Histoire et enjeux de l'IA

Gilbert Saporta

Histoire et enjeux de l'IA

Gilbert Saporta

In *L'IA éducative* (sous la direction de F.Guénot), Bréal, Paris, 2023

1. Définitions

On attribue généralement à John Mc Carthy, alors professeur au MIT, la paternité de l'expression « Artificial intelligence » utilisée en 1956 dans le titre du *Dartmouth Summer Research Project on Artificial Intelligence* [1] mais les racines de cette discipline remontent à bien plus loin comme on va le voir.

Selon le dictionnaire Larousse, l'intelligence artificielle ou IA est un « ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine ».

Le chercheur Yann LeCun[2] en définit l'objectif de manière moins ambitieuse et plus opérationnelle comme « faire faire aux machines des activités que l'on attribue généralement aux animaux et aux humains ».

La fortune de l'expression IA a connu des hauts et des bas (les « hivers ») à tel point que les chercheurs en informatique évitaient de l'utiliser à la fin des années 90 et jusqu'aux années 2010, de peur d'être pris pour des illuminés. Depuis ses succès récents dans de nombreux domaines, en particulier les jeux (Deep Blue pour les échecs et AlphaGo), l'IA est revenue au premier plan médiatique. Tout algorithme ou toute méthode statistique se retrouve qualifiée d'IA, le plus souvent abusivement. Ainsi de la détection de la fraude (aux impôts, à la carte bancaire etc.) ou de la notation des emprunteurs qui se pratiquent depuis plus de 50 ans avec des méthodes éprouvées.

L'IA suppose des processus largement autonomes et automatiques. IA et robotique sont liées mais ne doivent pas être confondus.

On distingue souvent l'IA faible ou étroite dédiée à la résolution d'un problème précis, de l'IA forte ou générale qui serait capable de résoudre toute sorte de problèmes à l'instar de l'esprit humain. Alors que les succès de l'IA faible (certains parlent *des* IAs) sont patents, L'IA forte n'est encore qu'un concept théorique que certains jugent impossible à réaliser.

Sur un plan technique on distinguera ces trois types d'IA :

- **L'IA symbolique** qui modélise le raisonnement humain par le biais de règles logiques ou probabilistes. L'IA symbolique triompha dans les années 80 avec les systèmes experts et les arbres de décision. Elle reste très présente dans le diagnostic médical.
- **L'IA numérique** ou *Machine Learning* qui sous sa version *supervisée* cherche à prédire un comportement ou réponse, à l'aide d'algorithmes basés sur des enchaînements de fonctions mathématiques et statistiques dont les paramètres sont optimisés par apprentissage sur de nombreux exemples. Les méthodes connexionnistes (réseaux de neurones, *deep*

[1] Workshop improprement appelé « conférence de Dartmouth » . À ne pas confondre avec la conférence de Dartmouth qui était un dialogue bilatéral américano-soviétique sur la paix débuté en 1959.
[2] Yann Le Cun a américanisé son patronyme en LeCun

learning) en sont les exemples les plus connus. On les qualifie souvent de *boîtes noires* en raison de leur complexité.

- **L'IA générative,** la plus récente, qui permet de créer des contenus à la demande (textes, images, etc.) à partir de données d'apprentissage.

2. Mécanisation de la pensée

Représenter l'homme (ou l'animal) comme une machine a permis à des penseurs d'imaginer des machines apprenantes.

2.a Les précurseurs

Ramon Llull (1232-1315), philosophe et théologien majorquin, est considéré comme un précurseur de l'IA avec ses machines conceptuelles: telles des règles à calcul souvent circulaires, elles devaient permettre en déplaçant des figures associées à des propositions d'en déduire leur vérité ou fausseté.

Pour René Descartes (1596-1650) « l'animal n'est rien d'autre qu'une machine perfectionnée, une horloge, composée de pièces mécaniques et de ressorts »[3]. Sa théorie de l'animal-machine en fait un des précurseurs de la cybernétique selon Warren Mc Culloch.

Un siècle plus tard, le philosophe et médecin Julien Offray de La Mettrie (1709-1751) alla encore plus loin que Descartes dans la doctrine du *mécanisme* avec son ouvrage intitulé «L'homme machine » (1748).

