Une brève introduction à l'intelligence artificielle

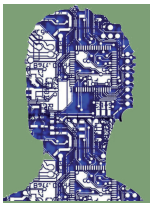
Aurélie Jean

médecine/sciences 2020 ; 36 : 1059-67

Une brève introduction à l'intelligence artificielle

Aurélie Jean

› Depuis plus d'une décennie, l'intelligence artificielle (IA) vit une accélération dans son développement et son adoption. En médecine, elle intervient dans la recherche fondamentale et clinique, la pratique hospitalière, les examens médicaux, les soins ou encore la logistique. Ce qui contribue à l'affinement des diagnostics et des pronostics, à une médecine encore plus personnalisée et ciblée, à des avancées dans les technologies d'observations et d'analyses ou encore dans les outils d'interventions chirurgicales et autres robots d'assistance. De nombreux enjeux propres à l'IA et à la médecine, tels que la dématérialisation des données, le respect de la vie privée, l'*explicabilité*[1] des algorithmes, la conception de systèmes d'IA inclusifs ou leur reproductibilité, sont à surmonter pour construire une confiance du corps hospitalier dans ces outils. Cela passe par une maîtrise des concepts fondamentaux que nous présentons ici ‹



In Silico Veritas, 4 rue Joseph Granier, 75007 Paris, France.
aurelie@silicoveritas.com

L'intelligence artificielle (IA) est une notion paradoxale car, comme le souligne Yoshua Bengio[2], on ne rend pas l'ordinateur plus intelligent mais on le rend au contraire moins stupide. L'IA est à la fois une discipline de recherche et une matière, à l'instar des mathématiques ou de la physique, utilisée dans de nombreuses autres disciplines de recherche, comme la médecine. Aujourd'hui plus que jamais, comprendre ce qu'est l'IA, ce qu'elle fait, ce qu'elle fera sûrement et ce qu'elle ne fera certainement jamais[3], est un moyen ingénieux d'en comprendre et d'anticiper, même partiellement, le champ des possibles, et donc de s'y préparer, en tirant tous les avantages tout en écartant les possibles menaces. Depuis plus de dix ans, cette discipline, ainsi que les technologies qui y recourent, vivent une certaine accélération dans leurs développements et les usages qu'elles en font, due, entre autres, aux avancées dans l'efficacité des algorithmes appliqués au traitement des données, dans la technologie des capteurs et des systèmes IoT (*Internet of Things*) pour la collecte, ou encore dans les procédés des systèmes de stockage de ces données. Cet article se veut être une brève introduction à l'intelligence artificielle, dans sa définition, sa classification, son application, son potentiel et ses limitations. Il présente des références et des exemples concrets non exhaustifs, qui illustrent les concepts, principalement dans le domaine de la médecine.

Analyser, comprendre et prédire

L'intelligence artificielle regroupe les méthodes de calculs numériques, sur ordinateur, qui reproduisent un certain type d'intelligence : l'*intelligence analytique (Glossaire)*. Le concept d'*intelligence computationnelle* a été concrétisé dans une publication de Alan Turing en 1950 [1], alors que le terme *intelligence artificielle* ainsi que le paradigme intellectuel qui lui est associé ont été proposés à l'issue de l'école d'été de l'université de Dartmouth de 1956 [32] (➜). Grâce à des modèles d'IA, des simulations sont réalisées sur ordinateur, selon trois objectifs : analyser, comprendre et prédire un phénomène. Ce dernier peut provenir de

(➜) Voir le Repères de J. Haiech, *m/s* n° 10, octobre 2020, page 919

Vignette (© Lightwise/123 RF).

[1] L'« explicabilité » est un principe de la régulation des algorithmes. Elle figure dans les recommandations que l'Organisation de coopération et de développement économiques (OCDE) a adoptées.

[2] Interview accordée au *MIT Technology Review* en 2016 : *Will Machines Eliminate Us?*

[3] Citation de 2019 de Antoine Bordes, directeur de recherche chez Facebook.

différents domaines : médical [33] (➜) économique, financier, éducatif, ou industriel.

(➜) Voir le Repères de C. Matuchansky, *m/s* n° 10, octobre 2019, page 797

Dépasser les contraintes du monde réel

On est amené à réaliser des simulations numériques pour plusieurs raisons. Les expériences dans la vie réelle sont souvent limitées, que ce soit par le temps d'exécution, par le coût financier, ou tout simplement l'impossibilité de les réaliser. Observer et analyser des mécanismes de déformation d'un tissu *in vivo*, à l'échelle nanoscopique, reste ainsi encore difficile pour ne pas dire impossible dans de nombreux cas, en raison principalement des limites dues aux techniques de microscopie. Conduite manuellement, une enquête médicale sur des dizaines de milliers de personnes peut prendre un temps significatif pour la collecte et l'analyse des données, rendant l'exercice difficilement réalisable dans de nombreuses situations. Réaliser des essais expérimentaux de tenue mécanique de toutes les pièces d'un équipement médical, telle qu'une machine d'imagerie par résonance magnétique (IRM) reste un exercice qui prend du temps et qui nécessite un budget souvent considérable. Les simulations sur ordinateur permettent de lever ces contraintes, soit en se substituant à des observations et des analyses difficiles ou impossibles à réaliser dans le monde réel, soit en complétant des expériences qui ont été réalisées concrètement. La simulation numérique permet donc de fournir des réponses rapides avec une certaine fiabilité à des questions ou des problèmes souvent difficilement solubles dans le monde réel. Les simulations numériques permettent en outre de révéler des mécanismes jusqu'alors insoupçonnables, ou encore largement incompris, dans de nombreux champs disciplinaires.

