Intelligence artificielle

L'**intelligence artificielle** (**IA**) est l'ensemble des programmes ou algorithmes permettant aux machines d'effectuer des tâches typiquement associées à l'intelligence humaine, comme l'apprentissage, le raisonnement, la résolution de problème, la perception ou la prise de décision. L'intelligence artificielle est également le champ de recherche visant à développer de telles machines ainsi que les systèmes informatiques qui en résultent.

Souvent classée dans le domaine des mathématiques et des sciences cognitives, l'IA fait appel à des disciplines telles que la neurobiologie computationnelle (qui a notamment inspiré les réseaux neuronaux artificiels), les statistiques, ou l'algèbre linéaire. Elle vise à résoudre des problèmes à forte complexité logique ou algorithmique. Par extension, dans le langage courant, l'IA inclut les dispositifs imitant ou remplaçant l'homme dans certaines mises en œuvre de ses fonctions cognitives[1].

Les applications de l'IA couvrent de nombreux domaines, notamment les moteurs de recherche, les systèmes de recommandation, l'aide au diagnostic médical, la compréhension du langage naturel, les voitures autonomes, les *chatbots*, les outils de génération d'images, les outils de prise de décision automatisée, les programmes compétitifs dans des jeux de stratégie et certains personnages non-joueurs de jeu vidéo[2].

Depuis l'apparition du concept, les finalités, les enjeux et le développement de l'IA suscitent de nombreuses interprétations, fantasmes ou inquiétudes, que l'on retrouve dans les récits ou films de science-fiction, dans les essais philosophiques[3] ainsi que parmi des économistes.

Intelligence artificielle

Partie de	Informatique, nouvelles technologies, raisonnement
Pratiqué par	Chercheur ou chercheuse en intelligence artificielle (d), ingénieur en intelligence artificielle
Champs	Intelligence artificielle symbolique connexionnisme ethical artificial intelligence (d) apprentissage automatique
Histoire	Histoire de l'intelligence artificielle

Définition

L'expression « intelligence artificielle », souvent abrégée par le sigle « IA » (ou « AI » en anglais, pour *artificial intelligence*) a été introduite en 1956 par John McCarthy, qui l'a définie en 2004[4] comme « la science et l'ingénierie de la fabrication de machines intelligentes, en particulier de programmes

informatiques intelligents. Elle est liée à la tâche similaire qui consiste à utiliser des ordinateurs pour comprendre l'intelligence humaine, mais l'IA ne doit pas se limiter aux méthodes qui sont biologiquement observables »[5].

Pour Marvin Lee Minsky, l'un de ses créateurs, l'IA est « la construction de programmes informatiques qui s'adonnent à des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisante par des êtres humains car elles demandent des processus mentaux de haut niveau tels que : l'apprentissage perceptuel, l'organisation de la mémoire et le raisonnement critique »[a, 6]. Cette définition combine l'aspect « artificiel » des ordinateurs et des processus informatiques, aux aspects « intelligents » d'imitation de comportements humains, notamment de raisonnement et d'apprentissage. Celui-ci est à l'œuvre dans les jeux, dans la pratique des mathématiques, dans la compréhension du langage naturel, dans la perception visuelle (interprétation des images et des scènes), auditive (compréhension du langage parlé) ou par d'autres capteurs, dans la commande d'un robot dans un milieu inconnu ou hostile.

Avant les années 2000, d'autres définitions sont proches de celle de Minsky, mais varient sur deux points fondamentaux[7] :

- les définitions qui lient l'IA à un aspect *humain* de l'intelligence et celles qui la lient à un modèle idéal d'intelligence, non forcément humaine, nommée *rationalité* ;
- les définitions qui insistent sur le fait que l'IA a pour but d'avoir *toutes les apparences* de l'intelligence (humaine ou rationnelle), et celles qui insistent sur le fait que le *fonctionnement interne* du système d'IA doit ressembler également à celui de l'être humain et être au moins aussi rationnel.

