HISTOIRE DE L'IA : PHILOSOPHIE, ÉPISTÉMOLOGIE ET FANTASMES

MARIE DAVID (X96)
coautrice de l'*Intelligence artificielle, la nouvelle barbarie* (éditions du Rocher, 2019) et cofondatrice de Carbometrix

Faire l'histoire de l'intelligence artificielle, c'est plus que faire de l'histoire, c'est toucher à la nature même de l'humain, à la nature de son intelligence. Les progrès réalisés depuis l'invention de l'IA, au lendemain de la dernière guerre, sont à la fois spectaculaires et frustrants dans leurs résultats. L'IA n'est encore qu'au stade de l'enfance.

L'histoire de l'intelligence artificielle est souvent mal connue, peut-être parce qu'elle échappe à la philosophie des sciences et que ses racines remontent à celles de la philosophie occidentale. C'est en effet Aristote qui, définissant le syllogisme, se pose en père de la logique, le terreau de l'intelligence artificielle. Un syllogisme part de prémisses pour s'assurer de la véracité d'un prédicat selon un raisonnement valide. Avec le syllogisme, Aristote tente de doter l'esprit de règles formelles capables de créer de la connaissance vraie. Le « Nul n'entre ici s'il n'est géomètre » de l'Académie annonce une tentation de la philosophie occidentale : les mathématiques, propédeutiques de la philosophie pour Platon, en constitueront toujours en filigrane l'étalon et le modèle invisible. Peut-on donner à la pensée la même rigueur qu'aux démonstrations mathématiques et comment faire ? Et réciproquement la pensée peut-elle être réductible à une formalisation de type mathématique ? Ces questions préoccuperont les philosophes de Descartes à Hobbes ou Leibniz. La logique se formalisera au XIXe siècle avec Boole et Frege, et Russell et Whitehead, avec les *Principia Mathematica*, chercheront à donner une nouvelle fondation – logique celle-là – aux mathématiques.

L'intervention fondamentale de Gödel

Mais, paradoxalement, c'est en pulvérisant l'espoir d'une axiomatisation parfaite des mathématiques que le mathématicien Kurt Gödel créera les conditions de la naissance de l'intelligence artificielle. Gödel démontre en effet que, pour tout système formel non contradictoire contenant l'arithmétique, il existe une proposition vraie non démontrable. Cela implique que les *Principia Mathematica* comportent des propositions indécidables, des propositions que l'on peut admettre comme vraies mais que l'on ne peut ni démontrer, ni réfuter. Ce résultat est révolutionnaire. D'une part il brise l'espoir d'obtenir un système d'axiomes complets et consistants permettant de fonder les mathématiques. Mais d'autre part c'est la technique employée par Gödel qui aura une descendance féconde : Gödel code non seulement les formules par

© lenscap50

→ Statue en ardoise du mathématicien Alan Turing à Bletchley Park Museum, Bletchley, Grande-Bretagne.

des nombres, mais aussi les démonstrations. Un nombre vaut pour sa valeur, mais aussi pour ce à quoi il fait référence, préfigurant l'usage de la mémoire en informatique, où une référence mémoire peut être utilisée pour elle-même mais pointe également vers la valeur qu'elle contient.

Alors advint Turing

Il reste une difficulté : Gödel ne parvient pas à définir précisément ce qu'est un système formel. Comment définir sans ambiguïté la façon dont un tel système fonctionne ? Cette question préoccupe le mathématicien Alan Turing. En effet pour Leibniz, comme pour les logiciens, les règles de la logique restent des règles abstraites, suivies par l'esprit de façon implicite. Turing a l'idée d'utiliser pour une démonstration mathématique l'artefact d'une machine, ce qui met pour la première fois les hommes en face d'une idée révolutionnaire : celle d'un raisonnement effectué par un mécanisme matériel. Turing montre ainsi que, étant donné une machine de Turing et un état des données que lira la machine, il n'existe pas de machine de Turing permettant de prédire si la première machine va s'arrêter ou non. Turing parachève ainsi ce que Gödel avait entamé : la mise en équivalence entre un raisonnement formel et un système mécanique. Tout calcul que nous effectuons en suivant des règles est susceptible d'être implémenté sur une machine. Mais, à partir du moment où étant donné une machine de Turing il n'existe pas de machine de Turing permettant de prédire si elle va s'arrêter ou non se pose la question suivante : si on ne peut pas prévoir de quoi une machine de Turing est capable, de quoi est-elle capable ?