« Analyser, comprendre et prédire » en médecine

Les récentes contributions utilisant l'IA s'articulent en grande majorité autour de son utilisation pour évaluer et prédire des situations, des grandeurs ou encore leurs évolutions. On cherche ainsi à évaluer la propagation d'une épidémie, ou à identifier une tumeur sur une radiographie, ou même à prédire des tremblements de terre et à anticiper les variations d'une action en bourse. Cela étant dit, ces modèles peuvent également contribuer à comprendre les mécanismes d'un phénomène, comme saisir les mécanismes d'une maladie, les raisons de catastrophes naturelles ou les origines d'une crise boursière. En médecine, on utilise des modèles pour analyser plus rapidement et plus efficacement des échantillons biologiques, pour détecter des cellules cancéreuses ou pour identifier des artéfacts sur des électrocardiogrammes. Un effort est réalisé autant dans l'identification de schémas répétitifs, dans les résultats et les données d'entrée, pour la mise en évidence d'une relation de causalité, que dans la simulation *stricto sensu* des mécanismes des maladies. On peut, par exemple, utiliser une approche contrefactuelle pour simuler une réponse à partir de différents scénarios. On identifiera ainsi le scénario « solution du problème », en comparant la réponse simulée à celle qui est observée dans le monde réel. On peut aussi modéliser et simuler directement les processus mécaniques, chimiques et/ou physiologiques d'un organe [2], d'un tissu [3] ou d'une cellule [4], pour comprendre une maladie, et donc la traiter.

Pour évaluer le potentiel de l'IA en médecine, autant dans l'analyse et la compréhension que dans la prédiction, il est fondamental d'en comprendre le fonctionnement de base ainsi que les défis à surmonter pour en extraire tous les bénéfices, tout en écartant les risques possibles. Pour cela, il est intéressant de différencier les IA afin d'en comprendre les tenants et les aboutissants, et ce, dans tous les domaines d'application.

GLOSSAIRE

Intelligence émotionnelle : l'intelligence émotionnelle d'une personne se réfère à l'identification et à la maîtrise de ses propres émotions et de celles des personnes avec lesquelles elle interagit. Ce concept, introduit par les psychologues Peter Salovey et John Mayer, a été largement déployé par le psychologue américain Daniel Goleman dans les années 1990.

Intelligence analytique : l'intelligence analytique d'une personne se réfère à sa capacité à traiter l'information qu'elle reçoit et à résoudre des problèmes. Elle est souvent présentée comme complémentaire aux deux autres intelligences du modèle triarchique de Robert Sternberg, l'intelligence créative et l'intelligence pratique.

Intelligence de situation : l'intelligence de situation d'une personne se réfère à la combinaison des deux des trois intelligences du modèle triarchique de Robert Sternberg : l'intelligence créative et l'intelligence pratique. Elle décrit la capacité d'une personne à s'adapter à son environnement, résoudre un problème dans un contexte nouveau, avec un certain pragmatisme mais aussi avec une originalité dans son approche et sa réflexion.

Intelligence artificielle : l'intelligence artificielle se réfère à une discipline qui consiste à reproduire par la simulation numérique sur un ordinateur les intelligences humaines.

Intelligence générale : l'intelligence générale se réfère à un système d'intelligence artificielle capable de reproduire l'ensemble des capacités cognitives d'un être humain, et qui constitue l'ensemble des intelligences : analytique, émotionnelle et de situation.

Conscience : en psychologie, la conscience d'un individu traduit le phénomène d'identification par lui-même de sa place dans le monde qui l'entoure, et de ses interactions avec celui-ci.

Point de singularité technologique : le point de singularité technologique se réfère à une théorie hypothétique qui stipule l'existence même lointaine d'un point de basculement dans le futur à partir duquel les machines maîtriseraient l'intelligence générale et aurait la conscience d'exister.

Intelligence artificielle explicite *versus* implicite

Il existe de nombreuses manières de classer les IA. On peut ainsi choisir de distinguer l'IA dite *faible*, actuellement développée et utilisée, de l'IA dite *forte*, qui suppose une maîtrise, par la machine, de l'intelligence dite générale *(Glossaire)*. Cette intelligence générale regroupe l'intelligence analytique, déjà présente dans l'IA faible, et les intelligences de situation et émotionnelle. Ce passage de l'IA faible à l'IA forte est régi par la théorie du *point de singularité technologique* [5]. Celui-ci suppose l'existence, dans un futur même lointain, d'un point de basculement technologique, où les machines auront une intelligence générale équivalente ou supérieure à celle de l'humain... avec la conscience d'exister. Cette théorie reste largement hypothétique et n'est pas soutenue consensuellement par la communauté scientifique. Néanmoins, chercher à atteindre un tel point de basculement est extrêmement stimulant intellectuellement. Cela permet de faire des progrès notables, autant dans le développement des technologies et des méthodes de calcul avancées, que dans la compréhension de notre monde [5]. Citons, par exemple, les travaux qui s'intéressent à comprendre, modéliser et simuler une émotion ou la conscience, et qui ont des implications technologiques et philosophiques pertinentes.

Une autre manière de classer les IA, qui est celle que nous retiendrons dans cet article, est la distinction entre une IA dite *explicite* et une IA dite *implicite*. Alors que la distinction entre IA faible et IA forte n'apporte aucun élément d'analyse supplémentaire, en raison de la non existence actuelle et future d'une IA forte (avec maîtrise, par la machine, de l'intelligence dite générale), différencier l'IA explicite de l'IA implicite permet d'approcher concrètement et efficacement de nombreuses questions, en traitant ces deux IA qui sont aujourd'hui développées et utilisées.

Intelligence artificielle *explicite*

L'intelligence artificielle *explicite*, également appelée IA symbolique, est celle traditionnellement développée et utilisée depuis plus de cinquante ans. La logique de cette IA est entièrement décrite explicitement par le(s) humain(s), et se traduit de différentes manières selon la complexité du processus à simuler *(Figure 1)*. On parle également de systèmes experts. Il existe des algorithmes qui comprennent des structures conditionnelles uniquement (qui définissent la décision à prendre), sans nécessité en amont d'une modélisation mathématique *stricto sensu*. On peut ainsi citer les arbres décisionnels explicites, qui sont des ensembles de nœuds et de branches définissant respectivement les conditions à satisfaire et les chemins à prendre selon la réponse à la condition antérieure. Un arbre décisionnel explicite peut, par exemple, être utilisé pour l'estimation des diagnostics, selon la nature et la fréquence de symptômes [6], ou dans l'évaluation de pronostics en réponse à un traitement [7]. Parmi les systèmes experts connus en recherche médicale, on peut citer Mycin, un système expert développé dans les années 1970 par l'université de Stanford en Californie (États-Unis), construit à partir de plus de 600 règles explicites et qui permet d'identifier la bactérie à l'origine d'une infection et d'optimiser le traitement antibiotique adéquat. Il existe également des IA explicites qui contiennent des modèles mathématiques, parfois complexes, pour décrire le problème à résoudre ou la question posée [8]. Cette approche a été utilisée, par exemple, dans la compréhension des mécanismes à l'origine d'un traumatisme crânien [9], dans la simulation atomistique[4] de protéines constitutives des tissus humains [10], ou dans la simulation du comportement mécanique du tissu constituant le col de l'utérus [11].

