# IA, explicabilité et défense

Olivier Kempf - Éloïse Berthier

Général de brigade (2S). Docteur en science politique, chercheur associé à la FRS et directeur de la collection « Cyberstratégie » chez Économica. Consultant en cyber et digital (OK Conseil). Dirige la lettre stratégique *La Vigie*.

Polytechnicienne, ingénieur de l'armement recherche, étudiante à l'École normale supérieure Paris-Saclay.

## Quel est le problème ?

L'intelligence artificielle (IA) est le concept « numérique » dont on parle le plus depuis ces derniers mois. Les grands noms (Elon Musk, Stephen Hawking) s'en émeuvent, les publicistes en vogue écrivent des livres dessus (Luc Ferry, Laurent Alexandre), le gouvernement appelle une médaille Fields pour écrire un rapport sur le sujet (rapport Villani) : autant dire que tout le monde a entendu parler d'IA, sous les atours les plus flatteurs et les plus inquiétants, d'ailleurs pour la même raison : ce serait capable de tout faire mieux que l'humain.

Il faut bien sûr raison garder et se méfier de ces modes qui animent régulièrement le débat public. Observons au passage qu'il s'agit là d'une résurgence (avec d'autres mots) d'un débat très ancien sur le progrès et son rôle dans nos sociétés humaines : le mythe de Prométhée est antique, lui qui vola le savoir divin pour le donner aux hommes. La tension entre le savoir et la connaissance (voire la sagesse) est une question philosophique classique qui trouve ici de nouveaux atours. Ajoutons le mythe de la créature qui prend le pas sur son créateur : là encore, de *Pygmalion* à *Frankenstein* puis *Dr Jekyll et Mr Hyde*, l'humanité a construit beaucoup de modèles inquiétants : sait-on d'ailleurs que Mary Shelley sous-titra son roman « Le Prométhée moderne » ?

Gardons donc la raison pour appréhender cet objet plus réduit qu'on ne le dit, l'IA.

### *Qu'est-ce que l'IA ?*

Beaucoup confondent, sciemment ou non, l'IA et le *Big Data*. De même, beaucoup l'assimilent à l'apprentissage machine qui ne constitue qu'une des techniques de l'IA, même si c'est celle qui connaît aujourd'hui les développements les plus rapides. Toutefois, rappelons que *Deep Blue*, l'ordinateur d'IBM qui a battu Kasparov en 1997, utilisait une technique de « force brute » : il calculait 200 millions de positions à la seconde quand Kasparov n'en cherchait probablement pas plus de 5. Ajoutons de plus que les échecs, comme le go, obéissent à un terrain de jeu et une gamme de règles finalement extrêmement réduite, ce qui rend la modélisation

aisée. Il reste qu'on observe aujourd'hui des machines qui apprennent le jeu d'échec en trois jours et sont alors capables de vaincre 98 % des joueurs : les nouvelles techniques d'IA, basées sur l'apprentissage, ont permis des développements radicaux qu'il serait vain de nier.

### *L'IA, une discipline scientifique*

Le rapport Villani se signale par l'absence de définition de l'IA ! On retrouve ici un biais courant : on en parle sans la définir. Mais définir impose certes de réduire et limiter, mais aussi de préciser et comprendre.

Selon le Larousse, l'IA est « l'ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence ». Selon Stuart Russel, auteur du manuel de référence sur l'IA (*Intelligence artificielle : une approche moderne*), c'est « l'étude des méthodes permettant aux ordinateurs de se comporter intelligemment ». Problème : rien n'est dit de ce qu'est l'intelligence…

On peut voir l'IA comme une finalité, un état à atteindre, à savoir la reproduction (voire le dépassement) de certaines facultés humaines par la machine. Ces facultés sont les mêmes que celles que les psychologues identifient chez l'homme pour définir l'intelligence. Mieux vaut donc dire que c'est une discipline scientifique réunissant plusieurs techniques (neurobiologie, logique mathématique, informatique…) qui cherche à produire des dispositifs imitant ou remplaçant certaines fonctions cognitives de l'homme.