Le grand public confond souvent l'intelligence artificielle avec l'apprentissage automatique (*machine learning*) et l'apprentissage profond (*deep learning*). Ces trois notions diffèrent et sont en réalité imbriquées : l'intelligence artificielle englobe l'apprentissage automatique, qui lui-même englobe l'apprentissage profond[8].

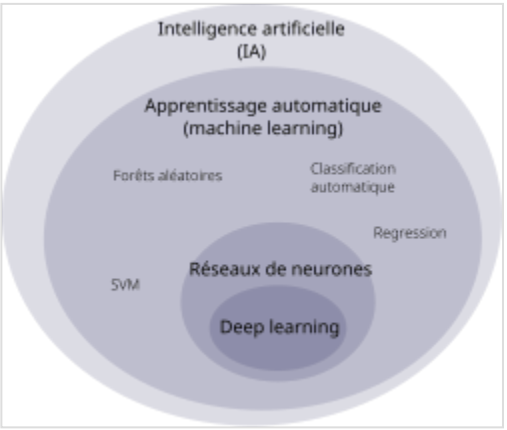


Diagramme de Venn montrant comment s'imbriquent les notions d'intelligence artificielle, d'apprentissage automatique et d'apprentissage profond.

Les définitions font souvent intervenir[9, 10] :

- une capacité à percevoir l'environnement et à prendre en compte la complexité du monde réel ;
- un traitement de l'information (collecter et interpréter des intrants, captés sous forme de données) ;
- des prises de décision (y compris dans le raisonnement et l'apprentissage), choix d'actions, exécution de tâches (dont d'adaptation, de réaction aux changements de contexte…), avec un certain niveau d'autonomie ;
- la réalisation d'objectifs spécifiques (raison ultime des systèmes d'IA).

Le groupe AI Watch note que les IA peuvent aussi être classées en fonction des familles d'algorithmes et/ou des modèles théoriques qui les sous-tendent, des capacités cognitives reproduites par l'IA, des fonctions exécutées par l'IA. Les applications de l'IA peuvent, elles, être classées en fonction du secteur socioéconomique et/ou des fonctions qu'elles y remplissent[9].

Une manière de définir l'intelligence artificielle est de considérer ses applications et les types de tâches qu'elle résout. Un rapport de la Commission européenne publié en 2020 présente une taxonomie classant les définitions de l'IA selon diverses tâches réalisées, telles que le raisonnement, l'apprentissage, la perception, etc.[11]. La même année, le professeur Jack Copeland propose une définition similaire, qui permet de distinguer plus clairement les facettes de l'IA, selon cinq catégories principales[11] :

1. L'apprentissage généralisé : l'IA apprend à partir de données diverses pour identifier des modèles et appliquer ces connaissances à de nouvelles situations, comme détecter des fraudes en ligne ;
2. Le raisonnement : cette capacité permet à l'IA de faire des prédictions et tirer des conclusions à partir des données, aidant dans des décisions comme la prédiction de comportements d'achat ;
3. La résolution de problèmes : l'IA trouve des solutions optimales pour des problèmes spécifiques, utilisée dans des contextes comme l'optimisation industrielle ou les stratégies de jeu ;
4. La perception : elle permet à l'IA de reconnaître et interagir avec son environnement, utilisée dans la robotique avancée et les véhicules autonomes pour naviguer et accomplir des tâches ;
5. La compréhension du langage : l'IA analyse et génère du langage à travers le NLP, utilisé dans des applications comme les assistants vocaux et les *chatbots* pour améliorer l'interaction utilisateur.

Techniques

Comportements prédéfinis (imitation d'intelligence)

Pour les intelligences artificielles servant principalement à donner une impression d'intelligence dans un cadre contrôlé, notamment pour les personnages non-joueurs des jeux vidéo, il est courant que l'apprentissage automatique ne soit pas utilisé. Un ensemble de fonctions et comportements plus précises et moins flexibles sont alors implémentés.

