Séance publique du 4 mars 2019

L'intelligence artificielle ? Un domaine de la science informatique

Michel CHEIN
Professeur émérite à l'Université de Montpellier

MOTS CLÉS
Intelligence artificielle (IA), IA symbolique, IA connexionniste, raisonner, apprendre

RÉSUMÉ
On présente les deux grands paradigmes de l'IA à partir de deux exemples simples, un en recherche d'information pour la représentation des connaissances et des raisonnements, et l'autre en reconnaissance de caractères manuscrits pour l'apprentissage statistique par réseau de neurones. La naissance de l'IA est ensuite rappelée, et on insiste sur le fait que le domaine scientifique conçu est un sous-domaine de la science informatique. La polysémie de la locution Intelligence Artificielle est expliquée ainsi que le risque, prévisible, d'un 3ième hiver de l'IA. Enfin, compte tenu des dangers inhérents à certaines applications de l'IA, la conclusion insiste sur l'importance de l'éthique dans ce domaine. Considérant qu'il n'existe pas, du moins aujourd'hui, de définition suffisamment scientifique aussi bien de l'intelligence humaine que d'une "intelligence" abstraite pour pouvoir y répondre autrement que par des croyances ou des conjectures, ce court texte n'abordera pas la question de leur comparaison.

1. D'où je parle ?

Le choix de la locution « intelligence artificielle », génial d'un point de vue marketing, est une malédiction pour les chercheurs travaillant dans ce domaine de l'informatique !

L'Intelligence Artificielle ne serait qu'un monstre, une chimère, stérile, résultat de l'accouplement improbable de la faculté la plus évoluée de l'homme et d'une machine !

Rien d'étonnant à ce que certains pensent que l'IA est impossible cependant que d'autres pensent que l'intelligence humaine n'est pas si intelligente que ça et qu'elle pourrait bien, un jour, être dépassée par celle des machines ! (le fameux point de singularité !)

Dans tous les cas, la contradiction entre intelligence et machine produit une tension qui fait parler, écrire et travailler.

« D'où tu parles toi ? » demandaient, parfois brutalement, aux environs de 68 des étudiants à leurs professeurs médusés. Si la forme de la question était violente son fond était, parfois, justifié. Certains ont appris et retenu la leçon et ils prennent maintenant les devants. Ceci est d'autant plus nécessaire pour moi aujourd'hui que la locution « Intelligence Artificielle » est éminemment polysémique.

Je ne suis pas philosophe (alors que je construis des ontologies), ni psychologue (alors que les réseaux sémantiques font partie de mon quotidien), ni linguiste (alors que

syntaxe, sémantique, pragmatique, synonymie, antonymie, hyperonymie, hyponymie sont des notions que j'utilise régulièrement), ni patron de start-up de la French Tech, ni chercheur dans une GAFAM[1], ni journaliste (alors que ce sont eux qui parlent le plus d'IA), ni etc, …, bref, la liste est longue des capacités que je n'ai pas et qu'il faudrait avoir pour appréhender dans toute sa complexité ce dont la locution « Intelligence Artificielle » est le nom.

Mais alors : quelle est ma légitimité pour parler d'IA ?

Pour simplifier, pendant une vingtaine d'années de 65 à 85 j'ai travaillé en mathématiques appliquées et en algorithmique combinatoire. Cependant – hasard et curiosité – j'ai eu de nombreuses discussions avec les fondateurs de l'IA en France. En 1972, je rejoins à Paris un laboratoire du CNRS composé de 3 équipes : Intelligence Artificielle, dirigée par Jacques Pitrat, le père de l'IA en France, Reconnaissance des Formes, dirigée par Jean-Claude Simon, le « général » de la RF en France, et la 3ème équipe, Théorie des Questionnaires était dirigée par Claude-François Picard, fondateur du laboratoire. En arrivant, je crée une 4 ème équipe « Graphes et optimisation combinatoire ». Le bureau voisin du mien était occupé par Jacques Pitrat et par Jean-Louis Laurière avec qui j'eus de très nombreuses discussions ainsi qu'avec Jean Sallantin, à l'époque jeune chercheur dans l'équipe de J. C. Simon. J'étais intéressé par leurs travaux et, pour être honnête, dubitatif.

Il y avait à cette époque tellement de jugements négatifs sur l'IA, que je fus chargé par le Comité national de la recherche scientifique, au début des années 80, de faire un rapport sur la recherche en IA en France. Après ce rapport – rapidement disparu dans les oubliettes habituelles – et malgré de nombreuses hésitations, je franchis le Rubicon ! Mon premier article en IA date de 1986 et le dernier, à ce jour, de 2017, je suis donc resté fidèle et c'est le domaine dans lequel j'ai travaillé le plus longtemps.

En arrivant à Montpellier j'ai retrouvé François Grémy, professeur d'épidémiologie à la faculté de médecine qui m'a aidé à quitter le terrain solide des mathématiques et de l'algorithmique pour les sables mouvants, à l'époque, des systèmes experts en médecine ! Deux Pierre, Pierre Dujols, polytechnicien et médecin, professeur à la faculté de médecine et Pierre Aubas, spécialiste de l'asthme, ainsi qu'un linguiste, Christian Baylon, m'ont entrainé dans le projet européen Ménélas dont le but était d'écrire un système de compréhension de lettres de sortie d'un service hospitalier de cardiologie. Plus tard, avec les deux Pierre, nous avons monté le Projet européen, Therapy Adviser in Oncology, dont l'objectif était d'écrire un programme d'aide au suivi de traitement en cancérologie.

Nous sortions du 2 ème hiver de l'IA, tout le monde voulait utiliser l'IA.

Avec F. Grémy, J. Sallantin et le soutien de Georges Frèche, nous avons organisé à Montpellier en 1986 le premier congrès international en IA, médecine et biologie AI BIOMED. Nous avons créé le groupement scientifique du CNRS DIALR (Développement de l'IA en Languedoc-Roussillon), dont le responsable fut J. Sallantin, qui développa des activités en médecine, en chimie avec B. Castro et C. Laurenço, en biologie avec O. Gascuel, en traitement de la langue naturelle avec J. Chauché, en jeu de Go avec H. Dicky, en droit avec G. Sabatier, etc.