Deux siècles furent nécessaires pour passer de l'utopie à la réalité avec l'avènement des ordinateurs.

2.b Biomimétisme et idées fausses

Si les premiers réseaux de neurones issus des travaux de McCulloch et Pitts s'inspiraient de ceux du cerveau humain, les algorithmes actuels en sont bien éloignés et il ne faut pas se laisser induire en erreur par un vocabulaire anthropomorphique [4]. Les réseaux de neurones utilisés en IA ne ressemblent pas plus à ceux du cerveau que les avions ne volent en battant des ailes comme des oiseaux.

Rappelons que le cerveau humain comporte environ 86 milliards de neurones connectés par plus de 10 000 milliards de synapses par cm^3 ! [5]. Même si notre cerveau consomme 20% de l'énergie du corps humain, on estime sa puissance à 20 watts, incomparablement moins que celle des machines dédiées à l'IA.

L'apprentissage par des machines est également bien différent de celui des êtres vivants : il faut entrainer un algorithme de reconnaissance d'images (par exemple reconnaitre des chats) avec bien plus d'exemples qu'un enfant ne pourra en voir, alors qu'il suffit à cet enfant de n'en voir que quelques uns pour pouvoir les reconnaitre tous. « Le mécanisme d'apprentissage chez l'enfant est

[3] Lettre au marquis de Newcastle, 23 novembre 1646

[4] « D'abord, petit rappel, l'intelligence artificielle ce sont des programmes informatiques dont l'architecture est calquée sur le fonctionnement du cerveau humain. » sic ! https://www.rtl.fr/actu/debats-societe/mac-lesggy-vous-explique-comment-les-intelligences-artificielles-sont-capables-de-lire-nos-pensees-7900264266 14 mai 2023

[5] https://www.futura-sciences.com/sante/actualites/biologie-votre-cerveau-15-chiffres-cles-51904/

très perfectionné, car il intègre des dimensions émotionnelles, cognitives, tactiles et des connexions innées, à la suite de milliers d'années d'évolution » [6].

Les algorithmes de l'IA sont capables de performances extraordinaires : sont-ils *intelligents*? Tout dépend du sens que l'on donne à ce mot. En tous cas l'intelligence est sûrement du côté des concepteurs des algorithmes.

3. Une histoire mouvementée

Le développement de l'IA est passé par des alternances de périodes d'enthousiasme, suivies de déceptions et de reculs dénommés « hivers ».

3.a Les réseaux de neurones et la naissance de l'informatique

Les réseaux de neurones et le connexionnisme qui s'en est suivi, ont une histoire intimement liée à celle de l'informatique. L'article fondateur[7] en 1943 du neurophysiologiste Warren Sturgis McCulloch (1898-1969) et du logicien William Pitts (1923-1969) décrivant mathématiquement le fonctionnement des neurones se présente explicitement comme une des premières tentatives d'appliquer la calculabilité définie par Alan Turing et Alonzo Church. Il a attiré l'attention de l'illustre mathématicien John von Neumann, le père de l'architecture des ordinateurs, qui en a fait son modèle dans sa description en 1945 [8] du futur ordinateur binaire EDVAC (Electronic Discrete Variable Automatic Computer). Pour réaliser cette machine, von Neumann suggérait d'utiliser des tubes à vide et de les relier selon le schéma des réseaux de McCulloch et Pitts. L'EDVAC ne sera opérationnel qu'en 1951.

McCulloch, Pitts et von Neumann formèrent autour de Norbert Wiener (1894-1964) le groupe des cybernéticiens. Le terme de *cybernétique* inventé par Wiener désigne une « théorie de la commande et de la communication, aussi bien chez l'animal que dans la machine » selon le titre de son ouvrage majeur publié en 1948 simultanément au MIT et à Paris. Si le mot cybernétique n'est plus guère utilisé en Occident, le préfixe *cyber* l'est énormément : cyber-défense, cyber-criminalité, cyber-sécurité etc.

3.b Succès et échecs du connexionnisme

Le psychologue Frank Rosenblatt (1928-1971) est à l'origine du *perceptron*, un modèle de réseau de neurones et aussi une machine bien réelle le *Mark 1 Perceptron* mise en service en 1958, utilisant ce modèle pour effectuer de la reconnaissance d'images. On peut voir au Smithsonian Institute à Washington cette machine de 5 tonnes basée sur un ordinateur IBM 704, qui possédait une couche d'entrée de 400 cellules photoélectriques simulant une rétine et des rangées de potentiomètres pour faire varier les poids (ou paramètres) du réseau.