En plus du choix des équations mathématiques régissant le phénomène à étudier, on est souvent conduit à établir des hypothèses, autant sur le modèle lui-même que sur ses conditions d'utilisation. Ces hypothèses sont souvent le résultats d'une méconnaissance de certains aspects du phénomène à simuler, à l'impossibilité de décrire tous les cas d'usage (ou scénario), ou à la volonté d'élargir le contexte d'application d'un modèle pré-établi. Dans le cas des calculs permettant d'étudier les mécanismes impliqués dans les traumatismes crâniens, par exemple [9], en accord avec des essais réalisés *in situ*, seules les sollicitations volumiques (c'est-à-dire les pressions exercées) dans le cerveau ont été considérées. Dans ce cas, les efforts de cisaillements induits sur le cerveau ont été écartés de l'analyse. Mais ces hypothèses de modèle et de son utilisation sont à surveiller, car elles peuvent à tout moment être identifiées comme fausses ou trop approximatives. Elles reposent sur des raisonnements qui sont fondés sur la physique, la biologie, la sociologie ou les mathématiques, qui présentent, dans certains cas, des biais, des failles ou des contradictions.

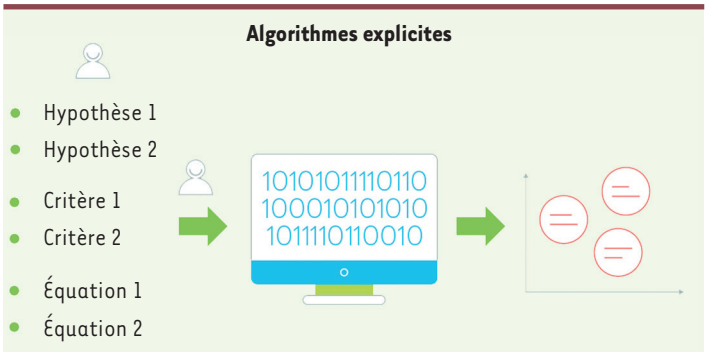


Figure 1. ***Schéma illustrant le fonctionnement général d'algorithmes explicites, dont l'ensemble des hypothèses, des critères et des équations sont définis explicitement par l'homme*** (© Cartoonbase).

[4] Simulation numérique d'arrangements moléculaires pour identifier une dynamique et des fonctions optimales, ou encore comprendre le comportement électronique et mécanique d'un matériau par exemple.

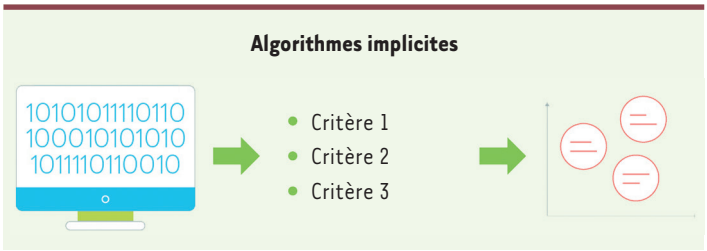


Figure 2. ***Schéma illustrant le fonctionnement général d'algorithmes implicites, dont la logique est traduite par un ensemble de critères implicitement définis à la suite d'un apprentissage*** (© Cartoonbase).

L'IA explicite présente néanmoins un avantage significatif par rapport aux autres approches, autant d'un point de vue scientifique, que de par son utilisation. On est en effet dans la capacité d'expliquer clairement et précisément l'ensemble des règles logiques de ce système d'IA, dans la mesure où on le décrit explicitement. On peut donc également décrire explicitement comment les résultats issus de simulations numériques sont obtenus. Ceci permet d'élargir son utilisation à des applications et des domaines pour lesquels la législation impose une certaine transparence et un haut niveau d'*explicabilité*, comme les milieux médical ou financier. L'IA explicite présente cependant une forte limitation : elle suppose que l'on est dans la capacité de *mathématiser*, ou, *a minima*, d'objectiver, le problème à résoudre. Or, il existe de nombreux cas pour lesquels aucune expression mathématique ne peut être définie pour expliquer un phénomène. Par exemple, il est quasiment impossible de décrire mathématiquement et explicitement, avec une grande fiabilité systématique, la présence d'une tumeur sur une radiographie. On peut en supposer la taille ainsi que la forme, mais on risque d'écarter un grand nombre de morphologies tumorales. Dans ce cas précis, on peut utiliser des IA dites *implicites*, qui ont la particularité et l'avantage de ne pas nécessiter de mathématiser *a priori* le problème à résoudre mais, au contraire, de déduire implicitement une logique de résolution à partir des données observées. Alors que dans les IA explicites, on utilise une approche dite *model driven* (orientée par le modèle), dans la mesure où les modèles mathématique et numérique orientent le raisonnement de la simulation, dans les IA implicites il est orienté par la donnée, on parle dans ce cas d'approche *data driven* (orientée par la donnée).

Intelligence artificielle *implicite*

L'intelligence artificielle *implicite* regroupe l'ensemble des systèmes d'IA dont la logique est traduite implicitement, par apprentissage à partir de données dites d'apprentissage *(Figure 2)*. Contrairement aux systèmes d'IA explicite, une IA implicite est capable de capturer des mécanismes avec un haut niveau d'abstraction dans l'exploitation des données d'apprentissage, ce qui permet de construire une logique par ailleurs difficilement traduisible explicitement. C'est ainsi que des IA implicites permettent de reconnaître une tumeur sur une image médicale [12], les signaux faibles d'un infarctus ou d'une crise d'épilepsie [13]. L'IA implicite est également utilisée pour l'analyse sémantique d'un texte, l'identification d'une entité précise sur une image, ou la traduction textuelle d'une conversation orale. L'apprentissage peut se réaliser par une simple analyse statistique sur des données : c'est le cas des IA de catégorisation, qui classent implicitement des ensembles de données en fonction de leurs similarités. Il peut également se réaliser par construction de réseaux neuronaux, comme dans le cas de nombreuses IA implicites décisionnelles.

