Vers la confiance, voire la certification, des systèmes à base d'intelligence artificielle

Par **Julien CHIARONI**
Directeur du grand défi sur « la sécurisation, la fiabilisation et la certification des systèmes à base d'intelligence artificielle », secrétariat général pour l'investissement (SGPI)

Le fonctionnement sûr des logiciels est depuis longtemps au cœur de nombreuses applications. Toutefois, la question reste ouverte lorsque les systèmes intègrent de l'intelligence artificielle (IA). Que l'on pense à la sûreté d'une prise de décision « autonome » en temps réel, à des domaines ne tolérant pas l'erreur de décision ou encore à des attentes d'équité des traitements qui exigent la garantie que ceux-ci ne sont pas biaisés, la confiance placée dans les systèmes intégrant de l'IA doit impérativement être développée.

Les enjeux de l'intelligence artificielle de confiance, voire de sa certification pour les applications qui le requièrent

Un fonctionnement de type « boîte noire » des algorithmes d'apprentissage profond impliquant que l'on ne sache ni les expliquer, ni les garantir

Les récents progrès en IA résultent des avancées dans un domaine numérique particulier, celui de l'apprentissage profond ou *deep learning*, associées à l'augmentation de la performance des architectures électroniques (essentiellement les *Graphics Processing Unit* ou GPU) et à la disponibilité de grandes quantités de données (ou *big data*).

Néanmoins, cette méthode présente un inconvénient majeur. Son fonctionnement est opaque et souvent qualifié de « boîte noire », « dans le sens où l'on peut juger des données qui entrent dans la boîte et des résultats qui en sortent, mais sans savoir ce qui se passe à l'intérieur » [(1)]. Cela implique que l'utilisateur final ne sait en général pas comment elle fonctionne et le concepteur comment en garantir le « bon » fonctionnement.

Un frein majeur à la diffusion de l'IA dans de futurs produits « industriels »

Aujourd'hui, l'intelligence artificielle occupe une place de plus en plus prépondérante dans notre quotidien, au travers par exemple des moteurs de recherche et de recommandations, ou des assistants vocaux. Nous mobilisons ainsi des algorithmes d'une complexité croissante pour produire des services de plus en plus personnalisés. Ces usages requièrent des garanties éthiques et de transparence, notamment inscrites dans la loi en 2016 [(2)]. Ils ne sont toutefois pas, pour la plupart, critiques au sens où elles ne présentent pas ou peu de risques pour les biens et les personnes. « Lorsqu'il s'agit de battre un champion de go ou bien de recommander le film du dimanche soir, la machine peut se tromper, ça n'est pas très grave, votre soirée peut en être gâchée, mais guère plus. » [(3)]

(1) « Les boîtes noires du "deep learning" », *Les Échos*, Benoit Georges, le 27 août 2018.
(2) Loi pour une République numérique (2016) : « Une décision individuelle prise sur le fondement d'un traitement algorithmique comporte une mention explicite en informant l'intéressé. Les règles définissant ce traitement ainsi que les principales caractéristiques de sa mise en œuvre sont communiquées par l'administration à l'intéressé s'il en fait la demande. »
(3) Discours sur l'intelligence artificielle de la ministre des Armées le 5 avril 2019 à Saclay.

Ce n'est pas le cas pour de futurs produits industriels mobilisant des algorithmes d'IA : véhicule autonome, assistant au pilotage, aide au diagnostic médical, contrôle commande industriel, etc. Prenons l'exemple du véhicule autonome : une erreur de décision du système peut potentiellement entrainer un accident grave, dont les conséquences seraient dramatiques pour le conducteur et les usagers ; ceci d'autant plus que des modifications de la perception réussissent parfois à tromper le guidage autonome, provoquant, par exemple, une circulation en sens inverse du véhicule [4]. Pour toutes ces applications dites « critiques », apporter des garanties de fonctionnement (fiabilité, sécurité, disponibilité), voire certifier quand cela s'avère nécessaire, est ainsi un impératif sociétal et économique majeur, mais aussi une véritable opportunité économique, en jetant les bases d'une « IA de confiance pour l'industrie ».

