Intelligence artificielle

Dominique Cardon

Dans Les petites humanités 2019, pages 385 à 398
Éditions Presses de Sciences Po

ISBN 9782724623659

INTELLIGENCE ARTIFICIELLE

C'est le nouveau fantasme de nos sociétés : des entités artificielles s'apprêtent à vivre parmi nous. Elles traitent des masses inimaginables d'informations, pilotent de grands systèmes techniques, font des hypothèses et arbitrent entre plusieurs stratégies. Mieux, elles nous parlent, ont des émotions et une conscience comme HAL, l'intelligence artificielle imaginée en 1968 par Stanley Kubrick dans le magnifique *2001 l'Odyssée de l'espace*, quand il lui faut décider de sacrifier des humains pour la survie de la mission spatiale. Dans *Her*, le film de Spike Jonze, c'est le même imaginaire qui se donne à voir lorsque Théodore tombe amoureux d'une entité artificielle, Samantha, un *operating system* doté de la voix de Scarlett Johansson. Samantha commence par apprendre les sensations, les comportements et les sentiments humains, puis elle abandonne Théodore afin de rejoindre une communauté regroupant d'autres operating systems, plus riche et plus épanouissante que celle des malheureux humains grâce à qui elle a amorcé son développement supra-humain.

Les représentations fascinantes de la science-fiction sont en grande majorité trompeuses. En réalité, HAL et Samantha ne ressemblent en rien aux machines « intelligentes » d'aujourd'hui, mais les prouesses qu'accomplissent les dernières générations d'algorithmes ont réveillé l'idée ancienne d'intelligence artificielle. Si l'on

voulait être rigoureux, il serait préférable de parler d'apprentissage automatique (*machine learning*) pour désigner la percée technologique que nous connaissons aujourd'hui et qui est en grande partie une conséquence de l'augmentation des capacités de calcul des ordinateurs (permises par les nouvelles cartes graphiques) et de l'accès à de très grands volumes de données numériques. Parmi les différentes techniques d'apprentissage, l'une d'elle, l'apprentissage profond (*deep learning*) à base de réseaux de neurones est, en réalité, le principal vecteur de la réapparition du terme d'« intelligence artificielle » dans le vocabulaire contemporain.

Pour y voir plus clair, commençons par décoder l'histoire tumultueuse et conflictuelle de l'intelligence artificielle. C'est Alan Turing qui, dans un article paru en 1950, évoque le premier l'intelligence des machines et invente un test, le test de Turing ou jeu de l'imitation. Dans ce test, une machine est dite intelligente quand elle parvient à tromper pendant cinq minutes un utilisateur discutant avec elle sans que cet utilisateur se rende compte qu'il échange avec une machine : la machine imite si intelligemment le raisonnement des humains que ceux-ci s'y laissent prendre. Mais le terme d'intelligence artificielle proprement dit est forgé en 1956 par l'informaticien John McCarthy lors d'un fameux workshop réunissant à Dartmouth des chercheurs comme Marvin Minsky et Herbert Simon, qui deviendront avec McCarthy les pères de l'intelligence artificielle.

Durant les années 1960, à l'Université de Stanford, deux laboratoires d'informatique se tiennent tête : le Stanford Artificial Intelligence Lab de John McCarthy et l'Augmentation Research Center de Doug Engelbart, l'auteur de la mère de toutes les démos que nous avons évoquée au chapitre 1. Pendant cette période d'effervescence intellectuelle, McCarthy et Engelbart développent deux conceptions radicalement différentes de la relation entre informatique et société. On se souvient que Doug Engelbart considère les ordinateurs comme une prothèse qui rend les humains plus intelligents (l'informatique nous « augmente »). À l'inverse, John McCarthy promeut l'idée de rendre les ordinateurs intelligents. Il veut fabriquer des automates qui parlent, raisonnent et s'animent. Tel des superhumains, ces robots se substitueront aux hommes dans des activités de plus en plus complexes, prendront des décisions à leur place, il se peut même qu'ils aient un jour une conscience. Cette dernière conception d'une machine autonome a longtemps présidé au développement de l'intelligence artificielle, mais sans grand succès. Jusqu'à présent, l'avènement du numérique a donné raison à Engelbart et à son projet d'augmentation des humains par les machines. L'histoire serait-elle en train de changer de nouveau de sens ? Les machines seraient-elles réellement en train de devenir intelligentes, comme l'imaginaient Minsky et McCarthy ?