Même si les IA implicites ont de nombreux avantages, principalement en raison de leur haut niveau d'abstraction qui leur permet de résoudre des problèmes complexes, elles présentent quelques limitations. Contrairement aux IA explicites, elles présentent un faible niveau d'*explicabilité* et d'*interprétabilité*, faisant ainsi référence à la fameuse « boîte noire » dans le fonctionnement des outils. De manière générale, un très grand nombre de données sont nécessaires pour assurer un apprentissage réaliste et juste. Également, les IA implicites ne permettent pas de comprendre aussi facilement un phénomène, et d'en extraire des mécanismes sous-jacents, que les IA explicites. Lors de la conférence internationale sur le *machine learning*, NeurIPS (*neural information processing systems*), de 2019, les scientifiques se sont accordés sur l'importance de faire cohabiter les deux types d'IA, implicite et explicite, dans les modèles numériques, afin de résoudre les grands problèmes de l'humanité, tels que ceux concernant les changements climatiques qui nécessitent une compréhension fine des phénomènes. On parle alors de modèles hybrides [14].

Implicite ou explicite, ces IA fonctionnent par l'exécution d'algorithmes qui capturent la logique de résolution du problème auquel on s'intéresse. On parle d'algorithmes computationnels, en comparaison des algorithmes historiques ou traditionnels, qui, eux, sont destinés à être manipulés à la main, bien loin de toute intelligence artificielle.

Les algorithmes

Algorithmes traditionnels

Un algorithme est, littéralement, une séquence d'opérations ou de tâches à exécuter selon une certaine logique pour répondre à une question ou résoudre un problème. Même si les algorithmes aujourd'hui couramment utilisés sont exécutés par un ordinateur, leur histoire est en réalité bien plus ancienne que la naissance du premier micro-processeur. Le mot algorithme vient du nom latinisé, *Algorithmi*, du mathématicien perse Muhammad ibn Musa al-Khwarizmi, père de l'algèbre au IX^{e} siècle de notre ère. Il faut cependant remonter un millénaire pour observer les débuts du raisonnement logique « algorithmisé »... dans les travaux d'Euclide, vers 300 ans avant notre ère. Dans son œuvre *Les Éléments*, Euclide propose en effet un ensemble d'objets

géométriques permettant, entre autres, l'élaboration et la démonstration de théorèmes par un raisonnement structuré. Alors que les algorithmes étaient traditionnellement conçus pour être utilisés à la main, ils sont aujourd'hui composés pour être programmés dans un code informatique afin d'être exécutés par un ordinateur. On les appelle *algorithmes computationnels*.

Algorithmes computationnels

Les algorithmes computationnels, que l'on appelle aujourd'hui plus simplement algorithmes, constituent l'ensemble des processus (en anglais, *process*) embarqués dans les outils numériques et servant à exécuter une série d'opérations dans un objectif précis. À l'instar de la distinction faite précédemment entre IA explicite et IA implicite, on différencie les algorithmes explicites des algorithmes implicites. Alors que les règles logiques des algorithmes explicites sont précisément définies par les concepteurs, la logique des algorithmes implicites procède d'une architecture (qui décrit les connexions logiques entre les paramètres établies pour le modèle) définie par un apprentissage reposant sur des scénarios réels. L'entraînement s'emploie ainsi à trouver les relations qui existent entre les variables d'entrée (mesurées) et le résultat final souhaité. Comme défini précédemment, les algorithmes explicites peuvent être décrits par des systèmes mathématiques, parfois complexes, et/ou ne contenir que des structures conditionnelles d'exécution d'opérations – *si (condition, résultat : oui, non) alors* – menant à un résultat final.

Parmi les algorithmes implicites, on recense, entre autres, les algorithmes d'apprentissage statistique (comme les algorithmes de catégorisation) et les algorithmes d'apprentissage par réseau neuronal. Les premiers, algorithmes de catégorisation, organisent implicitement, par un calcul d'optimisation, un ensemble de données, en fonction des valeurs qu'elles prennent, selon des classes statistiquement représentatives. Chaque classe définie représentera alors un comportement statistique précis : les individus d'une même classe possèderont donc des mesures statistiques (moyenne, covariance, etc.) établies sur des grandeurs (physiques, médicales, ou sociologiques, par exemple) qui seront similaires. Cette classification permet de comprendre davantage les données caractéristiques d'un phénomène, en extrayant les données statistiques qui ne seront pas intuitivement analysables, mais qui donneront néanmoins une information qui peut être pertinente. Cette technique a été utilisée pour classer les patients souffrant de la maladie d'Alzheimer, selon leurs similitudes phénotypiques, afin, entre autres, de personnaliser leur traitement [15]. Même si des caractéristiques définissant la maladie ont été reconnues consensuellement par la communauté médicale, une classification statistique permet de mettre en évidence implicitement des corrélations entre symptômes, ou encore entre profils phénotypiques des patients. Les algorithmes de catégorisation implicite, comme la méthode des *k-means*[5], sont également utilisés pour pré-segmenter des images médicales, telles que celles provenant d'IRM cérébrales, en séparant les différents milieux (cavités, muscles, cerveau, etc.) afin de faciliter la segmentation qui sera finalement réalisée [16]. L'algorithme identifie sur les images d'IRM, après optimisation, chaque ensemble tissulaire (les yeux, le cerveau, l'os crânien, etc.). Chacun étant défini par une ou plusieurs nuances de gris spécifiques incluant ses contours et son centre. Les opérations classiques de l'analyse d'image, comme le *watershed*[6], sont ensuite appliquées sur cette partition spatiale de l'image, ce qui facilite et améliore la segmentation finalement obtenue.

Il existe également des algorithmes décisionnels d'apprentissage qui fonctionnent sur un réseau neuronal construit sur des données représentant différents scénarios pour lesquels les décisions résultantes sont connues.