On évoque couramment la perception (de l'image, du son, du mouvement…) ; le langage (compréhension et production de texte, représentation de structures sémantiques, traduction…) ; le mouvement (la navigation en robotique, les mouvements précis des robots industriels ou de chirurgie) ; le raisonnement (construction de raisonnements logiques et de décisions, élaboration de stratégies pour résoudre des problèmes, dans les jeux ou en médecine par exemple) ; l'interaction sociale, interaction avec des humains (éprouver et faire éprouver de l'empathie, comme le robot *Nao*) ou interactions entre machines (comme dans les essaims de drones) ; l'apprentissage (l'adaptation du comportement au cours du temps en fonction des expériences passées).

L'IA a cependant une longue histoire. Alan Turing, le fameux mathématicien britannique qui mit au point la machine *Enigma* pendant la Seconde Guerre mondiale, publia en 1950 un article se demandant si une machine (l'informatique était émergente) pouvait penser. Une conférence réunissant les pionniers de ce qui allait devenir l'IA se tint en 1956 à Darmouth : elle allait essaimer dans plusieurs universités avec des auteurs comme John McCarthy, Marvin Minsky ou Herbert Simon.

Initialement cantonnée à un cercle de spécialistes, l'IA a connu des développements importants dès les années 1980 (premiers systèmes experts) puis sur-

tout à partir des années 2000, concomitamment aux vagues successives des révolutions informatiques. Les géants du numérique s'en servent de plus en plus tandis qu'IBM construit une IA qui devient un produit commercial (Watson) et finalement le pilier de la transformation de l'entreprise. Cette influence grandissante conduit à deux mouvements contraires : d'une part, l'idée née dans la Silicon Valley que nous sommes à l'aube d'un bouleversement technique et que l'IA va tellement se développer qu'elle va permettre à l'humanité de se dépasser (notion d'homme augmenté, de singularité, voir l'université de la singularité fondée par Google) ; d'autre part, un mouvement inquiet qui croit justement à ces possibilités et en dénonce les côtés totalitaires (l'humanité va être dépassée par l'IA et lui devenir subordonnée, ce qu'expriment les inquiétudes de Bill Gates ou d'Elon Musk).

#### *Des intelligences artificielles*

Compte tenu de la variété des procédés, il vaut mieux parler des IA au pluriel plutôt que de l'IA qu'on ne saurait réduire aux réseaux de neurones. Il y a en fait deux grandes catégories d'IA : une IA symbolique, fondée sur des ensembles de règles et d'équation ; et une IA connectiviste, fondée sur des statistiques et des grands nombres. Cette dernière est principalement basée sur des réseaux de neurones, types d'algorithme inventés dès les années 1980 mais elle resta longtemps improductive. Cette technique a connu un développement fulgurant au cours des dix dernières années grâce à deux facteurs : d'un côté, une masse de données incroyable permise par le *Big Data* ; de l'autre, la puissance de calcul disponible, grâce aux progrès des microprocesseurs mais aussi à la mise à l'échelle permise par l'infonuagique (*cloud*).

Parmi les techniques d'IA symbolique, on peut évoquer les systèmes experts, qui reproduisent des ensembles de règles et ont connu du succès grâce au caractère très précis de l'activité que l'on simulait. Le calcul formel (opposé au calcul numérique) permet de traiter les expressions symboliques, ce qui donne la « représentation des connaissances ». La simulation du raisonnement humain tente de formaliser le raisonnement humain (logiques modales, floues, temporelles, etc.). Le traitement du langage naturel, la résolution de problèmes (utilisée pour simuler les différents jeux comme les échecs ou le backgammon), la reconnaissance de la parole ou de l'écriture sont d'autres domaines de recherche. La robotique s'est beaucoup développée, avec plusieurs générations d'IA associées : reproduction de mouvements enregistrés, incorporation de capteurs pour prendre des décisions, développement de robots plus autonomes pouvant se déplacer en fonction de l'environnement. Parmi les techniques utilisées, mentionnons l'optimisation combinatoire, les méthodes algorithmiques issues des graphes, les moteurs d'inférences, la logique floue ou la programmation par contraintes (PPC).