L'histoire de l'intelligence artificielle est souvent comparée à un cycle de saisons, où s'enchaînent les promesses et les échecs. Si elle connaît aujourd'hui son troisième printemps, c'est parce qu'elle a déjà traversé

deux hivers. Après une première vague de succès, retombée à la fin des années 1960, elle a suscité un engouement considérable au cours des années 1980 avant de s'effondrer à nouveau au début des années 1990. Chaque vague ressemble en tout point à l'euphorie actuelle : des promesses de machines intelligentes, un débat sur l'automatisation du travail, des titres de journaux s'alarmant de la prise de contrôle de nos sociétés par des consciences artificielles, des financements massifs et des entreprises qui prophétisent que l'intelligence artificielle va changer la face du monde. Cependant, nous aurions tort d'analyser ce qui se déroule aujourd'hui comme un simple copié-collé des deux premières vagues, car entre-temps, la manière de définir l'intelligence de la machine a radicalement changé.

Le projet d'intelligence artificielle des deux premières vagues, celle des années 1960 et celle des années 1980 était d'ordre « symbolique ». À l'instar de HAL, dans *L'Odyssée de l'espace*, les promoteurs de l'intelligence artificielle imaginaient transférer vers la machine une capacité à raisonner comme les humains croient qu'ils le font. Ils voulaient que l'ordinateur reproduise les formes symboliques et logiques du raisonnement dit naturel. Durant les années 1980, ce projet a pris le nom de « système-experts ». On a alors créé des programmes informatiques en leur faisant ingérer des ensembles très sophistiqués de règles de raisonnement, par exemple celles d'un médecin dont on décompose le diagnostic : demander telle chose au patient, si le patient dit A, alors mesurer la constante C, et si la constante C est supérieure à 75, alors vérifier si le patient a des

antécédents familiaux de telle nature, et s'il a des antécédents familiaux alors faire ça, etc. Une fois ces règles transférées dans la machine, celle-ci devait pouvoir poser un diagnostic et prendre des décisions comme l'aurait fait un médecin. L'idée de faire raisonner la machine n'a cependant jamais fonctionné correctement. Les promesses des systèmes-experts n'ont pas été tenues, les entreprises qui les ont développés ont fait faillite, les financements de recherche se sont taris. L'intelligence artificielle est entrée dans son second hiver et, à partir des années 1990, elle était moribonde.

Qu'est-ce qui n'allait pas dans l'idée d'une machine raisonnant logiquement ? Tout simplement, que le fonctionnement de la pensée humaine est impossible à reproduire. Nous prenons très rarement des décisions à partir de règles de raisonnement que nous saurions expliciter. Nos jugements sont aussi faits d'émotions, d'éléments irrationnels, de spécifications liées au contexte et de toute une série de facteurs implicites ; bref, la décision ne se laisse pas capturer par des règles formalisables. Les systèmes-experts raisonnent certes logiquement, mais leur logique est froide, gauche et peu subtile quand elle doit être appliquée à des contextes de la vie sociale. Le monde social dans lequel les machines intelligentes doivent prendre des décisions, se mouvoir et arbitrer entre une multitude de signaux est changeant et pluriel, avec des variations de contextes presque infinies. Pour jouer aux échecs ou au go, la machine raisonne dans un monde clos, borné, simple et n'a pas à être attentive à la variabilité des situations. Mais la société ne ressemble pas à un échiquier sur lequel se déplacent des pièces.