Une autre machine de Turing

La machine de Turing dont on parle ici lit un ruban de papier, divisé en cases, contenant des symboles appartenant à une liste finie. En fonction de ce qu'elle lit et de son programme interne (qui détermine l'action à faire en fonction de l'état de la machine et du symbole imprimé), la machine peut se déplacer sur le ruban, imprimer un symbole dans une case vide, remplacer le symbole d'une case pleine ou changer d'état. Les symboles appartiennent à un alphabet fini mais le ruban, lui, est de taille infinie.

Et McCulloch vint

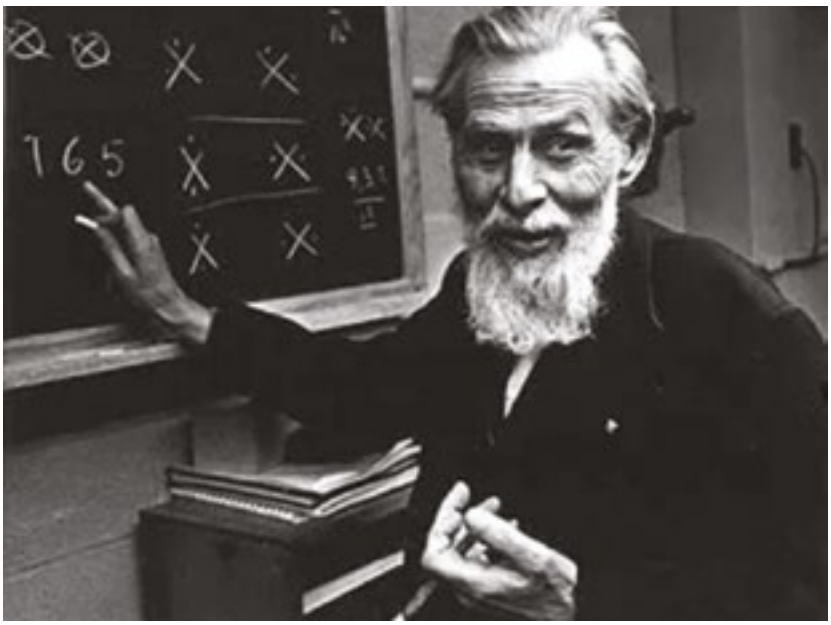
Dans *Computing Machinery and Intelligence*, Turing conclut ainsi qu'il est tout à fait possible qu'une machine fasse preuve d'une intelligence semblable à l'intelligence humaine. Pour Turing, il ne s'agit pas de reproduire, il s'agit de simuler. Le test de Turing appelé *Imitation Game* (mais le philosophe John Searle critiquera le test de Turing par l'expérience de pensée dite de la chambre chinoise, qui montre qu'une machine peut mimer l'intelligence sans pour autant faire preuve d'intelligence), qui sert encore d'étalon à la mesure d'intelligence artificielle, prévoit ainsi qu'un interlocuteur humain ne puisse distinguer si son interlocuteur est ou n'est pas une machine. Ayant pleinement intégré les conséquences des découvertes de Gödel et Turing, le neurologue McCulloch est ainsi persuadé d'une part que la pensée se réduit à des raisonnements logiques, d'autre part que ces raisonnements logiques sont intégrés dans la structure du cerveau. Il y a donc une équivalence entre la fonction et la structure : l'esprit est assimilé à une machine logique et la pensée peut se représenter par des mécanismes qui sont formels, mais qui ont en même temps une base →

matérielle. C'est là encore une idée révolutionnaire. Il y a donc équivalence (au sens mathématique d'une classe d'équivalence) entre les règles formelles de la pensée et la structure logique du cerveau (représentée par les neurones).