On s'aperçut rapidement que les réseaux de neurones les plus simples, basés sur des fonctions linéaires à seuil, étaient incapables de reconnaitre certaines formes ou fonctions, en particulier le *ou exclusif* et se limitait à des cas linéairement séparables. Mal interprété, le livre de Minsky et Papert[9], fut utilisé par ceux qui ne l'avaient pas lu pour déconsidérer les réseaux de neurones ce qui amena à

[6] Aimetti, J.P., Coppet, O., Saporta, G. (2022) *Manifeste pour une intelligence artificielle comprise et responsable*, Cent Mille Milliards, Paris

[7] McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*, 115-133.

[8] von Neumann, J. (1945). *First Draft of a Report on the EDVAC*. University of Pennsylvania

[9] Minsky, M. et Papert, S. (1969). *Perceptrons : an Introduction to Computational Geometry.* MIT Press

un assèchement des crédits et ce qu'il est convenu d'appeler le « premier hiver » de l'IA qui dura une dizaine d'années.

Le 11 juillet 1971, le jour même de ses 43 ans, Frank Rosenblatt périt d'un accident de bateau en solitaire dans la baie de Chesapeake.

3.c Le boom des années 80 et le « deuxième hiver » (1987-1993)

Les années 80 virent d'une part le retour en force de l'IA symbolique avec les systèmes experts dédiés à des applications spécifiques comme le diagnostic médical, et d'autre part une renaissance du connexionnisme avec l'algorithme de rétropropagation du gradient et des succès notables en reconnaissance des caractères.

L'intérêt pour les réseaux de neurones déclina cependant assez rapidement, d'une part en raison des faibles capacités de stockage et de l'absence de bases d'apprentissage suffisantes et d'autre part par leur difficulté de mise en œuvre et d'automatisation. Le réglage des modèles demandait une expertise humaine importante avec des tours de main que ne possédaient que de rares spécialistes.

Ce fut alors le grand moment du développement de l'apprentissage machine (*Machine Learning*) à partir des théories du statisticien russe Vladimir Vapnik, avec en particulier les SVM, une famille de méthodes de classification plus séduisantes sur le plan de la rigueur mathématique, développées avec Isabelle Guyon.

Les systèmes experts disparurent également des radars, n'ayant pas tenu selon certains les promesses excessives qui avaient pu être faites, ou aussi peut-être parce que leurs applications étaient devenues routinières, donc invisibles.

Le terme d'IA disparait du vocabulaire scientifique pendant au moins dix ans, ou reste cantonné à l'IA symbolique, mais reste présent en science-fiction.

3.d Le retour en fanfare des réseaux de neurones : le *deep learning*

Les réseaux de neurones n'avaient plus bonne presse mais des chercheurs y croyaient toujours en les perfectionnant avec de nouvelles architectures, comme les réseaux convolutifs. Les trois futurs lauréats du prix Turing 2018 (le « Nobel » de l'informatique) Yoshua Bengio, Geoffrey Hinton et Yann LeCun montèrent ainsi à la fin des années 90 ce qu'ils nommèrent eux-mêmes la « conspiration du Deep Learning ». Ce changement de nom associé à des succès spectaculaires dans les compétitions de reconnaissance d'images ou de textes remit les réseaux de neurones au premier plan de l'IA. Ces succès furent rendus possibles par les progrès inouïs de l'informatique avec la constitution d'immenses bases d'apprentissage et des gains de puissance obtenus en utilisant des puces spécifiques, les *Graphics Processing Units* ou GPU, conçus initialement pour les jeux vidéo, mais bien adaptés à certains calculs mathématiques.

3.e L'IA en France

Les chercheurs et praticiens français en IA sont nombreux et jouissent d'une forte réputation. On relève dès les années 80 des publications suivies d'échanges réguliers avec des chercheurs américains. Parmi les pionniers qui ont formé des générations d'étudiants, on trouve des physiciens comme Gérard Dreyfus à l'ESPCI, Werner Krauth et Marc Mézard à l'École Normale Supérieure et des mathématiciens comme Françoise Fogelman- Soulié à l'université d'Orsay qui ont mené de front recherche, formation et création d'entreprises que l'on n'appelait pas encore des start-ups, comme Mimetics, Netral, Neuristics.