Cette critique, émise par de nombreuses analyses, notamment celle d'Hubert Dreyfus dans *What Computer Can't Do*, s'est portée contre l'idée d'une machine qui raisonne de façon autonome. Mais une autre conception, très différente, de la machine intelligente a aussi pris forme au cours de l'histoire de l'informatique : au lieu d'essayer de la rendre intelligente en lui faisant ingérer des programmes, il serait préférable de la laisser apprendre toute seule à partir des données. La machine apprend directement un modèle des données, d'où le nom d'apprentissage artificiel (machine learning) donné à ces méthodes.

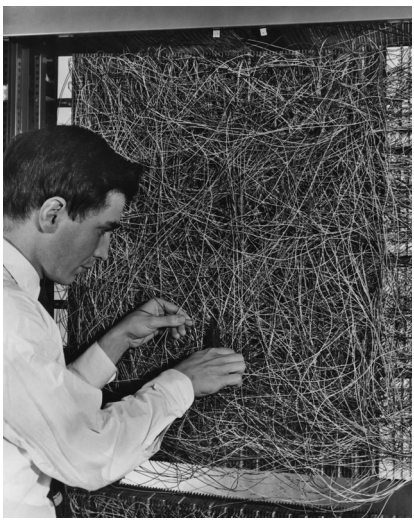
Prenons un exemple. Je veux écrire un programme qui convertit en degrés Celsius une température donnée en degrés Fahrenheit. Pour cela, il existe une règle simple : il faut soustraire 32 de la température en celcius et diviser le résultat par 1,8 (9/5). Une approche symbolique en intelligence artificielle consisterait à enseigner cette règle à la machine. Une approche par apprentissage propose une solution toute différente : au lieu de coder la règle dans la machine, on lui donne seulement des exemples de correspondance entre des températures en degrés Celsius et en degrés Fahrenheit ; on entre les données de cette liste d'exemples, et le calculateur s'en sert pour trouver lui-même la règle de conversion. Voilà, de manière très simplifiée, comment fonctionnent les méthodes d'apprentissage et ce sont principalement à ces méthodes que l'on fait référence quand on parle aujourd'hui d'intelligence artificielle.

Or, elles ont fait des progrès considérables au cours des dernières années, comme le montre le cas de la traduction automatique. Auparavant, on essayait d'inculquer à la machine des règles très sophistiquées qui lui permettent de raisonner comme un grammairien, en connaissant à la fois le vocabulaire, la grammaire, la syntaxe et un ensemble de dictionnaires de signification appelés ontologies. Au cours des années 2000, IBM puis Google ont changé de stratégie, ils ont enlevé du programme toutes les règles symboliques cherchant à rendre la machine intelligente pour les remplacer par des environnements de calcul statistique très puissants lui permettant d'apprendre à partir d'exemples de textes traduits par des humains. Tous les jours la Commission européenne produit des textes en 25 langues. Les traducteurs automatiques « avalent » tous ces textes pour améliorer leurs modèles. La machine ne cherche plus à comprendre la grammaire, elle fait des scores probabilistes sur les meilleurs exemples.

Voilà pour le principe général. Plus précisément, c'est une des différentes techniques d'apprentissage artificiel qui est à l'origine des progrès étonnants que l'on connaît actuellement : la méthode d'apprentissage profond, ou deep learning, qui fonctionne à partir d'une infrastructure dite de réseaux de neurones. L'idée n'est pas nouvelle. Elle a été émise en 1943 par Warren McCulloch et Walter Pitts dans un article devenu célèbre, « A Logical Calculus of Ideas Immanent in Nervous Activity », où les auteurs proposent de reproduire mathématiquement, de façon formelle, le fonctionnement d'un neurone qui s'excite lorsqu'un seuil électrique a été atteint par les

différents flux qui lui viennent de ses synapses. À la fin des années 1950, le psychologue américain Frank Rosenblatt se lance dans la conception d'une machine-cerveau fonctionnant sur le principe des réseaux de neurones, le Mark I, capable de détecter des formes à l'intérieur d'images à partir d'un calcul distribué au sein d'un réseau de neurones (document 66). Pour la faire fonctionner il invente le Perceptron, un algorithme destiné à régler le poids de la contribution des synapses qui inspirera le développement des techniques actuelles d'intelligence artificielle.