L'apprentissage est à la base de l'IA connectiviste (ou statistique, même si des techniques d'IA symbolique utilisent des méthodes statistiques sans qu'elles reposent

sur de l'apprentissage : arbres de décision, régression linéaire…). Par exemple, le calcul d'itinéraire d'un *GPS* se fait par une méthode déterministe d'optimisation de parcours de graphe. Mais le gros de l'apprentissage aujourd'hui est basé sur les réseaux neuronaux. Ce sont des couches superposées de neurones artificiels permettant de traiter successivement une information pour parvenir à déterminer, en fin de processus, une conclusion (un *output*). La performance de ces réseaux s'accroît à grande vitesse et donne des résultats étonnants dans plusieurs domaines (reconnaissance vocale, d'image, de texte, de vidéo).

Les progrès dans la faculté d'apprentissage irriguent d'autres facultés, en particulier aujourd'hui la reconnaissance d'image, la compréhension du langage ou la stratégie dans les jeux. Il n'est pas surprenant, même chez un humain, qu'en gagnant en faculté d'apprentissage, on puisse progresser dans d'autres domaines.

Quelques exemples assez parlants qui montrent cette tendance d'irrigation de l'apprentissage vers d'autres facultés :

• La reconnaissance d'image se fait par application de filtres (à peu près le même principe qu'un filtre photographique) pour en extraire des caractéristiques remarquables. Ces filtres ont été construits de façon « artisanale » par les chercheurs en traitement d'image durant des décennies. Dans un réseau de neurones convolutifs, les filtres sont appris à partir de l'expérience, en capitalisant des informations issues de centaines de milliers d'images.

• Traditionnellement, la traduction se faisait par règles logiques de correspondance dans un dictionnaire (si *cat* alors chat). Aujourd'hui (voir les travaux de Facebook AI Research Paris), on construit des modèles de langues en observant statistiquement les fréquences et les cooccurrences de mots dans d'énormes corpus, et on y repère des structures communes.

• *Deep Blue* a battu Kasparov en appliquant une méthode déterministe (tester toutes les combinaisons possibles en force brute). *Alpha Go* a battu Lee Sedol en apprenant sur des millions de parties, y compris contre lui-même.

• Quand un médecin doit prendre une décision (quel traitement appliquer ?), il utilise une série de règles implicites (si tel symptôme ou tel résultat d'analyse de sang, alors…) apprises pendant ses études et avec la pratique. Quand Watson s'attaque au même problème, il exploite un grand nombre de diagnostics passés, plus que le médecin n'en verra dans une vie.

Implicitement, dans la méthode déterministe, c'est l'humain qui utilise sa faculté d'apprentissage, forcément limitée par son expérience, et qui la transmet à la machine *via* des règles écrites. Avec l'apprentissage statistique, c'est la machine elle-même qui apprend pour construire son savoir à partir de données qui forment sa propre expérience. D'où les questions des biais introduits par les données et de l'explicabilité des algorithmes.

## Le défi de l'explicabilité

### *Difficultés formelles de l'explicabilité*

Un des grands défis de l'apprentissage statistique est celui de l'explicabilité du résultat. En effet, chaque réseau de neurones fonctionne comme une boîte noire et produit un résultat qui peut apparaître comme magique, sans que l'on puisse démontrer les processus qui l'ont produit.

Il est vrai que l'on fait alors une analogie avec l'IA symbolique, alors que le processus est différent.

Plus précisément, il n'est pas impossible que l'on se trompe de définition quand on parle d'explicabilité. On attendrait d'un modèle statistique qu'il puisse expliquer de façon intelligible pour un humain pourquoi exactement il a pris une décision : « Je classifie cette image comme une voiture parce que j'ai vu 4 roues, et qu'elle est sur une route… » Mais il est illusoire de demander à ce type de modèles de telles justifications sous forme de suite de décisions logiques, ou alors on perdrait ce qui fait leur efficacité. Ils sont efficaces précisément parce qu'ils manipulent une grande quantité de données, à l'aide de méthodes d'optimisation qui nécessitent une grosse puissance de calcul. On a affaire à des processus qui dépassent les capacités de l'humain seul en termes de complexité. Un réseau de neurones ainsi entraîné fournit un modèle puissant, mais forcément complexe (parfois plusieurs dizaines de couches, ce qui représente des centaines de Mo de paramètres… ). Chaque décision est donc motivée par une interaction très complexe entre le modèle et la donnée d'entrée, et on ne peut donc pas la réduire à une suite de justifications simples.