Parmi leurs étudiants, Léon Bottou, Isabelle Guyon et Yann LeCun pour n'en citer que trois, sont devenus des vedettes internationales. L.Bottou déclarait récemment : «Les chercheurs français ont toujours été présents dans l'IA, même quand le sujet n'était pas en vogue. C'est probablement lié à la liberté dont ils bénéficient».

4. Limites techniques

Les méthodes d'apprentissage ne sont ni parfaites ni aussi automatiques que l'on pourrait le croire et l'intervention d'humains est souvent indispensable.

4.a L'IA n'est pas infaillible

Les algorithmes actuels reposent sur un nombre phénoménal de paramètres, plusieurs millions dans certains cas , estimés sur de grandes quantités de données, ce qui peut conduire à des résultats instables. Des travaux de recherche récents[10] ont montré que la modification d'un seul pixel dans une image pouvait modifier totalement la prévision de sa catégorie: une image de cheval est ainsi classée comme étant celle d'une grenouille avec une confiance de 99,9% ! Cette instabilité peut être une source de risque de cyber-attaques par des individus malveillants.

L'insuffisance de certaines bases de données d'apprentissage peut conduire à des résultats choquants comme cet exemple très connu d'un algorithme de Google, corrigé depuis, qui confondait des images d'afro-américains avec des images de gorilles.

4.b Annotation et étiquetage humain

La reconnaissance d'images nécessite d'utiliser de grandes bases de données d'apprentissage où les catégories des images (chiens, chats, avions etc.) sont connues. Pour effectuer cet étiquetage, on fait intervenir des humains. Les entreprises de l'IA générative utilisent ainsi de nombreux travailleurs pour améliorer les performances de leurs outils en éliminant des contenus inappropriés ou toxiques.

Ces « travailleurs du clic » exerçant en général dans des pays à bas salaires, sont les soutiers invisibles mais indispensables de l'IA, dissimulés sous le vocable d'*apprentissage par renforcement à rétroaction humaine*.

5. Défis sociétaux

L'IA est en train de bousculer nos vies et suscite légitimement des inquiétudes par son côté massif et souvent obscur. Gouvernements, organismes supranationaux et entreprises en ont pris conscience et cherchent à promouvoir une « IA de confiance ».

5.a Le droit de comprendre

Les algorithmes d'IA sont souvent des *boîtes noires* auxquelles le citoyen n'a pas accès. Dans la vie quotidienne, nous faisons confiance à de nombreux processus que nous ne comprenons pas : voitures, télévision, smartphones, prévisions météorologiques et bien d'autres encore. Mais lorsque certaines décisions ont des implications sur notre vie : santé, emploi, argent, etc., le droit à une explication (*transparence*) est nécessaire. Ce droit est consacré par des recommandations d'organisations internationales telles l'OCDE et l'Union Européenne.

[10] Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, *23*(5), 828-841.

En attendant l'adoption du règlement européen sur l'IA (*IA Act*) , de nombreux codes de conduite et déclarations d'éthique ont fleuri depuis 2017 avec la « Déclaration de Montréal pour un développement responsable de l'intelligence artificielle », celle de 2020 du Conseil de l'Europe « Vers une régulation des systèmes d'IA » , la « Recommandation de l'UNESCO sur l'éthique de l'IA » (2021), jusqu'au « Projet de charte des droits de l'IA (*Blueprint for an AI Bill of Rights*) » publié par la Maison Blanche en octobre 2022.

Rendre explicables les algorithmes d'IA est devenu un enjeu et de nombreux travaux sont menés sous le vocable d'XAI pour *eXplainable Artificial Intelligence*. On cherche ainsi à mesurer l'importance des critères ayant abouti à une décision particulière, ou à rendre compréhensible cette décision en en cherchant *a posteriori* une approximation par des règles simples ou des modèles de substitution, d'où un certain regain d'intérêt pour l'IA symbolique. Un courant de pensée opposé, l'*IA interprétable*, consiste à refuser la complexité des boîtes noires en n'utilisant que des algorithmes simples. Ce courant est illustré par Cynthia Rudin, Professeure à Duke University, lauréate en 2019 du prix de « l'IA au service de l'humanité » de 1 million de dollars, décerné par l'Association Américaine pour le Progrès de l'AI (AAAI) .