Enfin en 1955 naquit l'intelligence artificielle

Les travaux de McCulloch vont ainsi inspirer les sciences cognitives naissantes et les fondateurs de l'intelligence artificielle. En 1955, le *Darmouth Summer Research Project on Artificial Intelligence* signe l'acte de naissance officiel de l'intelligence artificielle, qui se voit à la fois dotée d'un nom et d'un programme : doter les machines de facultés analogues à celles de l'esprit humain, entre autres utiliser le langage, construire des abstractions et des concepts, résoudre des problèmes… Résumer l'histoire d'une discipline aussi complexe comporte nécessairement son lot d'inexactitudes. Pour simplifier, on peut dire que l'intelligence artificielle suivra *grosso modo* deux grandes approches. D'un côté l'approche symbolique reprend l'idée que le raisonnement peut être modélisé par un ensemble de règles logiques et qu'il suffit d'apprendre ces règles à la machine. Cette approche passe par deux étapes qui ont chacune son lot de difficultés : d'une part la machine doit disposer de connaissances sur le monde (par exemple qu'un chien est un mammifère, qu'il a quatre pattes), d'autre part elle doit savoir comment combiner ces connaissances pour arriver à un raisonnement. De l'autre côté l'approche connectiviste, ainsi nommée car elle utilise des fonctions mathématiques inspirées des neurones biologiques. Ces « neurones » renvoient un signal dont l'intensité dépend des *inputs* reçus. À son tour un neurone peut être connecté à d'autres neurones, ce qui permet d'amplifier ou d'atténuer le signal. Les méthodes dites connectivistes reposent sur des méthodes d'apprentissages statistiques très variées – en anglais *machine learning* –, dont les réseaux neuronaux ne sont qu'une sous-famille. Il s'agit d'optimiser les paramètres d'une fonction plus ou moins complexe, afin de minimiser une fonction d'erreur sur les données d'observation. Une fois cette fonction optimale obtenue, on l'applique à de nouvelles observations, ce qui permet de répliquer l'apprentissage fait sur les données d'entraînement. Toute la difficulté d'utilisation de ces modèles reviendra à avoir des données d'entraînement suffisamment détaillées et génériques à la fois, pour que le modèle puisse se généraliser à de nouvelles données.

↓ Warren McCulloch.



Des résultats décevants, une mise en sommeil et un réveil récent

Dans les années 50 et 60, les chercheurs se concentrent ainsi sur des tâches précises : résoudre des énigmes, jouer à des jeux codifiés (comme les échecs ou les dames), reconnaître des images ou des lettres, répondre à des questions. L'approche dite symbolique reste majoritaire, même si les premiers modèles de réseaux de neurones apparaissent, expérimentés notamment par Frank Rosenblatt et son perceptron. Alors qu'elle est florissante durant deux décennies, la recherche en intelligence artificielle ralentit fortement dans les années 70. Les résultats obtenus sont décevants, les financements se tarissent. Après un hiver qui durera trente ans, l'intelligence artificielle revient sur le devant de la scène dans les années 2010, avec un retour en force des modèles dits de *machine learning*. Les progrès des années 2000 à 2010 n'ont en aucun cas constitué une révolution, tout au plus une optimisation d'algorithmes existants. Simplement le développement d'internet a permis la constitution d'immenses bases de données, permettant un meilleur entraînement des modèles. La baisse du coup du calcul, du stockage et de la mémoire a fait le reste, permettant

enfin aux modèles statistiques d'atteindre des performances satisfaisantes, impressionnantes, grâce notamment à l'utilisation de processeurs graphiques (GPU). Les succès de l'intelligence artificielle ont été aussi nombreux que médiatisés, de AlphaGo aux récents modèles GPT-3 ou DALL-E. Mais cela a eu un effet pervers : celui de discréditer complètement l'intelligence artificielle symbolique, par assèchement des financements.

