise en compte d'un contexte évolutif dans la coordination des opérations de secours. In *IIn Cinquième Workshop Interdisciplinaire sur la Sécurité Globale (WISG 2011)*, 2011.
8. Ch. Guettier and F. Lucas. A constraint-based approach for planning unmanned aerial vehicle activities. *The Knowledge Engineering Review*, 31(5):486, 2016.
9. M. Muller-Hannemann and S. Schirra. *Algorithm engineering: bridging the gap between algorithm theory and practice*. Springer-Verlag, 2010.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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.
15. 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.
16. J. Cohen, E. Rosenfeld, and JZ. Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019.
17. S. Mei and X. Zhu. Using machine teaching to identify optimal training-set attacks on machine learners. In *AAAI Conf. on Artificial Intelligence*, 2015.
18. L. Pipino, Y. Lee, and R. Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.
19. Y. Huh, F. Keller, Thomas C. Redman, and A. Watkins. Data quality. *Information and software technology*, 32(8):559–565, 1990.
20. R. Studer, VR. Benjamins, and D. Fensel. Knowledge engineering: principles and methods. *Data & knowledge engineering*, 25(1-2):161–197, 1998.
21. DGA (French MoD). Guide méthodologique pour la spécification et la qualification des systèmes intégrant des modules d'intelligence artificielle - version 1.0 b, 2019.
22. 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.
23. Darpa program assured autonomy, https://www.darpa.mil/program/assured-autonomy.