Le réseau neuronal se construit à partir d'un calcul d'optimisation des poids statistiques des nœuds du réseau qui dessine la logique de résolution du problème ou de la question auxquelles l'algorithme apportera une solution. Ces algorithmes « apprennent », à partir de données d'apprentissage dites labellisées (ou étiquetées) (voir plus loin) : on parle alors d'apprentissage supervisé (ou contrôlé). Des algorithmes d'apprentissage profond, conçus il y a trente ans [17], aussi appelés algorithmes de *deep learning*[7] [18, 19], et qui présentent plusieurs couches de neurones, sont aujourd'hui largement utilisés. Ils permettent d'augmenter le niveau d'abstraction des données et donc la complexité de la logique implicite.

Parmi ces algorithmes computationnels implicites, on trouve des algorithmes du langage, de l'analyse sémantique, ou de la vision par ordinateur (communément défini par le terme anglais *computer vision*). En médecine, ces algorithmes implicites d'apprentissage sont utilisés dans la conception de nouveaux médicaments [20], pour différencier les types de cancers dermatologiques [21], identifier et confirmer un diagnostic [22], ou pour la détection de tumeurs sur des radiographies.

Dans de nombreux cas d'apprentissage, l'accès aux données est limité ou leur *labellisation*[8] est difficile à réaliser en raison du temps nécessaire pour cette étape et de sa complexité. Pour surmonter cet obstacle, un apprentissage par transfert peut être réalisé. Il s'articule selon deux étapes. Dans un premier temps, on pré-entraîne un modèle sur un large jeu de données de référence étiquetées pour un problème plus général que celui initialement défini. Puis, dans un second temps, on

[5] La méthode des *k-means* appartient aux algorithmes de classification non supervisée (les groupes n'existent pas avant d'être créés). Les objets (sujets, sites, points, etc.) appartiennent à un seul groupe, chaque groupe étant différent et ne se chevauche pas.

[6] Se réfère à la ligne de partage des eaux : processus de délimitation des contours.

[7] Apprentissage profond.

[8] Annotation, étiquetage.

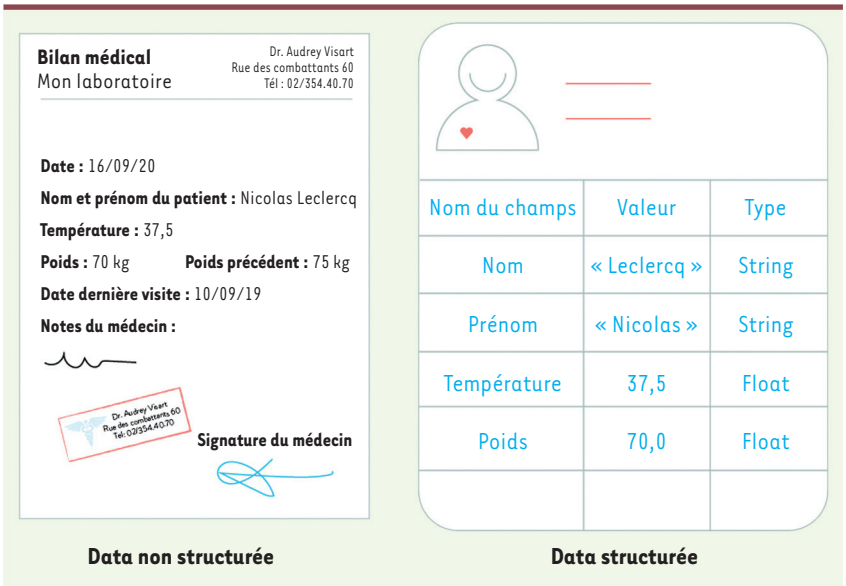


Figure 3. ***Schéma illustrant les différences entre une data non structurée brute et une data structurée formatée*** (© Cartoonbase).

réalise un apprentissage, dit fin, sur le modèle qui a été pré-entraîné, en utilisant le jeu de données disponible et limité en taille, spécifiquement choisi pour résoudre le problème posé. Une telle approche a été utilisée dans l'interprétation d'images de radiographies thoraciques [23], dans le diagnostic de rétinopathie diabétique, sur des examens de fond d'œil [24], ou dans le diagnostic de la maladie d'Alzheimer, à partir d'images de tomographie du cerveau [25]. Dans le cas de grands échantillons de données, mais qui restent difficile à étiqueter, des méthodes d'apprentissage, dites semi-supervisées, sont également utilisées. Dans ce cas, l'entraînement est réalisé sur des données faiblement labellisées, ou sans labels, selon des contraintes explicitement définies par le résultat obtenu. Une telle méthode a fait ses preuves dans la segmentation d'images provenant d'IRM cérébrales [26], ou dans la détection de lésions vasculaires, sur des images de tomodensitométrie [27].

Les limites et le potentiel des algorithmes explicites ou implicites résident, entre autres, dans le type, la taille et la diversité des données utilisées. Comprendre ce que sont ces données (ou *data*), comment les distinguer, et quels traitements leur appliquer, est fondamental pour envisager les futures avancées et le choix des techniques d'IA selon les applications.

Les *data*

Le terme *data* regroupe l'ensemble des données, que l'on appelle plus largement informations, sous la forme, par exemple, de chiffres, de graphiques, de photos, de bandes-son ou de textes. La nature (sa forme) et la source (son origine et le moyen de sa collecte) caractérisent cette *data* qui est très souvent traitée après avoir été extraite sous forme de données dites brutes. Cette donnée brute est donc ensuite *nettoyée* afin d'éliminer le bruit (la perturbation) et *formatée* pour en extraire les informations pertinentes pour l'analyse. Très souvent, la *data* brute collectée n'est pas structurée lors de sa collecte et elle nécessite un certain formatage afin de pouvoir être exploitée.

Data structurée *versus* non structurée

La *data* non structurée représente l'ensemble des données brutes obtenues après qu'elles aient été collectées *(Figure 3)*. Un questionnaire rempli par un patient avant une visite médicale constitue, par exemple, une donnée brute non structurée. On pourrait citer également un électrocardiogramme, une image d'échographie ou un bilan sanguin. Sous cette forme, la *data* est inexploitable *stricto sensu* par un algorithme de calcul. Il est donc nécessaire de structurer les informations recueillies que cette *data* contient pour obtenir une organisation, possiblement hiérarchisée, des valeurs qui lui sont associées. Dans le cas du questionnaire du patient, des champs décrivant la *data* « patient », qu'on appelle aussi *métadata*, comme l'âge, le genre, les antécédents chirurgicaux, les traitements médicamenteux, la situation familiale, ou la date du dernier bilan médical, sont extraits. Cette structuration peut être réalisée automatiquement, comme dans le cas d'un bilan sanguin à partir d'un document au format PDF dont on extrait (par un algorithme) les valeurs associées à chaque champ qui sont précisément localisés dans le document. La donnée structurée est donc une donnée à laquelle est associé un ensemble de métadonnées *(Figure 3)*.