En ce qui concerne l'IA, j'ai exercé, pendant ces années, en plus de mes recherches et de la direction de mon équipe, un certain nombre de fonctions (président du comité de programme de la Conférence nationale RFIA, membre de la direction du Programme de Recherches Coordonnées en IA, co-rédacteur de la Revue d'IA, …), j'ai

[1] Lorsque le nom complet d'un acronyme n'apparait pas dans ce texte on peut aisément le trouver sur le Web.

été invité en Allemagne, au Brésil, en Grande-Bretagne, aux USA et au Canada où je passais une année sabbatique en 1989-90. Tout ceci est banal, c'est l'activité normale de tout universitaire, cela prouve simplement que j'ai un certain âge … un peu de nostalgie, de l'agacement à la lecture de certains articles, et de l'exaspération face aux spécialistes auto-proclamés qui seraient bien incapables de comprendre, ne parlons pas de réussir, les examens de nos étudiants !

Depuis le milieu des années 90 la vie de l'IA à Montpellier est un long fleuve … vigoureux ! De nombreuses équipes travaillent en IA (contraintes, apprentissage, systèmes multi-agents, argumentation, préférences, ontologies, web sémantique, langue naturelle, représentation de connaissances et de raisonnements, etc.) et le Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier est depuis longtemps un laboratoire reconnu en IA au niveau international.

Je me limiterai au seul aspect dont je peux parler sans dire trop de bêtises et qui est celui d'une discipline scientifique, sous-domaine de la science informatique, qui s'appelle Intelligence Artificielle depuis le 31 aout 1955 mais qui était née quelques années auparavant.

Dans le paragraphe suivant les deux grands paradigmes de l'IA, l'IA symbolique et l'IA connexionniste sont présentés.

Cette unique partie un peu technique, est suivie d'une brève présentation de la naissance de l'IA.

Les deux derniers paragraphes concernent la situation présente et quelques inquiétudes face à certains usages de techniques d'IA.

2. Raisonner et Apprendre

2.1. Représentation de connaissances et de raisonnements

Pour illustrer l'IA symbolique, je présenterai un problème simple. Considérons cette photo avec quelques informations associées.

Il y a deux enfants, une fille et un garçon sur une plage de sable. Le garçon tient une pelle rouge, la fille un saut jaune. Ils sont très concentrés. La fille a une jupe et un tee-shirt blanc à points noirs, le garçon porte un short et un tee-shirt blanc. À côté d'eux il y a d'autres jouets, etc.

Si on interroge une base de photos pour trouver des photos avec des enfants jouant au bord de la mer, cette photo, accompagnée des informations précédentes, ne sera pas trouvée.

En effet, dans les informations associées à la photo il n'y a ni enfant, ni bord de mer, ni jeu. Pour répondre à la question il faut savoir qu'une fille ou un garçon est une sorte d'enfant, il faut aussi avoir des règles. Par exemple, une règle reliant la relation d'appartenance à une catégorie plus générale : si X est un élément de Y, et si Y est une sorte de Z alors X est un élément de Z. Cette règle permet de déduire, par exemple, que Julie est un enfant puisque Julie est une fille et qu'une fille est une sorte d'enfant (ceci en remplaçant dans la règle X → Julie, Y → Fille, Z → Enfant). De même, on peut

obtenir que Marc est un enfant. Ceci n'est pas suffisant pour répondre à la question, *on cherche des enfants au bord de la mer* or on sait seulement qu'ils sont sur une plage de sable.

On peut penser que si un enfant est sur une plage de sable alors il est au bord de la mer, une telle connaissance peut être formalisée par la règle :

Si X est un Enfant et X est sur une plage de sable alors X est au bord de la Mer. D'autres règles incertaines sont nécessaires, par exemple, que si un enfant tient une pelle alors il joue (et ne travaille pas …).

Voici un autre exemple illustrant les difficultés de la formalisation de raisonnements de « sens commun » [1] : « Moins de 3% de nos déplacements se font à vélo alors que les 3/4 des trajets font moins de 5 kms. Résultat : ça creuse deux fois le trou de la sécurité sociale. » Quelles connaissances et quels raisonnements permettraient de justifier cette conclusion ? Il faudrait avoir une chaine de déductions :

aller à vélo → faire du sport → améliore la santé → diminue les soins → diminue les charges de santé.

Il faut aussi supposer que si on ne va pas à vélo on prend un engin polluant.

pas de vélo → engin polluant → pollution → maladie → augmentation des charges de santé.

Dans ce type de connaissances et de raisonnements de « sens commun » le contexte est important (Fille prend des sens différents suivant le contexte, Méduse est une petite fille de Gaïa, lorsqu'on dit d'une personne « c'est une vraie jeune fille » elle peut être très âgée, etc.), et il est difficile pour un programme de jouer sur les contextes et les niveaux d'abstraction.

S'il s'agit de trouver des photos pour illustrer un article et si le système se trompe ce n'est pas très grave. Le programme n'est alors qu'un outil d'aide qui filtre des réponses, il peut donner des idées à un auteur ou lui permettre de gagner du temps. Par contre, si, comme pour certains systèmes, il s'agit de déterminer si un adolescent risque de faire une tentative de suicide, ou si une personne risque de faire un délit ou si un islamiste radicalisé risque de passer à l'acte, les conséquences d'une erreur du système peuvent être plus importantes.

On peut distinguer, lorsqu'on construit des systèmes à base de connaissances, trois grands types de connaissances des plus compliquées aux plus simples : des connaissances de sens commun, dont on vient de donner deux exemples et qui sont des connaissances essentiellement langagières, des connaissances scientifiques (utilisant des concepts précis et des modèles mathématiques), des connaissances techniques (utilisant des règles et des terminologies strictes). Dans tous les cas, il s'agit de modéliser et de formaliser connaissances et raisonnements par des logiques permettant de représenter différents types de connaissances (incomplètes, incertaines, des exceptions, des préférences, le temps, l'espace, …) et de faire différents types de raisonnement (déductif, inductif, analogique, causal, argumentatif, …).