5.b Biais et équité

Une abondante littérature de dénonciation[11] a mis en évidence des applications non éthiques des algorithmes d'IA conduisant à des discriminations selon le genre (outils d'aide au recrutement), la couleur de peau (prédiction de la récidive des condamnés aux USA) pour ne citer que quelques exemples. Il est devenu courant de parler des *biais* des algorithmes pour évoquer ces applications contraires à l'équité, mais un algorithme prédictif n'est pas inéquitable en soi : il est entraîné pour optimiser la prédiction de la réponse sur les données d'apprentissage. Si l'ensemble d'apprentissage est biaisé car non-représentatif, ou résultant d'une sélection, l'algorithme ne pourra que reproduire ces biais. Les « biais » des algorithmes ne sont souvent ainsi que la reproduction de ceux de décisions antérieures : attribution de prêts à des candidats jugés solvables, propension à condamner des accusés selon des critères ethniques inavoués, etc. [12]

5.c La meilleure et la pire des choses

A côté d'espoirs de progrès partagés en matière de santé (médecine personnalisée), de sécurité (véhicule autonome) et d'optimisation des ressources pour ne citer que quelques domaines, l'IA soulève des craintes tout aussi justifiées. La surveillance de masse par des états autoritaires, la désinformation par ciblage sont déjà des réalités. En échange de services réels, nos données et les traces que nous laissons pour chaque activité numérique font l'objet d'un commerce qui met en cause la vie privée. L'IA est un *business* qui met en jeu des investissements colossaux et les entreprises innovantes sont rapidement absorbées par des firmes de plus en plus puissantes, créant des inégalités économiques.

Des régulations se mettent cependant en place pour éviter les distorsions de marché et les monopoles et les atteintes à la vie privée, tel le RGPD européen qui fait des émules là où l'attendait le moins mais pour des raisons différentes: en Californie et en Chine.

[11] Voir le livre de Cathy O'Neil (2016) *Weapons of Math Destruction,* Crown. Sa traduction en 2018 sous le titre moins provocateur de *Algorithmes : la bombe à retardement,* Les Arènes, est préfacée par Cédric Villani

[12] Saporta, G. (2023). Équité, explicabilité, paradoxes et biais. *Statistique et Société*, 10(3), 13-23.

L'impact sur l'emploi est un sujet très discuté : s'il est clair que des métiers disparaitront comme dans les révolutions industrielles précédentes, et que d'autres apparaitront, quel sera le bilan ? La mondialisation avec le partage Nord-Sud que nous connaissons évoluera t-elle vers plus ou moins d'inégalités ?

5.d Impacts environnementaux

L'IA peut améliorer le fonctionnement de la société par sa recherche de solutions optimales : les applications *smart cities* permettent de fluidifier les trafics, d'économiser de l'énergie en optimisant l'éclairage etc. Mais le fonctionnement de l'IA est énergivore. On a estimé à plus de 4 GWh la puissance nécessaire pour l'entrainement de GPT3, soit une tranche complète de centrale nucléaire pendant quatre heures, sans parler des émissions de CO_2. Un bilan objectif reste à établir.
Ces données sont à rapprocher du fait que les *Data Centers* représentent 2 à 3 % de la consommation électrique mondiale. Le refroidissement des *Data Centers* devient une préoccupation majeure et les industriels sont à la recherche de solutions innovantes : Microsoft en collaboration avec Naval Group (l'héritière des arsenaux français et de la Direction des constructions navales) a ainsi testé l'immersion sous-marine avec le projet Natick.

5.e Formation

Lutter contre l'ignorance est un des meilleurs moyens de faire accepter l'IA dans ce qu'elle a de meilleur et d'en maîtriser ses usages. Former les citoyens et les jeunes à l'IA est un défi immense, alors que le système éducatif peine déjà à enseigner l'informatique à tous. Mais c'est une nécessité qui devra s'appuyer sur des technologies innovantes, dont l'IA fait bien sur partie!