Souvent, par cette question d'explicabilité, on tente de calquer le type de raisonnements que construisent les méthodes d'IA symbolique (une suite de si… alors…) qui sont, de fait, parfaitement explicables, sur des raisonnements statistiques qui sont d'une nature différente et qui ne se prêtent pas à cette analyse.

Un deuxième problème est celui de la qualité et de la sélection des données que l'on place dans la machine (autrement dit, un jeu différent de données change-t-il le résultat ?) et de ce fait la reproduction du résultat, sa neutralité et sa fiabilité. Mais les performances d'un modèle s'expliquent en grande majorité par la qualité des données d'entraînement. L'explicabilité doit d'abord être cherchée du côté des données : qualité des données fournies, biais éventuels… ?

Une autre difficulté tient à la structure et à la qualité de l'algorithme. Un même algorithme entraîné pour une tâche et utilisé pour une autre ne fonctionnera pas. Il fonctionne parce qu'il a été conçu pour une tâche précise : par exemple, des réseaux de neurones en vision ont une architecture différente de ceux utilisés pour le traitement du langage.

### *Difficultés militaires liées à l'explicabilité*

Pourtant, toutes ces difficultés formelles posent la question de la responsabilité du décideur. Si une IA propose une décision, c'est toujours un humain qui la prend. Dans le cas militaire, cela conduit à engager des soldats au combat et à leur faire prendre un risque létal. Or, toute décision militaire doit pouvoir se justifier, pour plusieurs raisons : éthique puisque la prise de risque létal, sur ses troupes, l'ennemi ou la population, soulève un problème éthique avant celui de l'efficacité, notamment si « les choses se passent mal ». On se souvient ici de la plainte formulée par des familles de soldats français tués dans l'embuscade d'Usbine, en Afghanistan, et qui mettaient en cause la responsabilité du commandement français.

La deuxième difficulté a trait à la sécurité de l'IA : comment être sûr qu'une IA donnera toujours les mêmes résultats et aura donc une permanence de sa performance ? L'IA sera en effet incorporée dans plusieurs dispositifs militaires (systèmes d'armes ou aide à la décision d'état-major). Cette utilisation de plus en plus importante entraîne une exigence de qualité. On peut ici ajouter les risques de cybersécurité liés à l'IA : une IA est un système numérique susceptible d'être attaqué par d'autres systèmes numériques.

La dernière difficulté se réfère à la confiance, aussi bien du commandement que des troupes engagées. Or, cette confiance repose d'abord sur des liens humains et systémiques. Certes, les forces modernes ont appris depuis longtemps à utiliser des machines puissantes auxquelles elles « font confiance », mais cela résulte d'un long « apprentissage ». Il est nécessaire de rassurer les forces quant à l'usage de l'IA, dans une sorte d'apprentissage de l'apprentissage.

### *Vers de nouvelles formes d'explicabilité*

Lancé en 2016 par la DARPA (Defense Advanced Research Projects Agency), le programme *Explainable Artificial Intelligence* (*XAI*) a pour but de stimuler la recherche de nouvelles formes d'explicabilité pour les systèmes d'apprentissage statistique.

Un point d'étape présentant les premières contributions fin 2017 [(1)] a mis en évidence deux catégories d'approches : la première consiste à développer des outils d'analyse pour les méthodes d'apprentissage existantes, la seconde, plus ambitieuse, à construire des modèles statistiques intégrant directement une composante d'explicabilité. Cette dernière pourrait par exemple s'appuyer sur des techniques d'inférence causale, qui sont utilisées aujourd'hui, entre autres, dans les études épidémiologiques. Néanmoins, l'intégration de telles méthodes dans des modèles d'apprentissage statistique est loin d'être un problème résolu. La première approche a, quant à elle, déjà donné des résultats concrets, sous la forme d'outils

(1) David Gunning : « Explainable Artificial Intelligence (XAI) », DARPA, 2017.

de communication du système vers l'utilisateur. Par exemple, pour un réseau de neurones qui classifie des images, il est possible de visualiser les zones de l'image qui sont responsables du résultat fourni (figure ci-dessous).