On entend habituellement par *connaissance* une relation entre quelqu'un qui sait et une proposition concernant l'état du monde qui est ainsi et pas autrement. Le terme de représentation de connaissances est mal choisi car il ne s'agit pas de re-présenter mais de modéliser informatiquement, c'est-à-dire de choisir des aspects que l'on modélisera par une chaine de caractères, donc très simplement, et que l'on manipulera ensuite d'une certaine manière. Pour modéliser on simplifie, on s'intéresse à un cas idéal qui n'existe pas, donc on supprime, on mutile, car dans de nombreuses situations il est impossible de déterminer un ensemble complet de paramètres, le réel n'est pas réductible à un ensemble fini de symboles.

Comment représenter des sentiments : l'amitié serait représentée par un nombre de likes ? Comment représenter des concepts tels que la beauté ou l'équité par quelques

symboles ? Prétendre re-présenter une personne par un ensemble de chiffres n'est ni prétentieux ni ridicule, c'est dangereux !

Où est l'intelligence dans un tel système ? Elle est dans des humains qui ont fait preuve d'intelligence dans de nombreux domaines pour modéliser connaissances et raisonnements, mais alors, que fait la machine ? Bof, elle calcule …

Pour raisonner il faut avoir des connaissances. Construire manuellement des connaissances explicites peut être compliqué voire irréalisable. Il est par exemple impossible d'envisager l'annotation manuelle de toutes les photos disponibles sur le web. On va donc construire des programmes qui apprennent.

2. 2 Apprentissage automatique

On présente le principe des méthodes d'apprentissage supervisé par réseaux de neurones à partir du problème de la reconnaissance de chiffres manuscrits. Ce problème, facile pour un humain, est resté longtemps difficile pour un programme (les premiers captcha étaient composés de lettres ou des chiffres). Il suffit d'essayer d'écrire un programme reconnaissant des 1 manuscrits pour s'en convaincre. Un tel 1 est parfois composé de deux segments, le plus long, proche de la verticale, a son extrémité supérieure voisine de celle du segment le plus court. Ces deux segments, dans un rapport approximatif de 1 à 3, font un angle d'environ 45 degrés ; ils peuvent être complétés par une base horizontale, de longueur proche de celle du segment le plus court, accueillant en un point proche de son milieu l'extrémité inférieure du segment le plus long. Certains 1 manuscrits sont aussi composés d'un unique segment, etc. Bref, les nombreux cas, les exceptions, les approximations font que ce qui est simple pour un humain est difficile pour un programme, sauf s'il est capable d'apprendre.

L'apprentissage supervisé consiste à considérer de nombreux exemples, par exemple des 1, et à améliorer les paramètres du programme jusqu'à ce qu'il reconnaisse les 1. Les récents succès pour résoudre ce type de problèmes ont été obtenus avec des réseaux de neurones artificiels.

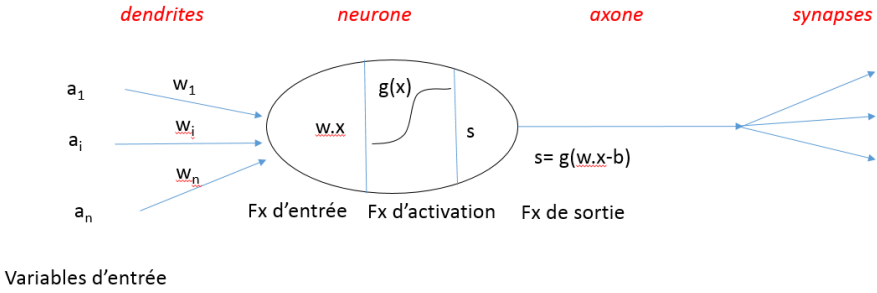

Figure 1 : Un neurone artificiel

Un neurone artificiel simple fonctionne comme un mécanisme d'admission à un examen. On a une note par matière, chaque matière est affectée d'un coefficient, et une personne est reçue à l'examen si la somme des notes pondérées dépasse un seuil sinon elle est collée. Formellement, un neurone artificiel a n variables réelles (les matières, a_1, a_2, …, a_n), n poids (les coefficients, w_1, w_2, …, w_n), une fonction d'entrée (somme pondérée des notes, $\mathbf{w}.\mathbf{x} = w_1.a_1 + w_2.a_2 + \ldots + w_n.a_n$), une fonction d'activation

(la fonction g(x) = **w.x** – *b*, où *b* est le total qu'il faut avoir pour être reçu), une fonction de sortie (1 lorsque **w.x** – $b \geq 0$ et 0 autrement)[2].

Un réseau de neurones comme celui de la figure 2 est composé de tels neurones, les sorties d'un neurone pouvant être reliées aux entrées d'autres neurones. Dans l'exemple de la reconnaissance du chiffre *1*, on donne en entrées les valeurs des pixels d'une image d'un chiffre *1* et on modifie les poids et les biais jusqu'à ce que pour un chiffre, par ex. 1, à force de fournir de nouveaux exemples de *1*, le réseau ait une sortie toujours supérieure à 0. 5 sur le neurone de sortie correspondant au chiffre *1* et pour les neuf autres neurones de sortie toujours une sortie inférieure à 0. 5.

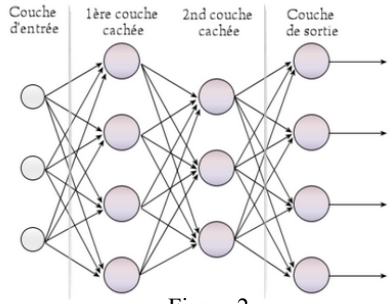

Figure 2

La phase d'apprentissage est terminée lorsque le réseau donne les bons résultats sur les exemples. Le réseau

est alors évalué sur un autre échantillon de données et si les résultats sont satisfaisants (s'il donne suffisamment souvent la bonne sortie) on considère que le réseau a appris, on peut alors l'utiliser pour reconnaitre des données quelconques (cf. [2] pour plus de détails).