Document 66 — Une machine-cerveau, le Mark I Perceptron



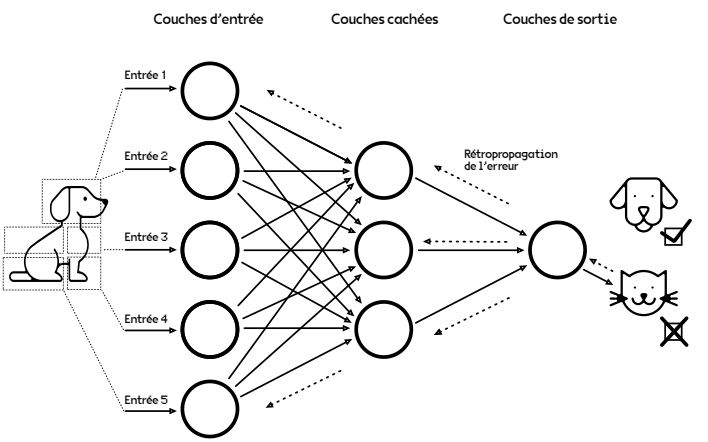
Le Mark I Perceptron, conçu par Frank Rosenblatt en 1957, était une machine d'apprentissage inspirée du réseau neuronal humain. Elle regroupait 400 cellules photoélectriques connectées à des neurones artificiels. Les poids synaptiques étaient encodés dans des potentiomètres, et les changements de poids pendant l'apprentissage étaient effectués par des moteurs électriques. Cette machine physique peut être considérée comme l'ancêtre des techniques actuelles de l'intelligence artificielle. Mais aujourd'hui, le calcul est effectué virtuellement dans la mémoire des ordinateurs dont les capacités se sont considérablement accrues.

La mise en œuvre concrète de ce type de machine apprenante est cependant restée très rare en raison des limitations techniques de l'époque. L'idée que l'intelligence artificielle devait être « connexionniste » et reproduire les mécanismes perceptifs du cerveau humain a été jugée irréaliste et mise au placard par les promoteurs de l'intelligence artificielle symbolique des années 1980. C'est pourtant elle qui revient aujourd'hui sur le devant de la scène.

Comment marche l'apprentissage profond ? En s'inspirant de la biologie : la méthode reproduit informatiquement le fonctionnement des neurones et des synapses du cerveau. Les données sont décomposées en entrées de la façon la plus élémentaire possible. Par exemple, pour les images, le signal en pixels est décomposé à l'aide des chiffres des trois valeurs RVB (rouge, vert, bleu) qui les caractérisent. Chaque donnée en entrée est associée par une synapse à un neurone, une sorte d'automate à deux valeurs. Il s'oriente vers le 0 ou le 1 en fonction des coefficients qui ont été donnés aux synapses auxquels il est connecté. La première couche de neurones est ensuite liée à une autre couche par un nouveau système de synapses qui, elles aussi, vont se voir attribuer des coefficients qui font tourner en 0 ou 1 la nouvelle couche de neurones, et ainsi de suite. On ajoute un nombre plus ou moins important de couches cachées pour arriver à un résultat final, car au bout des couches de neurones, il y a un objectif à satisfaire. On parle alors d'apprentissage supervisé, car c'est un humain qui dit à la machine s'il y a un « chat » ou s'il n'y a pas de « chat » sur l'image. Si l'on donne plusieurs

milliers d'images de chats à un réseau de neurones, avec un système statistique à la fois simple et mystérieux, les coefficients oscillent et se règlent pour apprendre au réseau de neurones la forme « chat ». Une fois cette forme apprise, le système parviendra à reconnaître les images comportant un chat et les images n'en comportant pas. Cet outil statistique simple et mystérieux est l'algorithme de rétropropation de gradient (ou *backprop*) qui a été mis au point par David Rumelhart et Geoffrey Hinton en 1986 (même s'il avait déjà été imaginé précédemment) pour permettre de répartir le poids de tous les coefficients des synapses des réseaux de neurones et fabriquer ainsi des modèles d'un genre très particulier **(document 67).**