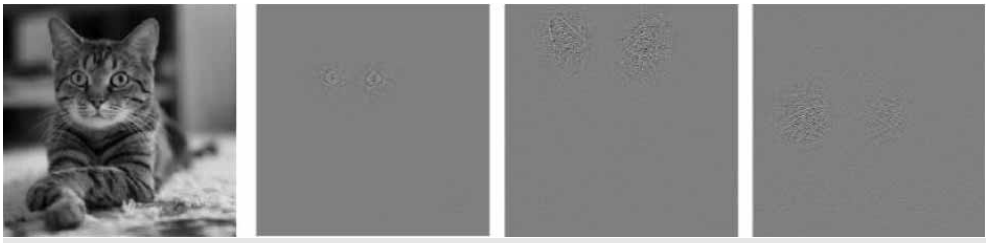

**Figure 1** - Reproduction de l'expérience de Zeiler & Fergus [2] pour visualiser les zones d'une image ayant influencé le résultat retourné par un réseau de neurones. Ici, pour classifier correctement l'image d'entrée (à gauche) comme un chat, le réseau s'est concentré sur plusieurs zones particulièrement discriminantes, comme les yeux, les oreilles ou les moustaches.

Dans le même esprit, une équipe de Google Brain [3] propose une méthode générique pour construire automatiquement des concepts justifiant les résultats des réseaux de neurones.

## Quels critères pour assurer la sécurité d'un système apprenant ?

Au-delà de la question de l'explicabilité se pose celle du niveau de confiance qui peut être accordé à un système d'intelligence artificielle. L'explicabilité n'est évidemment pas suffisante pour garantir la sécurité et la performance. Il s'agirait plutôt d'évaluer le système selon des critères objectifs et quantifiables dont nous proposons ici trois exemples. Ces critères s'appliquent aux systèmes d'IA qui s'appuient sur des données et ils mettent l'accent sur quelques défis scientifiques dans le domaine de l'apprentissage statistique.

### *Mesure des performances*

Le premier critère assurant la sécurité d'un système est celui de la mesure des performances. Supposons que l'on souhaite construire un système détectant tous les piétons sur des images. Ses performances peuvent être quantifiées en termes de précision (parmi les piétons réels, combien ont été détectés), de rappel (parmi les détections annoncées, combien étaient réellement des piétons) ou de taux d'erreur (combien de réponses correctes ont été données). L'utilisation d'une de ces mesures seule n'a pas de sens, car elle peut refléter des réalités bien

(2) Zeiler Matthew D. et Fergus Rob : « Visualizing and Understanding Convolutional Networks », *European Conference on Computer Vision*, 2014, p. 818-833.

(3) Kim Been, Wattenberg Martin, Gilmer Justin, Cai Carrie, Wexler James, Viegas Fernanda et Sayres Rory : « Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV) », *International Conference on Machine Learning*, 2018, p. 2673-2682.

différentes. Ainsi, si un système se contente de ne presque jamais détecter de piéton, il aura une précision proche de 100 %. De même, le taux d'erreur peut se révéler trompeur lorsque l'objet à détecter est rare. Il arrive qu'un certain flou soit entretenu quant aux performances réelles ou que les métriques les plus favorables soient systématiquement choisies *a posteriori*. Le choix du critère d'évaluation doit être décidé en amont de la conception du système, puisqu'il fait partie de la définition du besoin.

De plus, certains modèles fonctionnent très bien sur des jeux de données contrôlés, mais dès que les données d'application s'éloignent du contexte de l'entraînement, les performances sont significativement dégradées. Il est donc nécessaire, avant le déploiement du système, de procéder à des essais sur des données réelles, afin d'en estimer les limites et le domaine de validité. Par exemple, un traducteur automatique entraîné sur des textes littéraires fonctionnera sur le même type de textes, mais généralement pas sur de l'argot.