Les banques, pour reconnaitre les chèques, et les postes, pour reconnaitre les adresses, ont été les premiers grands utilisateurs commerciaux de ces techniques qui ont depuis été utilisées avec succès dans de nombreux domaines (reconnaissance d'images, de sons, de mots, dans des jeux, …) Trois raisons expliquent ces nouveaux succès, premièrement, internet, les réseaux sociaux, les data centers, bref l'informatisation de la société permet d'avoir de très nombreux exemples, deuxièmement, les machines qui sont de plus en plus rapides permettent de traiter ces grandes masses de données, troisièmement, des algorithmes efficaces ont été inventés. Nous sommes bien ici au cœur de l'informatique dont nous retrouvons les trois ingrédients principaux : des données, des méthodes de calcul et des machines pour supporter ces calculs résolvant un problème particulier.

On vient d'expliquer très succinctement, *comment ça marche*, ce qu'est un système apprenant à reconnaitre un chiffre manuscrit, ou un chat, ou un bateau de pêche, ou une tumeur cancéreuse ou etc. *Pourquoi ça marche ?* est une des grandes questions ouvertes concernant ce type de systèmes. La DARPA aux USA finance de nombreux projets de recherche sur le sujet et une chaire Sciences des données vient d'être créée au Collège de France dont le titulaire est Stéphane Mallat [3].

3. La naissance de l'IA

Pour comprendre un concept il faut en faire la généalogie, connaitre et comprendre son histoire, voici quelques éléments concernant la naissance de l'IA.

[2] La fonction discrète 0-1 n'étant pas dérivable on utilise pour fonction d'activation des fonctions plus compliquées, comme la fonction sigmoïde $\sigma(x) = 1/ (1 + e^{-x})$, la fonction de sortie étant $\sigma(w. x - b)$. Il serait plus correct de parler d'un neurone *mathématique* au lieu d'artificiel, puisqu'un tel neurone n'est qu'un modèle mathématique simpliste d'un neurone biologique.

J. McCarthy, M. Minsky, N. Rochester et C. Shannon demandèrent, le 2 septembre 1955, à la *Rockfeller Foundation* de financer un séminaire pour 10 personnes. Ils obtinrent 7. 500 $, les 10 furent 11 et il y eut une quarantaine de participants au séminaire organisé à Dartmouth l'été suivant.

Qui étaient ces 11 et que faisaient-ils *à l'époque* **?**

- J. McCarthy 1927-2011 mathématicien, Assistant Professor au MIT ;
- M. Minsky 1927-2016 mathématicien, physicien, Harvard Junior Fellow ;
- C. Shannon 1916-2001 mathématicien aux Bell Labs. Dans son master en 1937 il avait utilisé l'algèbre de Boole pour décrire l'aspect fonctionnel des circuits électriques (relais), il avait fait de la cryptographie pendant la guerre, et en 1948 il avait publié son article fondateur de la théorie de l'information ;
- N. Rochester 1919-2001 ingénieur à IBM responsable du groupe pattern recognition, information theory and switching circuit theory, concepteur de l'IBM-701 ;
- J. Bigelow 1913-2003 chercheur à The Institute for Advanced Study à Princeton, collaborateur de von Neumann, concepteur de l'ordinateur IAS prototype de IBM-701, coauteur, avec Wiener, and Arturo Rosenblueth de l'article, "Behavior, Purpose and Teleology, " article fondateur de la cybernétique ;
- J. Holland 1929-2015, Master de math en 1954 doctorant à l'U. du Michigan à Ann Arbor ;
- D. Mackay 1922-1987 physicien et neuroscientifique, au King's College ;
- A. Newell 1927-1992 mathématicien, en thèse avec H. Simon à Carnegie Mellon University, il construit avec Simon et Shaw The Logic Theorist en 1955-56 qui démontre 38 des 52 premiers théorèmes des Principia Mathematica ;
- O. Selfridge 1926-2008 mathématicien, collaborateur de N. Wiener au MIT ;
- H. Simon 1916-2001, économiste, Professeur à Carnegie Mellon University, travaille sur l'aide à la décision introduit la rationalité procédurale qui lui vaudra le prix d'économie de la banque de Suède ;
- R. Solomonoff 1926-2009, mathématicien.

On peut remarquer : une très grande majorité de mathématiciens et deux générations, les xeux animateurs principaux n'ont même pas 30 ans, alors que les plus âgés C. Shannon, J. Bigelow, H. Simon, ont déjà une carrière bien remplie. On peut aussi se demander où est A. Turing ? Né en 1912, mathématicien génial, inventeur de l'informatique, animateur de l'équipe qui pendant la guerre avait découvert le fonctionnement d'ENIGMA, la machine qui servait à coder les messages de l'armée allemande, sauvant ainsi des centaines de milliers de vies, se suicida le 7 juin 1954, suite à une gynécomastie due à sa condamnation à la « castration chimique » (cette année-là des milliers d'autres homosexuels anglais durent subir comme lui des injections d'anti-androgènes).

McCarthy, inventeur de l'expression Artificial Intelligence, dit que c'est Turing qui, lors d'une conférence en 1947, proposa le programme de recherche qu'ils reprirent à Dartmouth [4]. Difficile de trouver une trace de cette conférence mais Turing avait publié en 1950, 6 ans avant Dartmouth, l'article fondateur de l'IA [5].

Dans cet article Turing remplace la question *une machine peut-elle penser ?* (qui ne mérite pas une discussion, écrit-il) par *une machine peut-elle agir intelligemment ?* et il propose deux exemples de comportement : un jeu (le jeu de l'imitation ou le jeu d'échecs) et l'apprentissage de la langue. Dans cet article Turing considère *machine* et programme comme des synonymes (rappelons qu'une machine de Turing est un modèle formel d'un programme et une machine de Turing universelle celui d'un ordinateur).