Il s'agit souvent de liste de textes ou paroles prédéfinis, aux déclenchements parfois conditionnels, par exemple un choix de mouvements suivant une série de règles et des déplacements de *pathfinding*.

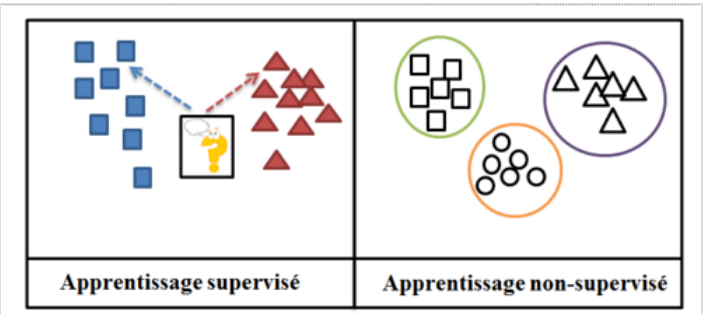
Apprentissage automatique

L'apprentissage automatique consiste à permettre au modèle d'IA d'apprendre à effectuer une tâche au lieu de spécifier exactement comment il doit l'accomplir[12]. Le modèle contient des paramètres dont les valeurs sont ajustées tout au long de l'apprentissage. La méthode de la rétropropagation du gradient est capable de détecter, pour chaque paramètre, dans quelle mesure il a contribué à une bonne réponse ou à une erreur du modèle, et peut l'ajuster en conséquence. L'apprentissage automatique nécessite un moyen d'évaluer la qualité des réponses fournies par le modèle[13]. Les principales méthodes d'apprentissage sont :

Apprentissage supervisé
: Un jeu de données annoté est utilisé pour entraîner l'algorithme. Il contient des données d'entrée fournies au modèle et les réponses correspondantes attendues, que le modèle est entraîné à produire[12]. Il est parfois difficile de se procurer suffisamment de données annotées avec les réponses attendues[14].

Apprentissage non supervisé

Un jeu de données est fourni au modèle, mais n'est pas annoté avec les réponses attendues. Le but peut par exemple être de regrouper les données similaires entre elles[12] (*clustering*).



En apprentissage supervisé, les réponses sont connues, tandis qu'en apprentissage non supervisé, l'algorithme découvre des structures dans les données par lui-même[12].

Apprentissage auto-supervisé

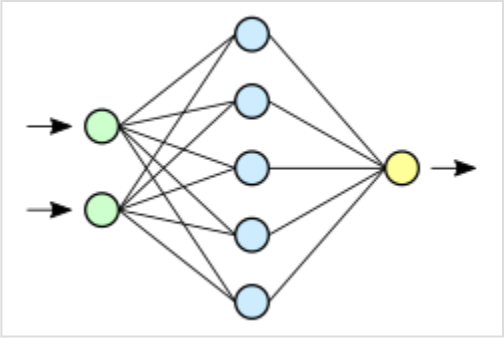
Un problème d'apprentissage supervisé est *automatiquement* généré à partir d'un jeu de données non annoté. Cela fonctionne souvent en cachant une partie des informations (des mots d'un texte, des morceaux d'images…) afin d'entraîner le modèle à les prédire[15].

Apprentissage par renforcement

L'agent est plongé dans un environnement où ce qu'il fait est évalué. Par exemple, un agent peut apprendre à jouer aux échecs en jouant contre lui-même, et le résultat (victoire ou défaite) permet à chaque itération d'évaluer s'il a bien joué. Il n'y a dans ce cas pas besoin de jeu de données[12].

Réseaux de neurones

Les réseaux de neurones artificiels sont inspirés du fonctionnement du cerveau humain : les neurones sont en général connectés à d'autres neurones en entrée et en sortie. Les neurones d'entrée, lorsqu'ils sont activés, agissent comme s'ils participaient à un vote pondéré pour déterminer si un neurone intermédiaire doit être activé et ainsi transmettre un signal vers les neurones de sortie. En pratique, pour l'équivalent artificiel, les « neurones d'entrée » ne sont que des nombres et les poids de ce « vote pondéré » sont des paramètres ajustés lors de l'apprentissage[16, 17].