L'impératif d'adopter une approche « système », et pas uniquement algorithmique

Un système à base d'IA est un assemblage de briques logicielles et matérielles permettant de réaliser une fonction ou un service, soit « purement » numérique, soit produisant un actionnement sur le monde physique.

Ce système est un « objet » réalisant une fonction plus ou moins complexe et résultant de l'intégration de trois principales composantes :

- Des données en quantité très importante ou *big data* servant non pas à programmer, mais à estimer un modèle à partir d'« observations » fournies ainsi que des connaissances. La fonction réalisée en est fortement dépendante ce qui implique que ces données et connaissances soient intégrées au système.
- Un algorithme ou une somme d'algorithmes d'IA qui peuvent être de « natures » ou de typologies variables. Par exemple, dans le cas d'un système d'aide à la décision, le concepteur peut avoir recours à des algorithmes d'apprentissage machine, ceux-ci générant des règles à partir de bases de données importantes, ces dernières servant ensuite à alimenter un moteur de règles.
- Un composant ou une architecture électronique sur lesquels sont implémentés les algorithmes avec des contraintes et spécifications associées à l'application visée.

L'algorithme s'intègre ainsi dans un système pour lui conférer de nouvelles fonctions ou propriétés. Il ne doit donc pas être considéré indépendamment de cet ensemble plus large, sans quoi tous les éléments nécessaires ne pourront être pris en compte. C'est pourquoi la confiance, et à terme la certification, ne peut être menée qu'au travers d'une approche « systémique » et pas seulement algorithmique, qui va considérer l'ensemble des composantes, à la fois techniques et fonctionnelles donc dépendantes du contexte d'utilisation et du cycle de vie du produit ; ce dernier point est particulièrement essentiel, dans la mesure où le contexte dans lequel évolue le produit doit être qualifié, mais aussi qu'il peut évoluer sans pour autant que son fonctionnement ne soit impacté.

Un système à base d'intelligence artificielle doit respecter un ensemble d'exigences pour garantir la confiance des utilisateurs et prétendre, suivant les applications, à une certification.

Les terminologies pouvant être amenées à varier, ainsi nous proposons de retenir le regroupement d'exigences suivant, bien qu'il existe évidemment des liens, voire des recoupements entre eux :

(4) Tesla.

- Responsable (éthique, valeurs) : « L'emploi de l'IA, comme l'utilisation de toutes les autres technologies, doit toujours être aligné sur nos valeurs fondamentales et respecter les droits fondamentaux. L'objectif des lignes directrices en matière d'éthique est de s'en assurer dans la pratique [5]. » Citons l'exemple des « biais » par lesquels l'algorithme peut discriminer un groupe de personnes en particulier. Plusieurs initiatives ont proposé des principes permettant de garantir une IA responsable ; parmi ces initiatives figurent les lignes directrices posant le « cadre d'une IA digne de confiance » par la France (GPAI), les travaux de l'Union européenne au travers du High Level Expert Group [6] (HLEG AI) ou encore la déclaration de Montréal [7].
- Fiable et explicable (ou auditable) : le système est capable d'expliquer et d'informer les utilisateurs (ou les concepteurs) de propriétés ou limites de fonctionnement, de ses choix ou raisonnements relatifs à la décision. À cela s'ajoute l'instauration d'un mode de communication entre l'individu et l'IA, afin de permettre une assistance à l'opérateur, à l'utilisateur ou au concepteur de systèmes. Notons que l'« auditabilité » est aussi une notion centrale, notamment pour les systèmes autonomes, qui permettent de comprendre et corriger *a posteriori* une erreur de décision. Enfin, d'autres propriétés techniques doivent également être prises en compte comme la fiabilité, dans laquelle nous inclurons notamment les concepts de robustesse (à savoir l'évaluation de l'aptitude du système à fournir des réponses correctes y compris face à des situations inconnues ou à des malveillances), la contrôlabilité (à savoir la garantie que le système ne fait que ce que l'on attend de lui et rien d'autre, notamment le maintien dans son domaine d'emploi), etc.
- Certification : le produit, service ou système doit être conforme à des exigences spécifiques. Pour ce faire, il faut garantir le fonctionnement du système (fiabilité, robustesse, reproductibilité ou sécurité, soit un ensemble d'éléments qui fournissent des informations sur le système et qui permettent donc en partie de l'expliquer) avec un niveau d'exigence et de criticité définis dans le cahier des charges. Enfin, des normes et standards doivent être édictés. Cela dépend évidemment des secteurs, applications et risques afférents, qui ne concernent donc pas tous les produits, services ou systèmes. À ce titre, le livre blanc de la Commission européenne sur l'IA [8] souligne notamment ce qu'elle considère comme des applications à « haut risque », selon deux critères : le secteur de marché (santé, mobilité, etc.) et le risque pour les bien et les personnes. Citons aussi la publication de l'European Union Aviation Safety Agency [9] (EASA) qui présente une première feuille de route proposant un schéma qui repose sur des *trustworthy AI building blocks*.