Le jeu de l'imitation consiste à remplacer, dans un jeu à trois personnes, l'un des protagonistes par un programme et à évaluer si ce programme jouerait aussi bien que

la personne qu'il remplace. Plus précisément, les joueurs sont un homme (A), une femme (B) et un interrogateur (C). C ne voit ni A ni B et ne peut communiquer avec eux que via un clavier. En posant des questions à l'un ou l'autre des joueurs, C doit déterminer qui est l'homme et qui est la femme. A cherche à tromper C, alors que B essaie de l'aider. Turing propose alors le problème suivant : est-il possible d'écrire un programme remplaçant A et jouant aussi bien que A ? Et il propose la conjecture suivante : « Je crois que dans une cinquantaine d'années il sera possible de faire jouer des ordinateurs, ayant une mémoire de 10^9, au jeu de l'imitation de telle sorte qu'un interrogateur moyen n'aura pas plus de 70 % de chance de faire la bonne identification après 5 minutes. »

Comme Turing conjecture qu'il faudrait 50 ans à 60 programmeurs pour programmer ce jeu … il propose de faire de l'apprentissage machine en décomposant le problème en deux : *The child programme* et *The education process*. Et il décompose à nouveau *The child programme* en deux cas extrêmes ; un programme très simple ne comprenant que des principes généraux ou un système à base de connaissances constitué de faits, de théorèmes, de conjectures, de dires d'experts et de règles d'inférence. Pour *The education process* il propose l'approche supervisée rapidement décrite en **2. 2**.

> « Nous pouvons espérer que les machines pourront finalement concurrencer les hommes in all purely intellectual fields. » écrit Turing dans la conclusion de son article et il propose à nouveau d'essayer de programmer « une activité très abstraite comme jouer aux échecs » et « apprendre à comprendre et à parler anglais ».

Que retenir ?

Que le programme de Dartmouth : « Nous essaierons de trouver comment une machine pourrait utiliser la langue, construire des abstractions et des concepts, résoudre des problèmes à l'heure actuelle réservés aux humains, et s'améliorer elles-mêmes. » a été proposé par Turing.

Que l'IA a été créée par des mathématiciens, des physiciens, des ingénieurs, et un économiste (qui a écrit, avec Newell et Shaw, le premier démonstrateur de théorèmes) et que les créateurs de l'informatique (Shannon, Turing, von Neumann) ont participé plus ou moins directement à sa naissance.

Que l'IA a été créée comme un domaine de l'informatique puisqu'il s'agissait de représenter des informations d'un certain type (connaissances et données) et de construire des procédures de manipulation de ces données – i. e. des programmes – en vue d'un certain but (raisonner, apprendre).

On peut aussi noter que le prix Turing, qui est le prix le plus prestigieux en informatique, a été attribué à de nombreux chercheurs en IA, M. Minsky, J. McCarthy, A. Newell, H. Simon, E. Feigenbaum, R. Reddi, J. Pearl…

4. Aujourd'hui

Depuis environ 70 ans l'IA[3], en tant que discipline scientifique avec ses nombreux journaux scientifiques (Journal of AI, JAIR, …, en France RIA) et ses nombreux congrès, (IJCAI, KR&R, AMAS, ECAI, …, en France CNIA), avec ses sociétés savantes (tous les pays économiquement développés ont une telle société savante, AAAI, …, en France AFIA), s'est développée suivant les grands axes : Apprendre, Acquérir, Représenter des connaissances et Raisonner, avec des ruptures, des échecs et des succès.

L'IA faible, qui consiste à construire des programmes simulant des tâches intellectuelles spécifiques, a obtenu de nombreux succès.

[3] On peut oublier que IA est un acronyme …

L'IA forte, qui a pour objectif de construire un programme simulant toutes les capacités cognitives humaines, n'est toujours qu'un projet, ce n'est qu'une idée, une conjecture qui fait travailler certains chercheurs.

Les trois tomes du *Panorama de l'Intelligence Artificielle* coordonnés par P. Marquis, O. Papini et H. Prade proposent une vision assez complète de l'IA aujourd'hui [6].

- Le tome 1, Représentation des connaissances et formalisation des raisonnements, concerne : une histoire de l'émergence de l'IA, les différents cadres de représentation, logiques, quantitatifs, ou graphiques, susceptibles de prendre en compte des informations incomplètes, incertaines, des exceptions, des préférences, le temps, l'espace, des taxonomies, etc., les différents types de raisonnement, la description des actions et de leurs conséquences, argumentation, décision, diagnostic, interaction entre agents, apprentissage, acquisition et validation de bases de connaissances, etc.
- Le contenu du tome 2, Algorithmes pour l'IA, est le suivant : déduction automatique, satisfiabilité en logique des propositions, programmation logique, satisfaction de contraintes, algorithmique des modèles graphiques de représentation, réseaux de type bayésien, algorithmes pour la planification, l'apprentissage symbolique, numérique, supervisé, non supervisé, par renforcement etc.
- Dans le tome 3, L'IA frontières et applications, sont abordées les interfaces de l'IA avec différents champs de recherche de l'informatique : les bases de données, le Web sémantique, la linguistique computationnelle, etc. Des chapitres s'intéressent aussi à des questions philosophiques, en particulier épistémologiques, ainsi qu'à la psychologie cognitive. Un plaidoyer pour la recherche en IA clôt ces trois volumes.

Le terme IA est surchargé de significations, ce qui fait que l'IA peut parfois apparaitre comme une espèce d'auberge espagnole où chacun apporte ce qu'il souhaite. En se limitant à ses usages scientifiques alors qu'il y a d'autres usages commerciaux, politiques, voire idéologiques, on peut constater quatre phénomènes qui peuvent expliquer cette impression :

Un phénomène de *restriction* : l'IA c'est l'apprentissage numérique (c'est l'usage dominant aujourd'hui dans la presse, l'économie et la politique. Pour Cédric Villani l'IA c'est « des grandes masses de données et des statistiques »), à ce phénomène de réduction est associé un phénomène concomitant *d'expansion*, l'IA c'est toute l'informatique plus la robotique, plus l'aide à la décision, plus …

Il y a également un processus *d'évaporation* et ce depuis le début **:** des domaines s'autonomisent ou s'immergent dans l'informatique. Un exemple ancien est celui de la *linguistique computationnelle*, qui a commencé comme un domaine autonome principalement en traduction automatique. Puis, après les critiques de Chomsky, le rapport de Bar Hillel puis celui de l'ALPAC en 1966, les crédits ont été brutalement coupés. Les travaux ont continué en IA, ont repris de l'importance et le domaine s'est autonomisé. Un exemple actuel est celui des *sciences des données*. Le premier cours de Stéphane Mallat, titulaire de la chaire *sciences des données* au collège de France depuis 2017, s'intitulait « L'apprentissage face à la malédiction de la grande dimension » et celui de cette année s'intitule « L'apprentissage par réseaux de neurones profonds ». Ce qui était jusqu'à ces dernières années une partie de l'IA est devenue une science autonome.