### *Robustesse*

Le deuxième critère de sécurité est celui de la robustesse. On s'attend notamment à ce que pour deux entrées très proches, le système renvoie deux sorties proches. Ce n'est pas le cas pour les réseaux de neurones, connus pour leur instabilité. Ils sont donc vulnérables à des attaques par « exemples contradictoires », consistant à perturber légèrement la donnée d'entrée pour leurrer le système (figure ci-dessous).

L'identification et la lutte contre de telles attaques préoccupent la communauté des chercheurs et de gros efforts sont déployés sur ce sujet. Les attaques sont facilitées quand l'attaquant connaît l'architecture du réseau de neurones cible. Dans un domaine où la plupart des modèles sont disponibles en *open source*, cela pourrait pousser le défenseur à garder son architecture secrète pour se protéger. Certains systèmes, notamment ceux qui utilisent l'apprentissage par renforcement, sont conçus pour continuer à apprendre au contact de leur environnement. Ils sont alors particulièrement sensibles à l'injection malveillante de données pour les faire dévier de leur but. On se

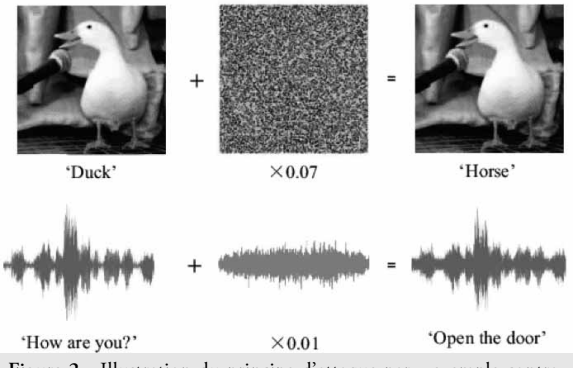


**Figure 2** - Illustration du principe d'attaque par « exemple contradictoire », pour un système de reconnaissance d'images et un système de reconnaissance de parole. Dans les deux cas, un bruit imperceptible est ajouté à la donnée d'entrée, ce qui suffit à rendre la prédiction incorrecte. Source : Yuan Gong & Christian Poellabauer : « An Overview of Vulnerabilities of Voice Controlled Systems », *1st International Workshop on Security and Privacy for the Internet-of-Things* (*IoTSec*), 2018.

souvient de l'exemple de *Tay*, le *chatbot* de Microsoft manipulé sur Twitter pour publier des messages racistes et misogynes.

#### *Protection des données d'entraînement*

Un troisième critère à prendre en compte est celui de la protection des corpus de données d'entraînement. Connaissant un modèle entraîné à partir de données inconnues, on ne sait pas estimer dans quelle mesure il serait possible que des malfaisants le régénèrent pour retrouver des informations sur ces données. Cela pose également la question du degré de sensibilité d'un système apprenant sur des données : est-il plus ou moins sensible que ses données d'entraînement, et si on mélange des données avec plusieurs niveaux de confidentialité, sur lequel faut-il s'aligner ? C'est aujourd'hui une question ouverte qu'il conviendra de traiter, plus fortement encore dans le monde de la défense, mais qui préoccupe déjà le domaine civil pour des enjeux de protection de la vie privée.

Enfin, l'entraînement de systèmes sur de grands volumes de données potentiellement sensibles nécessite de les centraliser, ce qui engendre un risque cyber supplémentaire. Une piste de recherche actuelle consiste à entraîner le modèle de façon décentralisée, c'est-à-dire en distribuant les données sur différentes machines d'un réseau.

## Conclusion

Les questions de l'emploi de l'IA dans la défense se heurtent très souvent à la question de l'explicabilité. Les méthodes d'apprentissage statistique et en particulier les réseaux de neurones sont accusés d'être des « boîtes noires », ce qui est objectivement vrai, dans le sens où ils ne fournissent pas d'explication complète de leurs décisions. Ce n'est pas une raison suffisante pour s'opposer à leur utilisation dans le contexte de la défense, alors qu'ils ont prouvé leur supériorité dans de nombreux domaines (traitement des images, des langues, etc.). Il reste à déterminer comment on peut résoudre la question de l'explicabilité dans un contexte militaire. ◆