Revisiter l'ensemble de la chaîne de conception et d'évaluation des systèmes à base d'intelligence artificielle en vue d'apporter les garanties de confiance nécessaires à leur déploiement

Pour parvenir à développer « une IA de confiance pour l'industrie », il est essentiel de répondre aux problématiques suivantes : concevoir de manière sûre les systèmes, savoir les évaluer pour en garantir le fonctionnement et créer l'environnement normatif adéquat pour, à terme, les certifier.

(5) Discours de Mariya Gabriel, commissaire chargée de la politique numérique, à la Commission européenne.
(6) Lignes directrices en matière éthique pour une intelligence artificielle de confiance, groupe d'experts de haut niveau, avril 2019.
(7) « Déclaration de Montréal pour un développement responsable de l'intelligence artificielle », initiative de l'Université de Montréal lancée en décembre 2018.
(8) Livre blanc (2020), « Intelligence artificielle : une approche européenne axée sur l'excellence et la confiance », commission européenne, février.
(9) EASA (2020), "Artificial Intelligence Roadmap 1.0", février.

Un impératif de conception « systémique » pour le déploiement sûr, voire certifiable, des produits à base d'intelligence artificielle

La réalisation de systèmes critiques à base d'IA nécessite de revisiter les ingénieries classiques (ingénierie de la donnée et de la connaissance, ingénierie algorithmique et ingénierie système) et de les enrichir. Il faut être en mesure de s'assurer de la conformité du système aux besoins et aux contraintes du client, définir des méthodes et outils pour sécuriser l'ensemble des phases de conception, mais aussi garantir des propriétés de type fiabilité, sécurité et cybersécurité, et « 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 la démarche de « génie de l'IA », en prenant en compte les dimensions algorithmiques, logicielles et systèmes, mais aussi celles de la donnée et des connaissances, afin de faire émerger les bases d'« une IA de confiance pour l'industrie ».

Il s'agit ainsi d'outiller la chaîne de conception en vue de respecter les exigences et spécifications, d'apporter le maximum d'éléments de preuve et d'expliciter le processus, pour au final réduire la quantité de tests nécessaires à la qualification, voire à la certification, du ou des systèmes. Il est indispensable de disposer ou de revisiter, afin de tenir compte de l'introduction de l'IA, de plusieurs sous-ensembles de la chaîne de conception : les outils d'ingénierie de la donnée et de la connaissance permettant de maitriser l'ensemble des étapes nécessaires à l'obtention d'une base de données qualifiée et représentative du domaine d'emploi concerné par le système, les outils d'ingénierie algorithmique tels que définis par Peter Sanders [10] et d'autres [11] au début des années 2000, et incluant la validation et la vérification des algorithmes, les outils d'ingénierie système, sans oublier la prise en compte au besoin des contraintes d'« embarquabilité » et d'interface entre l'homme et la machine.