On assiste aussi à un phénomène de neutralisation de certains domaines de l'IA par l'informatique, en les absorbant, ces domaines perdent l'étiquette IA. En effet, les chercheurs en IA sont des explorateurs de l'informatique. Ce sont des chercheurs en IA qui sont à l'origine de nombreuses notions devenues banales en informatique : le temps partagé, les interfaces homme-machine, les langages fonctionnels, langages

objets, le backtracking, alpha-beta pruning, garbage collector, les systèmes multi-agents, la programmation logique, la programmation par contraintes, etc. Ce que Nick Bostrom (directeur du *Future Humanity Institute*) décrivait en 2006 de la manière suivante : « Beaucoup d'IA de pointe a filtré dans des applications générales, sans y être officiellement rattachée car dès que quelque chose devient suffisamment utile et commun, on lui retire l'étiquette d'IA », aujourd'hui ce serait plutôt l'inverse ! Quand l'IA a le vent en poupe, comme en ce moment, tout le monde fait de l'IA parce que c'est excitant, à la mode, et c'est aussi pour bénéficier des financements des plans nationaux ou pour conquérir des marchés.

Quid de l'avenir de l'IA ?

L'IA a connu un premier hiver au milieu des années 70. Au début de l'IA, les chercheurs avaient faits des prévisions optimistes (H. Simon : « Les machines seront capables, dans vingt ans, de réaliser tout ce qu'un homme peut faire » ou M. Minsky : « Dans une génération, le problème de créer de l'intelligence artificielle sera pratiquement résolu. ») qui avaient crispé philosophes, linguistes, … et de nombreux informaticiens ! Leurs espoirs furent déçus, les crédits furent restreints voire supprimés. Mais, les recherches continuèrent. Aujourd'hui les chercheurs en IA ne font plus de telles prévisions, ce sont plutôt des non-spécialistes qui font des prévisions hyper-optimistes ou apocalyptiques !

Une dizaine d'années plus tard le semi-échec des systèmes experts, l'échec japonais dans la construction d'un ordinateur de 5° génération et celui d'un centre de recherche industriel européen à Munich, conduisit à un deuxième hiver de l'IA, ceci malgré de nombreux progrès scientifiques : création de PROLOG, de la programmation par contraintes, construction d'interfaces « intelligentes » etc. Une fois de plus ce fut un coup de frein brutal sur les moyens de la recherche.

Et maintenant, « AI Winter is coming ! » comme le prédisent certains ?

En 2016 La Tribune posait la question 2016 l'année de l'IA ?

En mai 2017 le mot numérique fut IA, et La Tribune titrait L'IA va être un énorme tsunami !

En 2018 l'été des sciences dans Le Monde consacra 5 articles à l'IA … était-ce le chant du cygne puisque ce même journal publiait le 28 juin 2018 un article de Caroline Talbot intitulé Les premiers ratés de l'IA ?

Le 1er février 2019 Les Échos signalaient le risque d'une bulle financière. Une bulle spéculative autour de l'IA est-elle en train de se former ? se demandait Maxime Nicolas dans *The Conversation* tandis que Le Monde Informatique publiait la liste des 100 startups les plus prometteuses en IA que la Chine, les USA, le Japon, …, la France annonçaient des plans IA avec des sommes plus ou moins pharamineuses en IA, alors que penser ?

Un 3ème hiver de l'IA est-il à l'horizon ?

En 1969, Minsky et Papert ont publié un livre, *Perceptrons*, dans lequel ils démontraient les limites des réseaux de perceptrons. Résultat : l'essentiel des crédits de recherche est allé vers l'intelligence artificielle symbolique ! Aujourd'hui on assiste à la revanche de l'apprentissage numérique et on peut craindre que l'IA symbolique, la GOFAI (*Good Old-Fashioned Artificial Intelligence*) telle que la nomme H. Levesque [1], en subisse les conséquences, comme cela avait été le cas pour la traduction automatique : des espoirs démesurés, des crédits importants, des limites évidentes, une coupure des crédits, mais comme les chercheurs travaillent même sans beaucoup de moyens et que le problème est important, de nouveaux espoirs sont apparus et un

nouveau cycle a recommencé. Le problème c'est qu'il n'y a pas souvent concordance entre les cycles politiques ou économiques et les cycles scientifiques !

Après les succès spectaculaires de l'apprentissage statistique et la tempête médiatique, il est possible que la folie actuelle ne dure pas ! Les techniques d'apprentissage statistique seront à la base d'outils informatiques standards qui iront enrichir les bibliothèques de programmes et les plugins des plateformes. Mais si on oublie les limites des techniques de l'apprentissage statistique, et que l'on cherche à l'appliquer n'importe comment à n'importe quel domaine, ceci conduira à des échecs et c'est toute l'IA qui en subira les conséquences alors qu'il reste tant de choses à faire !

Voici un exemple de problème, en raisonnement de sens commun, pour lequel l'IA a des progrès à faire. On s'est aperçu que le *test de Turing* (le jeu de l'imitation dans sa version simpliste) conduisait à écrire des programmes consistant à tromper une personne plus que d'apporter des réponses directes et claires aux questions posées, ceci en utilisant jeux de mots, plaisanteries, citations, débordements émotifs, apartés…, pour expliquer des réponses étranges un programme se faisait passer pour un jeune Ukrainien de 13 ans maitrisant mal l'anglais ! Très critiquée la *Loebner Competition* entre programmes prétendant passer le test de Turing s'arrêta en 2016. Mais l'écriture d'un programme de *questions/réponses* en langue naturelle est un problème scientifique qui reste fondamental. Winograd proposa un cadre, que H. Levesque proposa d'appeler schéma de Winograd [1], dont voici quelques exemples.