Data de calibration *versus data* d'apprentissage

On distingue également les *data* de calibration et les *data* d'apprentissage *(Figure 4)*. Dans le cas d'un algorithme explicite, de nombreux paramètres, rattachés à la modélisation mathématique du phénomène à simuler, sont à identifier. Pour cela, on calcule la valeur des constantes du modèle, contenues dans l'algorithme, en confrontant les résultats issus de la simulation à la valeur des données obtenues à partir de mesures réalisées *in situ* ou d'expérimentations concrètement réalisées. On parle alors de calibration algorithmique et de données de calibration identifiées par un calcul d'optimisation. La taille et la nature des données utilisées pour cette étape doivent être choisies de telle sorte que l'échantillon de calibration

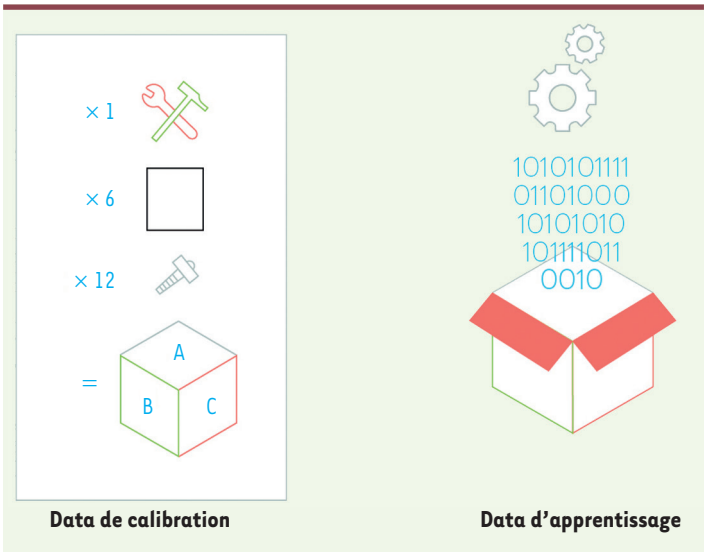


Figure 4. ***Schéma illustrant la différence entre la data de calibration utilisée pour paramétrer un algorithme explicite, et la data d'apprentissage utilisée pour entraîner un algorithme implicite*** (© Cartoonbase).

soit représentatif. Dans le cas contraire, les paramètres identifiés et l'algorithme qui y est associé, peuvent présenter des biais. L'algorithme risque ainsi de ne fonctionner que pour une partie des scénarios possibles.

Dans le cas des algorithmes implicites, les règles logiques sont définies implicitement par apprentissage sur des données dites d'apprentissage. La représentativité des données d'apprentissage est encore plus sensible car elle touche le fonctionnement logique de l'algorithme qui possède un plus haut niveau d'abstraction, ce qui induit un risque de biais plus pervasif[9]. Les données d'apprentissage sont étiquetées pour permettre à l'algorithme de dessiner une logique implicite à partir de la construction du réseau neuronal. Cette labellisation de pré-apprentissage consiste, en particulier, à définir les valeurs des métadata, ainsi que d'autre valeurs, qui définissent le résultat recherché par l'algorithme. Dans le cas d'un apprentissage à partir de résultats biologiques afin d'établir un diagnostic, par exemple, un échantillon représentatif de résultats déjà diagnostiqués et pour lequel une liste de métadata est associée, incluant les métadata du patient (telles que l'âge, le genre, ou les antécédents médicaux), est considéré. À chaque *data* de cet échantillon est également rattaché le diagnostic, afin de contrôler le bon apprentissage de l'algorithme, qui apprendra donc des différents scénarios qui auront été mis à sa disposition. Aujourd'hui, de nombreux travaux s'intéressent au développement de nouvelles techniques d'apprentissage, dites non supervisées, qui permettraient de faire apprendre un algorithme à partir de données non labellisées [28]. Des méthodes semi-supervisées, combinant *data* labellisées et non labellisées au sein des données d'apprentissage, sont également utilisées.

[9] Pervasif : enfoui, caché.

Enjeux de l'intelligence artificielle en médecine

Dématérialisation des données de santé

En intelligence artificielle, et en particulier pour les algorithmes implicites, de grandes quantités de *data* sont nécessaires pour la phase d'entraînement. En effet, taille et diversité des données d'apprentissage permettent d'extraire l'ensemble des corrélations et des schémas répétitifs et représentatifs présent dans les scénarios du problème à résoudre. En médecine, comme dans d'autres disciplines, la première étape consiste à numériser l'ensemble des informations pouvant être utilisées dans les modèles d'IA. Que ce soient des antécédents médicaux, des prescriptions de médicaments, ou des résultats d'examens, ces informations doivent être dématérialisées, puis éventuellement structurées pour pouvoir être exploitées par un algorithme. En 2019, la France a déployé l'ordonnance électronique avec plusieurs objectifs, dont le suivi numérisé des informations concernant les noms des médicaments, la prise de génériques, l'observance du traitement, la gestion de maladies chroniques, et une meilleure communication entre le médecin et le pharmacien. Pour une complète dématérialisation des données de santé, à la numérisation des informations médicales, doit s'ajouter la mise en commun et le partage éthique et responsable de ces données. Même si depuis 2018, le dossier numérique médical partagé est généralisé en France, un effort est encore à opérer pour le partage, encore une fois responsable et probablement anonymisé, des données de santé au sein de l'Union européenne, pour affiner et rendre plus efficaces les modèles d'apprentissage.