L'évaluation de la conformité des algorithmes et des systèmes à base d'IA au cœur du processus de certification

La confiance dans les systèmes à base d'intelligence artificielle requiert également le développement de plateformes d'évaluation aux besoins des applications et services. En effet, en se focalisant sur l'évaluation des « performances » d'une IA à base de données, celle-ci peut se résumer à évaluer la « qualité » d'une fonction dont l'apprentissage est statistique. se pose ainsi la question de la représentativité des données utilisées, de leur couverture par rapport au domaine d'emploi, mais aussi de possibles problèmes de surreprésentation pouvant induire un biais dans le système. À cela s'ajoutent deux questions centrales : quelles sont les métriques les plus adaptées pour réaliser les évaluations ? Comment réaliser une quantité suffisante de tests ?

Pour ce faire, revisiter nos approches techniques est fondamental, notamment en ayant davantage recours à des chaînes de simulation générique, interopérables et évolutives, associées à des schémas de tests (ou *scenarii*) représentatifs du contexte et du domaine d'emploi, voire même des potentielles évolutions de ces derniers. Toutefois, se pose alors la question de la qualification des outils et des modèles eux-mêmes, mais aussi de la reproductibilité des tests. Cela constitue un élément central pour que la simulation puisse être utilisée comme une preuve « acceptable » dans le cadre d'un processus règlementaire d'évaluation et de pré-certification. La définition d'exigences, l'injection de données réelles, voire la réalisation de tests réels comparés constituent des pistes pour permettre de valider cette conformité et par là-même les résultats.

(10) SANDERS P. (2010), "Algorithm engineering: an attempt at a definition using sorting as an example", *Proceedings* of the Twelfth Workshop on Algorithm Engineering and Experiments (ALENEX), *Society for Industrial and Applied Mathematics* – SIAM, pp. 55-61.

(11) ZHANG H., WENG T. W., CHEN P. Y., HSIEH C. J. & DANIEL L. (2018), "Efficient neural network robustness certification with general activation functions", Advances in neural information processing systems (NIPS), pp. 4939-4948.

L'importance de normes et de standards pour le monde numérique, et tout particulièrement l'intelligence artificielle

Développer des normes volontaires est l'une des actions essentielles pour accompagner le développement de l'IA de confiance, qui prennent en considération les enjeux socio-économiques et certains enjeux spécifiques tels que l'éthique (transparence des plateformes, égalité de traitement, etc.), la sureté et la sécurité des biens et des personnes, ceci dans l'objectif d'assurer la souveraineté numérique et de préparer un cadre répondant aux besoins industriels sectoriels comme la santé, l'industrie du futur ou la mobilité, notamment en termes de responsabilité. Certaines normes essentielles sont en cours de réflexion à l'International Organization for Standardization (ISO) (concepts et terminologie ISO/IEC 22989 sur l'éthique, ISO/IEC 24372 ou le management du risque, ISO/IEC 23894), en parallèle de la publication par certains pays de feuilles de route, comme récemment par le German Institute for Standardization (DIN) en Allemagne [12].

Conclusion

La confiance est un élément central pour la diffusion de l'IA dans de nombreux produits et services. À cette fin, une approche règlementaire et la définition de « grands principes » (biais, robustesse, etc.) sont essentielles, mais leur « application » directement dans les systèmes se doit d'être démontrée et prouvée, afin que cette exigence soit effectivement respectée et implémentée, comme cela a pu être fait dans d'autres secteurs. Il nous faut donc développer un cadre « technique », au travers de la revisite des chaînes d'ingénierie ou au travers d'approches outillées plus adaptées pour l'évaluation, permettant l'introduction de l'IA dans les systèmes critiques, et, plus largement, d'une IA pour l'industrie. L'ensemble de ces briques logicielles contribueront ainsi à un « socle » plus large, celui de la confiance numérique, incluant les infrastructures, les données ou l'électronique.

Remerciements

Je tiens à remercier toutes les équipes impliquées dans le grand défi du conseil de l'innovation, tout particulièrement le consortium « Confiance.ai », initiative du grand défi sur la conception d'outils logiciels et des méthodologies associées.

(12) "The German Standardization Roadmap on Artificial Intelligence", novembre 2020.