Exemple de réseau de neurones comprenant deux neurones d'entrée (en vert), une couche « cachée » de neurones (en bleu) et un neurone de sortie (en jaune).

Hormis la fonction d'activation, les réseaux de neurones artificiels n'effectuent en pratique que des additions et des multiplications matricielles, ce qui fait qu'ils peuvent être accélérés par l'utilisation de processeurs graphiques[18]. En théorie, un réseau de neurones peut approximer n'importe quelle fonction[19].

Pour de simples réseaux de neurones à propagation avant (*feedforward* en anglais), le signal ne passe que dans une direction. Avec les réseaux de neurones récurrents, le signal de sortie de chaque neurone est réinjecté en entrée de ce neurone, permettant d'implémenter un mécanisme de mémoire à court terme[20].

Les réseaux neuronaux convolutifs, qui sont particulièrement utilisés en traitement d'images, introduisent une notion de localité. Leurs premières couches identifient des motifs relativement basiques et locaux comme des contours, là où les dernières couches traitent de motifs plus complexes et globaux[17].

Apprentissage profond

L'apprentissage profond (*deep learning* en anglais) utilise de multiples couches de neurones entre les entrées et les sorties, d'où le terme « profond »[21]. L'utilisation de processeurs graphiques pour accélérer les calculs et l'augmentation des données disponibles a contribué à la montée en popularité de l'apprentissage profond. Il est utilisé notamment en vision par ordinateur, en reconnaissance automatique de la parole et en traitement automatique des langues[22] (ce qui inclut les grands modèles de langage).

Grands modèles de langages

Les grands modèles de langage sont des modèles de langage ayant des milliards de paramètres. Ils reposent très souvent sur l'architecture transformeur[23].

Les transformeurs génératifs préentraînés (*Generative Pretrained Transformers* ou *GPT* en anglais) sont un type particulièrement populaire de grand modèle de langage. Leur « pré-entraînement » consiste à prédire, étant donnée une partie d'un texte, le *token* suivant (un *token* étant une séquence de caractères, typiquement un mot, une partie d'un mot, ou de la ponctuation). Cet entraînement à prédire ce qui va suivre, répété pour un grand nombre de textes, permet à ces modèles d'accumuler des connaissances sur le monde. Ils peuvent ensuite générer du texte semblable à celui ayant servi au pré-entraînement, en prédisant un à un les *tokens* suivants. En général, une autre phase d'entraînement est ensuite effectuée pour rendre le modèle plus véridique, utile et inoffensif. Cette phase d'entraînement (utilisant souvent une technique appelée RLHF) permet notamment de réduire un phénomène appelé « hallucination », où le modèle génère des informations d'apparence plausible mais fausses[24].

Avant d'être fourni au modèle, le texte est découpé en *tokens*. Ceux-ci sont convertis en vecteurs qui en encodent le sens ainsi que la position dans le texte. À l'intérieur de ces modèles se trouve une alternance de réseaux de neurones et de couches d'attention. Les couches d'attention combinent les concepts entre eux, permettant de tenir compte du contexte et de saisir des relations complexes[25].

Ces modèles sont souvent intégrés dans des agents conversationnels, aussi appelés chatbots, où le texte généré est formaté pour répondre à l'utilisateur. Par exemple, l'agent conversationnel ChatGPT exploite les modèles GPT-3.5 et GPT-4[26]. En 2023 font leur apparition des modèles grand public pouvant traiter simultanément différents types de données comme le texte, le son, les images et les vidéos, tel Google Gemini[27].

Recherche et optimisation

Certains problèmes nécessitent de chercher intelligemment parmi de nombreuses solutions possibles.

Recherche locale

La recherche locale, ou recherche par optimisation, repose sur l'optimisation mathématique pour trouver une solution numérique à un problème, en améliorant progressivement la solution choisie[28].