Respect de la vie privée et des données personnelles

Le corps médical est, par sa formation et le principe du secret médical, respectueux de la vie privée des patients et de leurs données de santé. Le texte européen de protection des données personnelles (RGPD) impose, depuis 2018, des règles strictes pour la collecte, le traitement, le stockage et l'effacement de ces données. Dans ce texte, les données médicales entrent dans le cadre des données dites « sensibles ». Des enjeux concernent la protection contre leur collecte et leur utilisation malveillantes avec, entre autres, l'utilisation et l'amélioration des techniques de cryptage et d'anonymisation. La confidentialité différentielle consiste ainsi à bruiter les données et les modèles, afin de brouiller le signal pour anonymiser les données du patient, sans en altérer la qualité et la précision, et donc l'apprentissage algorithmique qui en résulte.

Explicabilité, *interprétabilité* et transparence des algorithmes

L'émergence des modèles d'IA implicites par apprentissage diminue, par définition, leur niveau d'*explicabilité*, faisant référence à la fameuse « boite noire ». En médecine, il est fondamental de pouvoir expliquer *a minima* la logique de l'algorithme décisionnel utilisé. Ceci permet de garantir une certaine transparence afin d'assurer une compréhension de la part du médecin et du patient, de pouvoir contredire, en connaissance de cause, le résultat donné par l'algorithme, ou d'analyser les causes d'une erreur médicale et d'en définir, le cas échéant, les responsabilités. Plusieurs facteurs contextuels aident à déterminer le degré d'*explicabilité* approprié [29] : le(s) destinataire(s) de l'explication (médecin, patient, autorités de santé, etc.) ; l'impact du modèle d'IA et les possibles dangers de dysfonctionnement du modèle ; ou la conformité à la réglementation. On distingue les méthodes d'*explicabilité* globales et les méthodes dites locales. Dans les méthodes globales, on cherche à définir le fonctionnement global de l'algorithme implicite et la manière dont il apprend à partir des données d'entrée. On examine, en particulier, les métadata qui sont les plus influentes sur la suggestion algorithmique. Dans les méthodes locales, on analyse une donnée en particulier, ou un sous-ensemble de données d'apprentissage, comme un dossier médical ou une image de radiologie. On utilise également une approche de raisonnement contrefactuel pour tester la logique de l'algorithme. En pratique, on étudie l'impact du changement de la valeur d'une métadata sur la décision finale suggérée par l'algorithme. On peut également utiliser une méthode inverse, afin d'analyser les modification à opérer sur les métadata pour obtenir un résultat algorithmique particulier.

Inclusion, égalité et équité au sein des algorithmes

Les modèles d'IA implicite et explicite présentent des risques variables de discrimination technologique[10], en raison de la présence de biais algorithmique(s). Ces biais proviennent des critères explicites de l'algorithme, des données de calibration, dans le cas d'algorithmes explicites, ou des données d'apprentissage, dans le cas d'algorithmes implicites. Cette discrimination se traduit par un traitement différent et parfois mauvais des utilisateurs ou des personnes à l'origine des données traitées. Dans le cadre de l'apprentissage algorithmique, on risque également de transformer une certaine représentativité statistique en une condition systématique, qui est par définition injustifiée pour une minorité. C'est le cas, par exemple, aux États-Unis [30], d'un algorithme d'évaluation du coût d'une couverture maladie qui, entraîné sur les prix des assurances de la population, a considéré un surcoût supplémentaire systématique pour une personne noire, s'appuyant sur les coûts moyens plus importants observés dans ces populations. Les principes d'égalité et d'équité, importants chez les médecins sont un enjeu en IA et doivent être garantis selon différentes méthodes, incluant une co-conception algorithmique par les scientifiques et les médecins, des tests d'échantillonnage sur les données de calibration ou d'apprentissage, ou des tests sur l'apprentissage lui-même en entraînant l'algorithme sur des données synthétiques différentes des données d'origine.

Reproductibilité du comportement algorithmique

Reproduire le comportement d'un modèle permet de garantir la robustesse de l'outil afin de réaliser des tests réguliers ou encore de rechercher les causes d'une réponse erronée. La robustesse se caractérise entre autres par la capacité de l'algorithme à reproduire dans le temps des résultats pour un jeu de données d'entrée identique, et à ne pas modifier sa réponse pour une variation légère de ce jeu de données. Dans les techniques d'apprentissage, le risque de ne pas pouvoir reproduire exactement le comportement algorithmique existe en raison de différences de données d'apprentissage, par exemple, ce qui est également symptomatique d'un apprentissage approximatif, voire faux. Parmi les articles de recherche sur l'IA appliquée à la médecine, seuls 19 % utilisent plusieurs jeux de données pour estimer la performance de leurs systèmes d'IA (ils sont 83 % pour les articles portant sur l'IA appliquée à la vision assistée par ordinateur et 66 % pour ceux portant sur l'IA appliquée à l'analyse du langage naturel) [31]. Les moyens mis en œuvre pour garantir la reproductibilité des systèmes d'IA en médecine représentent un enjeu pour le déploiement à grande échelle d'outils mais aussi pour le développement de la confiance du corps médical dans l'adoption de ces systèmes.

Conclusion

L'IA intervient dans pratiquement tous les domaines d'application de la médecine : de la recherche fondamentale et clinique [33] (➜) à la pratique hospitalière, aux examens médicaux, aux soins et à la logistique. Les avantages pour la médecine sont nombreux, comme l'amélioration de la précision des diagnostics et des pronostics, une médecine encore plus personnalisée et ciblée, des avancées dans les technologies d'observations et d'analyses, la démocratisation d'une médecine de qualité au niveau mondial, ou le développement d'outils chirurgicaux et autres robots d'assistance. Les médecins sont, par leur formation à l'éthique et leur multidisciplinarité, les plus à même de comprendre les enjeux de l'IA dans leur domaine et de (ré)agir, et éventuellement de participer à la construction de modèles et d'outils avec les scientifiques et les ingénieurs. Les défis sont technologiques et scientifiques mais également humains, avec le développement d'enseignements de l'IA pour le corps médical[11]. Cela permettrait aux médecins d'utiliser

(➜) Voir le Repères de C. Matuchansky, *m/s* n° 10, octobre 2019, page 797

[10] Traitement différent jugé comme injuste, dû en particulier à une non considération de certaines populations d'individus, provenant de l'usage d'une technologie numérique.