Considérons l'énoncé suivant : *La statue n'entre pas dans la valise car elle est trop grande*. Si on pose la question qu'est-ce qui est trop grand ? la réponse est : *la statue.*

Avec l'énoncé : *La statue n'entre pas dans la valise car elle est trop petite* et la question *qu'est-ce qui est trop peti*t ? la réponse est : *la valise*.

Un tel schéma consiste en un énoncé et une question avec les contraintes suivantes :

- répondre doit être très facile pour des humains,
- la réponse doit faire appel à des connaissances sur le monde,
- on doit utiliser des raisonnements de bon sens,
- on ne doit pas pouvoir répondre en utilisant des méthodes statistiques (*google-proof*),
- la question consiste à demander quel est le référent du pronom de l'énoncé ?
- deux réponses sont possibles suivant le mot spécial tiré au hasard (petit ou grand)

L'évaluation est faite automatiquement et on ne compte pas de la même façon une bonne et une mauvaise réponse[4]

Voici un autre exemple, les deux formes du mot spécial sont entre crochets :

Énoncé : Paul a essayé de joindre Georges sur son téléphone, mais il [n'a pas réussi / n'a pas répondu].

Question : Qui [n'a pas réussi / n'a pas répondu] ?

Réponses : Paul / Georges

Une compétition existe sur ce sujet depuis 2011 (à laquelle ni IBM ni Google ne participent) dont les résultats montrent bien les difficultés posées par le raisonnement de sens commun (cf. [7] pour un ensemble de schémas de Winograd).

4 Pour N questions sur un domaine la note est : max(0, N-k. NombreRéponsesFausses)/N, où k≥2 est un coefficient pénalisant les mauvaises réponses (par exemple avec k=2 si une réponse sur 2 est fausse la note est 0).

5. Conclusion

Le débat entre H. G. Wells qui croyait que l'espoir de l'humanité résidait dans la création d'un gouvernement mondial supervisé par des ingénieurs et des scientifiques et G. Orwell qui insistait sur les limitations de la science en tant que guide pour les affaires humaines, est toujours d'actualité en ce qui concerne l'IA.

Inutile d'insister sur les aspects positifs de nombreuses applications de l'IA, ceci est très présent dans les médias *(en santé – par exemple dans l'analyse d'images, l'analyse des EEG, le monitoring, la robotique médicale –, en recherche d'informations et recommandation-personnalisation, pour traquer le harcèlement en ligne, calculer des itinéraires en utilisant l'état du trafic, la conduite d'avion, les anti-spams, etc.)*.

Inutile également de dénoncer certaines pratiques bien connues comme : l'hameçonnage personnalisé (c'est-à-dire la détermination de cibles intéressantes et la génération de messages crédibles en fonction des profils de ces cibles) ; les atteintes aux droits de l'homme, en particulier à la vie privée ; tout ce qui conduit à une véritable perversion du débat démocratique, les manipulations politiques par des campagnes de propagande automatisées et personnalisées, en utilisant des données personnelles vendues ou volées (dans la campagne de Donald Trump, la société Cambridge Analytica détermina des États pouvant basculer, puis détermina dans ceux-ci des électeurs incertains, et enfin envoya des messages ciblés, le Brexit aurait été entaché lui aussi par ce type de manipulation) ; la fabrication de « fake news » en utilisant des fausses photos ou la manipulation de vidéos avec des systèmes permettant la synthèse de la parole. Tous ces dangers sont renforcés parce que les techniques de l'IA sont un monopole des géants de l'informatique : les GAFAM sans oublier Alibaba, Baidu, Tencent, Xiaomi … et lorsqu'elles sont aux mains d'un régime autoritaire ou totalitaire *1984* pourrait devenir une réalité comme c'est le risque en Chine où « le crédit social (*social ranking system*) pourrait conduire à un contrôle total de la population.

Il y a tellement d'applications inquiétantes que de nombreux organismes, dans les pays démocratiques, se sont dotés de commission d'éthique. En France l'alliance Allistene, qui regroupe CEA, CNRS, CPU, INRIA, Mines-Télécom, Écoles d'ingénieurs, a créé la CERNA la Commission de réflexion sur l'Éthique de la Recherche en sciences et technologies du Numérique. Tout programme, tout système informatique, rappelle la CERNA dans un rapport sur l'*Éthique de la recherche en apprentissage machine* [8] devrait respecter de nombreuses propriétés : *loyauté* (un système informatique est loyal s'il se comporte comme ses concepteurs le déclarent), *équité* (l'équité d'un système informatique consiste en un traitement juste et équitable des usagers), *transparence* **(**la transparence d'un système signifie qu'un utilisateur peut vérifier son comportement), *traçabilité* (la mise à disposition d'informations sur ses actions suffisamment détaillées pour qu'il soit possible après coup de suivre ses actions), *explicabilité (*le fonctionnement d'un système doit être compris par un utilisateur), *responsabilité* (le donneur d'ordre ou le concepteur étant responsable si le système est mal conçu, l'utilisateur étant responsable s'il a mal utilisé le système), *conformité* (un système numérique doit être conforme à son cahier des charges ce qui signifie que le système est conçu pour effectuer les tâches spécifiées en respectant les contraintes explicitées dans ce cahier).

Toutes ces propriétés devraient être vérifiées ***avant*** que le système soit utilisé, en analysant son code et ses données. Cependant, ces diverses propriétés peuvent être difficiles voire impossibles à réaliser dans des systèmes utilisant des techniques d'IA (en particulier l'explicabilité pour les systèmes d'apprentissage numérique). De plus, il y a une contradiction entre la rapidité des inventions et des innovations techniques et la lenteur de leur évaluation qui nécessite : la compréhension des conséquences, le contrôle

non seulement la conformité d'un programme aux lois mais aussi les propriétés que devraient avoir tout système informatique, ainsi que des procédures démocratiques, nécessairement lentes, afin que les citoyens puissent décider librement de leur destin.

L'urgence, en particulier de conquérir des marchés, a dicté sa loi, de plus, après la crise de 2008, dans de nombreux secteurs les crédits ont été restreints, voire drastiquement supprimés, d'où le remplacement de personnes par des programmes.