Document 67 — Ce chien n'est pas un chat…



La méthode de deep learning (ou réseaux de neurones) consiste à décomposer des données en entrées pour obtenir la granularité la plus fine possible (ici les pixels de l'image) et à fixer un objectif en sortie (ici par exemple : est-ce un chat ou pas ?). Entre l'entrée et la sortie, des couches de milliers de neurones formels s'activent (ou pas), en fonction du poids des synapses qui les relient. L'algorithme de la *backdrop* (rétropagation de l'erreur) règle les coefficients de synapses afin de former une grande matrice de chiffres correspondant au modèle et de permettre à la machine d'identifier des chats dans des images.

De cette description succincte du fonctionnement du deep learning, il faut surtout retenir que le modèle du chat qu'a appris le réseau de neurones n'est pas une description intelligible du chat comme essayait de le faire l'intelligence artificielle symbolique : « Un chat est une forme de 20 à 40 centimètres de long qui a quatre pattes, des oreilles pointues et des moustaches ». Le chat appris par le réseau de neurones est une immense matrice de coefficients qui donnent le poids chiffré de chaque synapse. Si l'on ouvrait le capot d'un réseau de neurones, on n'y verrait pas un chat et on comprendrait encore moins comment il a pu se muer en une matrice de chiffres. C'est une véritable boîte noire. À la suite de Chris Anderson annonçant « la fin de la théorie » dans un article à succès, « The End of Theory », beaucoup d'observateurs ont déploré que ces nouvelles formes de calcul ne permettaient plus de connaître le monde. Les réseaux de neurones sont efficaces, mais on ne sait pas pourquoi. On leur fournit des données en entrée, on leur fixe un objectif en sortie, mais entre les deux, il est difficile de comprendre comment la prédiction s'est formée (les programmateurs eux-mêmes ne comprennent pas vraiment comment s'opère le calcul). Mais il est vrai que les recherches en cours sont en train de trouver les moyens de rendre intelligibles ce que font les réseaux de neurones.

Le deep learning accomplit actuellement une percée sensationnelle. Il permet aux machines d'identifier les images, de lire les adresses des courriers, de trier les spams. On voit se profiler les domaines dans lesquels il jouera un rôle décisif. La traduction automatique a

soudain fait des progrès étonnants, les voitures se conduisent de mieux en mieux toutes seules, les robots s'orientent beaucoup mieux dans leur environnement, la génération de langage et de textes est bien meilleure, etc. Ces machines sont extrêmement performantes pour réaliser des tâches de perception liées au son, à l'image ou au langage. En revanche, elles sont peu adaptées au raisonnement et aux tâches complexes pour lesquelles elles doivent s'hybrider avec les règles symboliques de l'intelligence artificielle traditionnelle.

Les techniques d'apprentissage actuelles sont supervisées – on donne aux machines un objectif (« Ceci est un chat »), pour qu'elle puisse apprendre leur modèle –, mais personne ne sait encore concevoir une intelligence artificielle non supervisée. Les machines sont spécialisées dans le domaine d'apprentissage qu'elles ont appris. Si l'intelligence est la capacité à varier les heuristiques, les cadres d'interprétations et les visions du monde, c'est-à-dire à faire des prédictions de façon non pas modulaire mais méta-modulaire, alors les machines spécialisées n'ont pas cette intelligence.

L'histoire de l'intelligence artificielle nous apprend que la trajectoire des innovations technologiques n'est jamais rectiligne. Elle passe toujours par des détours et des bifurcations. Des stratégies de recherche restent dans l'impasse et d'autres que l'on croyait stériles se débloquent. Ainsi vont les technologies. À tous ceux qui annoncent qu'une intelligence autonome sera parmi nous en 2025, tels les transhumanistes, tenants de la théorie de la singularité de Ray Kurzweil, il faut

rappeler que la science en train de se faire suit un parcours beaucoup plus original et subtil qu'une simple ligne droite.