[11] Comme le diplôme universitaire (DU) « Intelligence artificielle appliquée en santé » qui a été ouvert en 2020 à l'université Paris-Descartes.

les systèmes d'IA de manière éclairée afin de compléter, défier, voire contredire, les résultats fournis par la *Machine*. ◊

SUMMARY

A brief history of artificial intelligence

For more than a decade, we have witnessed an acceleration in the development and the adoption of artificial intelligence (AI) technologies. In medicine, it impacts clinical and fundamental research, hospital practices, medical examinations, hospital care or logistics. These in turn contribute to improvements in diagnostics and prognostics, and to improvements in personalised and targeted medicine, advanced observation and analysis technologies, or surgery and other assistance robots. Many challenges in AI and medicine, such as data digitalisation, medical data privacy, algorithm explicability, inclusive AI system development or their reproducibility, have to be tackled in order to build the confidence of medical practitioners in these technologies. This will be possible by mastering the key concepts *via* a brief history of artificial intelligence. ◊

LIENS D'INTÉRÊT

L'auteure déclare n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.

RÉFÉRENCES

1. Turing A. Computing machinery and intelligence. *Mind* 1950 ; 49 : 433-60.
2. Costabala FS, Yaob J, Kuhla E. Predicting the cardiac toxicity of drugs using a novel multiscale exposure-response simulator. *Computer Methods Biomechanics Biomedical Engineering* 2018 ; 21 : 232-46.
3. Doblare M, Garcia JM, Gomez MJ. Modelling bone tissue fracture and healing: a review. *Engineering Fracture Mechanics* 2004 ; 71 (13–14).
4. Shim J, Grosberg A, Nawroth JC, *et al*. Modeling of cardiac muscle thin films: pre-stretch, passive and active behavior. *J Biomechanics* 2012 ; 45 : 832-41.
5. Shanahan M. *The technological singularity*. Essential knowledge series. Cambridge (MA) : The MIT Press, 2015.
6. De Dombal FT, Leaper DJ, Staniland JR, *et al*. Computer-aided diagnosis of abdominal pain. *Br Med J* 1972 ; 2 : 9-13.
7. Ravdin PM, Siminoff LA, Davis GJ, *et al*. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *J Clin Oncol* 2001 ; 19 : 980-91.
8. Velten K. *Mathematical modeling and simulation: introduction for scientists and engineers*. New York : Wiley, 2009 : 362 p.
9. Jean A, Nyein MK, Zheng JQ, *et al*. An animal-to-human scaling law for blast-induced traumatic brain injury risk assessment. *Proc Natl Acad Sci USA* 2014 ; 111 : 15310-5.
10. Yeo J, Jung GS, Tarakanova A, *et al*. Multiscale modeling of keratin, collagen, elastin and related human diseases: Perspectives from atomistic to coarse-grained molecular dynamics simulations. *Extreme Mechanics Letters* 2018 ; 20 : 112-24.
11. Febvay S, Socrate S, House MD. Biomechanical modeling of cervical tissue: a quantitative investigation of cervical incompetence. *Int Mechanical Engineering Congress Exposition* 2003 ; 399-400.
12. Tang A, Tam R, Cadrin-Chenevert A, *et al*. Canadian association of radiologists white paper on artificial intelligence in radiology. *Can Assoc Radiol J* 2018 ; 69 : 120135.
13. Bou Assi E, Nguyen DK, Rihana S, Sawan M. Towards accurate prediction of epileptic seizures: a review. *Biomedical Signal Processing Control* 2017 ; 34 : 144157.
14. Marcus G. The next decade in AI: four steps towards robust artificial intelligence. *arXiv* 2002 ; 06177 : 2020.
15. Alashwal H, El Halaby M, Crouse JJ, *et al*. The application of unsupervised clustering methods to Alzheimer's disease. *Front Comput Neurosci* 2019 ; 13-31.
16. Ng HP, Ong SH, Foong KWC, *et al*. Medical image segmentation using K-means clustering and improved watershed algorithm. *Proc IEEE Southwest Symposium Image Analysis Interpretation* 2006 ; 61-5.
17. LeCun Y, Boser B, Denker JS, *et al*. Backpropagation applied to handwritten zip code recognition. *Neural Comput* 1989 ; 1 : 541-51.
18. Erhan D, Bengio Y, Courville A, *et al*. Why does unsupervised pre-training help deep learning? *J Machine Learning Research* 2010 ; 11 : 625-60.
19. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015 ; 521 : 436-44.
20. Ekins S, Puhl AC, Zorn KM, *et al*. Exploiting machine learning for end-to-end drug discovery and development. *Nat Mater* 2019 ; 18 : 435-41.
21. Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017 ; 542 : 115–8.
22. Deo RC. Machine learning in medicine. *Circulation* 2015 ; 132 : 1920-30.
23. Majkowska A, Mittal S, Steiner DF, *et al*. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology* 2019 ; 294(2).
24. Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016 ; 316 : 2402-10.
25. Ding Y, Sohn JH, Kawczynski MG, *et al*. A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain. *Radiology* 2018 ; 290(2).
26. Xie Y, Ho J, Vemuri BC. multiple atlas construction from a heterogeneous brain MR image collection. *IEEE Trans Med Imaging* 2013 ; 32 : 628-35.
27. Zuluaga MA, Hush D, Edgar JF, *et al*. Learning from only positive and unlabeled data to detect lesions in vascular CT images. medical image computing and computer-assisted intervention – MICCAI 2011. *Lecture notes in computer science*. Berlin-Heidelberg : Springer. 2011 ; 6893.
28. Martin L, Muller B, Ortiz Suárez PJ, *et al*. CamemBERT: a tasty French language model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 2020.
29. Beaudouin V, Bloch I, Bounie D, *et al*. Flexible and context-specific AI explainability: a multidisciplinary approach. 2020. arXiv:2003.07703 [cs.CY].
30. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019 ; 366 : 447-53.
31. McDermott MBA, Wang S, Marinsek N, *et al*. Reproducibility in machine learning for Health. *International Conference on Learning Representations* 2019.
32. Haiech J. Parcourir l'histoire de l'intelligence artificielle, pour mieux la définir et la comprendre. *Med Sci (Paris)* 2020 ; 36 : 919-23.
33. Matuchansky C. Intelligence clinique et intelligence artificielle : une question de nuance *Med Sci (Paris)* 2019 ; 35 : 797-803.

TIRÉS À PART

A. Jean