Certains ont poussé des cris d'orfraie contre les robots tueurs, ils avaient raison. Le contrôle d'armes automatiques par des personnes n'est pas toujours faisable, le chemin suivi par un drone dans un essaim de drones menant une attaque ne peut pas être contrôlé par des humains, et les conséquences peuvent être dramatiques, mais il n'y a pas que les robots tueurs… voici un extrait de ce même rapport de la CERNA :

« De façon caricaturale, un véhicule autonome qui se trouverait à choisir entre sacrifier

son jeune passager, ou deux enfants imprudents, ou un vieux cycliste en règle, pourrait être programmé selon une éthique de la vertu d'Aristote – ici l'abnégation - s'il sacrifie le passager, selon une éthique déontique de respect du code de la route s'il sacrifie les enfants, et selon une éthique conséquentialiste s'il sacrifie le cycliste – ici en minimisant le nombre d'années de vie perdues.

Le propos n'est pas ici de traiter de telles questions qui relèvent de la société toute entière … »

On peut se demander s'il est éthiquement responsable de continuer à essayer de construire des voitures autonomes, et plus généralement des véhicules autonomes ne circulant pas sur des voies spécialement aménagées. En effet, s'il y a un obstacle Météor, Orlyval, un tramway ou un train n'ont pas à choisir, ils freinent pour éviter le choc ! Il semble inadmissible qu'on puisse demander à un algorithme de choisir les personnes à sacrifier, et donc demander à un informaticien de programmer de tels choix !

Indépendamment des applications facilement critiquables qui viennent d'être citées rapidement, il faut dire un mot sur des systèmes d'aide à la décision qui concernent directement des personnes. Certains de ces systèmes censés n'être que des systèmes d'aide ont tendance à devenir de plus en plus automatisés, que ce soit pour un diagnostic médical, l'obtention d'un emprunt ou d'un contrat d'assurance, le recrutement d'un employé ou l'inscription dans un cursus universitaire, l'évaluation ou la notation de personnes, la détermination de personnes potentiellement dangereuses (risque de passage à l'acte ou de récidive), une condamnation, etc.

Ce genre d'applications pose d'autres questions fondamentales. Dans tous les cas, la construction d'un tel système nécessite de réduire une personne à un ensemble fini de caractères ceci sans qu'elle puisse rendre compte de sa singularité. De plus, même dans les cas où un recours est prévu, même s'il y a une personne responsable de la décision, l'utilisateur du système risque de devenir un simple intermédiaire acceptant la décision du système « je n'y suis pour rien c'est l'informatique ! » entend-on dire souvent. En effet, remettre en cause la décision d'un programme nécessite de comprendre comment le programme fonctionne, et comme, d'une certaine façon, le faire reviendrait à mettre en cause un outil pouvant être important dans un organisme ou une société, seule une personne occupant une position relativement élevée dans l'organigramme pourrait le faire.

Tout ceci est particulièrement critique dans les systèmes d'apprentissage numérique dont les résultats sont non seulement difficilement explicables mais qui, de plus, présentent des biais lorsque la phase d'apprentissage du programme est réalisée sur des données biaisées.

Une chercheuse du MIT*, Joy Buolamwini,* a démontré que les systèmes de reconnaissance du genre d'une personne à partir d'une photo faisaient beaucoup plus

d'erreurs pour le classement des femmes *à la peau sombre* que pour celui des hommes blancs. Par exemple, le système de Microsoft classifiait incorrectement 1% d'hommes blancs et 35% de femmes *à la peau sombre*. Pourquoi ? Parce que leurs données d'apprentissage comprenaient beaucoup plus d'hommes blancs que de femmes noires. Le système de recrutement d'Amazon avait été entrainé sur des données de personnes ayant déjà candidatées à Amazon. La majorité de celles-ci étaient des hommes blancs. Et le résultat fut de sous-noter les femmes et de sur-noter les hommes. Une fois cette analyse rendue publique en 2018, Amazon cessa d'utiliser ce système.

Il existe de nombreuses études bien documentées sur ce sujet, en particulier dans le livre de Cathy O'Neil *Weapons of Math Destruction*, en français *Algorithmes – La bombe à retardement* [9], qui explique, par exemple, pourquoi « les innocents entourés de criminels se font malmener, tandis que les criminels entourés de bons citoyens respectueux des lois passent au travers ; et compte tenu de la forte corrélation entre les situations de pauvreté et le signalement de délits, ce sont les plus modestes qui continuent d'être pris au piège de ces filets numériques ».

Cédric Villani termine ainsi la préface de ce livre : « Elle [Cathy O'Neil] est bien décidée à ne pas baisser les bras, et à faire porter sa voix autant qu'il le faut pour que nous puissions conserver notre humanité. Écoutez-la attentivement. » La lecture du livre de Cathy O'Neil ne demande aucune compétence technique, il faut absolument le lire !

Pour éviter les dangers inhérents à de nombreuses applications de l'IA, et plus généralement de l'informatique, une formation à l'éthique des étudiants, enseignants, ingénieurs, concepteurs et utilisateurs serait nécessaire, ainsi que la création d'un comité national d'éthique pour l'informatique à l'image du comité consultatif national d'éthique pour les sciences de la vie et de la santé, afin que l'ensemble des citoyens puisse participer à des choix essentiels concernant la société dans laquelle nous vivons.

RÉFÉRENCES

[1] H. Levesque, Common Sense, the Turing Test, and the Quest for real AI, The MIT Press, 2017

[2] M. Nielsen, http: //neuralnetworksanddeeplearning. com/

[3] S. Mallat, https : //www. college-de-france. fr/site/stephane-mallat/_course. htm

[4] J. McCarthy, http: //jmc. stanford. edu/articles/whatisai. html

[5] A. Turing, Computing Machinery and Intelligence, Mind, 1950

[6] P. Marquis, O. Papini, H. Prade (éditeurs), Panorama de l'Intelligence Artificielle, Cépaduès Éditions, 2014

[7] Winograd Schemas, https : //cs. nyu. edu/faculty/davise/papers/WinogradSchemas/WSCollection. html

[8] http: //cerna-ethics-allistene. org/digitalAssets/53/53991_cerna___thique_apprentissage. pdf

[9] C. O'Neil, Algorithmes – La bombe à retardement, Les Arênes, 2018