À LIRE, À VOIR, À ÉCOUTER

● Une histoire originale de l'intelligence artificielle, organisée autour de l'opposition entre ceux qui veulent rendre les machines « intelligentes » et ceux qui veulent rendre les humains plus « intelligents » grâce aux machines, assortie d'un récit complet des progrès de la robotique : John Markoff, *The Machine of Loving Grace. Between Human and Robots*, New York (N. Y.), HarperCollins, 2015.

● Le livre qui a porté la critique la plus forte de l'intelligence artificielle des années 1960 et des années 1980 et qui a été à l'origine d'une vaste réflexion philosophique sur les raisons de l'échec des approches symbolique : Hubert C. Dreyfus, *What Computers Can't Do: The Limits of Artificial Intelligence*, New York (N. Y.), Harper & Row, 1972.

● Sur l'histoire du conflit entre l'approche symbolique (la machine raisonne avec des symboles et des règles) et l'approche connexioniste (la machine produit des calculs de très bas niveau dans un système statistique distribué mimant le fonctionnement des neurones) : Dominique Cardon , Jean-Philippe Cointet et Antoine Mazières, « La revanche des neurones. L'invention des machines inductives et la controverse de l'intelligence artificielle », *Réseaux*, 211, 2018, p. 173-220.

● L'article qui a déclenché le débat sur la « fin de la théorie », dans lequel Chris Anderson défend l'idée que les nouvelles machines n'ont pas besoin des hypothèses humaines sur les données pour produire des résultats, et que les nouveaux calculateurs peuvent fonctionner efficacement sans avoir besoin de « théories » : Chris Anderson, « The End of Theory. The Data Deluge Makes the Scientific Method Obsolete », *Wired Magazine*, 16 juillet 2008, **https://www.wired.com/2008/06/pb-theory/**

● Un long article du *New York Times* qui raconte la manière dont le système de traduction automatique de Google a été complètement transformé et dont il est devenu soudainement performant lorsque les ingénieurs ont décidé d'y intégrer les techniques de deep learning : Gideon Lewis-Kraus, « The Great AI Awakening », *New York Times*,

14 décembre, 2016,
https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html?smprod=nytcore-ipad&smid=nytcore-ipad-share&_r=0

● Deux vidéos pour comprendre le fonctionnement du machine learning : l'une, très simple, présente la manière dont une machine apprend des données, CGP Grey, « How Machines Learn » (8'),
https://www.youtube.com/watch?v=R9OHn5ZF4Uo
l'autre, plus longue mais claire et minutieuse, entre dans le mécanisme du deep learning, notamment la descente de gradient, qui est au cœur de la mécanique statistique des systèmes d'apprentissage à base de réseaux de neurones : Brandon Rohrer, « How Deep Neural Networks Work » (25'),
https://www.youtube.com/watch?v=ILsA4nyG7I0

● Une redoutable critique des théories de la singularité et une vision réaliste et informée des avancées de l'intelligence artificielle, par un des meilleurs experts du domaine : Jean-Gabriel Ganascia, *Le Mythe de la singularité. Faut-il craindre l'intelligence artificielle ?*, Paris, Seuil, 2017.

● Un ouvrage expert mais lisible qui détaille les différentes familles de machine learning (symboliste, analogique, connexionniste, bayesien, etc.), présente les enjeux techniques – principalement statistiques – des progrès en la matière et constitue une bonne introduction aux mécanismes de l'intelligence artificielle : Pedro Domingos, *The Master Algorithm. How the Question for the Ultimate Machine Will Remake Our World*, Londres, Penguin Random House UK, 2015.

● Sur le débat sur l'automatisation, une analyse éclairée par de nombreux exemples : Nicholas Carr, *Remplacer l'humain. Critique de l'automatisation de la société*, Paris, L'échappée, 2017 [*The Glass Cage. Automation and Us*, New York (N. Y.), W. W. Norton & Company, 2015].