Changer de modèle ?

Malgré les progrès fulgurants obtenus, les algorithmes actuels restent cependant extrêmement loin d'une intelligence artificielle comparable à la pensée humaine. Ils sont performants sur des tâches très spécialisées sur lesquelles ils ont été entraînés, mais il est encore difficile de se passer de supervision humaine dans de nombreux cas (la détection de contenus problématiques sur les réseaux sociaux par exemple). Des attaques célèbres ont également montré les faiblesses de modèles de reconnaissance visuelle. Le chercheur Gary Marcus reproche ainsi à la recherche en intelligence artificielle de se limiter à l'utilisation des modèles de *machine learning*. Pour Gary Marcus, ils sont bons pour faire de l'interpolation et non de l'extrapolation, c'est-à-dire que leur « connaissance » est limitée à l'ensemble des données sur lesquelles ils ont été entraînés. Il montre qu'il est par exemple extrêmement difficile à un tel algorithme de reconnaître la fonction identité, alors qu'un enfant devinera aisément la suite de f(1)=1, f(2)=2, f(3)=3… Par ailleurs, il leur manque un système de représentation, ce que nous appelons le sens commun, un réseau de connaissance permettant de lier des concepts. Ces modèles détectent des corrélations, mais ils ne comprennent pas la façon dont le monde fonctionne de façon fondamentale. Ainsi le modèle GPT-3 donne l'illusion qu'il maîtrise le langage naturel, mais il ne le comprend pas en réalité. Il ne fait que reproduire des occurrences statistiques présentes dans des corpus. Pour Gary Marcus, il faut revenir à l'intelligence artificielle symbolique et construire des modèles hybrides qui seuls pourront dépasser les limites des modèles actuels de *deep learning*.

Changer carrément de paradigme ?

Le philosophe Hubert Dreyfus, lui, critique de façon plus fondamentale la façon dont l'intelligence artificielle a été développée. Pour lui, la théorie de l'intelligence artificielle repose sur une vision dualiste – héritée de Descartes – d'un intellect distinct du corps, dont le fonctionnement serait indépendant de celui du corps. Pour cette tradition philosophique, l'esprit reçoit passivement des éléments de l'extérieur et les trie ensuite. Cette représentation philosophique qui assimile l'esprit à une machine de Turing, qui traiterait des données issues des sens ou de nos représentations internes, pour répondre par des influx nerveux, manque la vraie nature de l'intelligence qui est d'être incarnée. Dreyfus se réfère ainsi au philosophe Merleau-Ponty pour qui la perception est une expérience active, dans laquelle le corps est engagé, et non le résultat d'un traitement d'information séparée de toute incarnation corporelle. Cette conception de l'intelligence est représentée par un courant des sciences cognitives dites de l'*embodied cognition* (dont les représentants sont Francisco Varela, Evan Thompson et Eleanor Rosch), qui conçoit la cognition comme le résultat de l'activité d'un corps en rapport permanent avec son environnement. C'est bien parce que nous sommes avant tout des êtres vivants, en permanence en mouvement et en interaction avec notre environnement physique, que notre intelligence s'est développée. L'esprit devient alors un attribut du corps, son prolongement, en aucun cas une fonctionnalité distincte. En particulier l'intelligence ne peut se développer que dans un corps en contact avec l'extérieur, qui interagit avec son environnement. Il faudrait dans ce cas changer totalement de paradigme de l'intelligence artificielle. L'intelligence artificielle apparaît ainsi comme une science encore dans son enfance, jamais les réalisations de l'intelligence artificielle n'ont été aussi impressionnantes (que l'on pense à DALL-E ou GPT-3), mais jamais l'écart avec la pensée humaine n'a été aussi prononcé, et avec lui l'espoir qu'on parvienne à le combler.

« Jamais les réalisations de l'intelligence artificielle n'ont été aussi impressionnantes, mais jamais l'écart avec la pensée humaine n'a été aussi prononcé. »