Intelligence artificielle

Pour les articles homonymes, voir A.I. Intelligence artificielle (film) et IA .

modifier - modifier le code - modifier Wikidata

L'intelligence artificielle (IA) est la capacité des machines à 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 comprennent 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.

Le terme « intelligence artificielle », souvent abrégé par le sigle « IA » (ou « AI » en anglais, pour artificial intelligence) a été créé par John McCarthy, qui l'a défini 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 »[4].

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],[5]. 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[6] :

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[7].

Les définitions font souvent intervenir[8],[9] :

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[8].

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.[10]. 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[10] :

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.

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[11]. 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[12]. Les principales méthodes d'apprentissage sont :

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[15],[16].

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[17]. En théorie, un réseau de neurones peut approximer n'importe quelle fonction[18].

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[19]. 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[16].

L'apprentissage profond (deep learning en anglais) utilise de multiples couches de neurones entre les entrées et les sorties, d'où le terme « profond »[20]. 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[21] (ce qui inclut les grands modèles de langage).

Les grands modèles de langage sont des modèles de langage ayant des milliards de paramètres. Ils reposent très souvent sur l'architecture transformeur[22].

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[23].

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[24].

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[25]. 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[26].

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

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[27].

En particulier, en apprentissage automatique, la descente de gradient permet de trouver une solution localement optimale, étant donné une fonction de coût à minimiser en faisant varier les paramètres du modèle. Elle consiste, à chaque étape, à modifier les paramètres à optimiser dans la direction qui permet de réduire le mieux la fonction de coût. La solution obtenue est localement optimale, mais il se peut qu'il y ait globalement de meilleures solutions, qui auraient pu être obtenues avec différentes valeurs initiales de paramètres[27]. Les modèles d'IA modernes peuvent avoir des milliards de paramètres à optimiser, et utilisent souvent des variantes plus complexes et efficaces de la descente de gradient[22].

Les algorithmes évolutionnistes, inspirés de la théorie de l'évolution, utilisent une forme de recherche par optimisation. À chaque étape, des opérations telles que la « mutation » ou le « croisement » sont effectuées de manière aléatoire pour obtenir différentes variantes, et les variantes les mieux adaptées sont sélectionnées pour l'étape suivante[27].

La recherche dans l'espace des états vise à trouver un état accomplissant l'objectif à travers un arbre des états possibles[28]. Par exemple, la recherche antagoniste est utilisée pour des programmes jouant à des jeux tels que les échecs ou le go. Elle consiste à parcourir l'arbre des coups possibles par le joueur et son adversaire, à la recherche d'un coup gagnant[29]. La simple recherche exhaustive est rarement suffisante en pratique vu le nombre d'états possibles. Des heuristiques sont utilisées pour prioriser les chemins les plus prometteurs[30].

La logique formelle est utilisée pour le raisonnement et la représentation des connaissances. Elle se décline en deux principales formes, la logique propositionnelle et la logique prédicative. La logique propositionnelle opère sur des affirmations qui sont vraies ou fausses, et utilise la logique connective avec des opérateurs tels que « et »,

« ou », « non » et « implique ». La logique prédicative étend la logique propositionnelle et peut aussi opérer sur des objets, prédicats ou relations. Elle peut utiliser des quantificateurs comme dans « Chaque X est un Y » ou « Certains X sont des Y »[31].

L'inférence logique (ou déduction) est le processus qui consiste à fournir — à l'aide d'un moteur d'inférence — une nouvelle affirmation (la conclusion) à partir d'autres affirmations connues comme étant vraies (les prémisses). Une règle d'inférence décrit les étapes valides d'une preuve ; la plus générale est la règle de résolution. L'inférence peut être réduite à la recherche d'un chemin amenant des prémisses aux conclusions, où chaque étape est une application d'une règle d'inférence[31]. Mais à part pour de courtes preuves dans des domaines restreints, la recherche exhaustive prend beaucoup de temps.

La logique floue assigne des valeurs de vérité entre 0 et 1, permettant de gérer des affirmations vagues, comme « il fait chaud »[32]. La logique non monotone permet d'annuler certaines conclusions[31]. Diverses autres formes de logique sont développées pour décrire de nombreux domaines complexes.

De nombreux problèmes en IA (raisonnement, planification, apprentissage, perception, robotique, etc.) nécessitent de pouvoir opérer à partir d'informations incomplètes ou incertaines[33].

Certaines techniques reposent sur l'inférence bayésienne, qui fournit une formule pour mettre à jour des probabilités subjectives étant données de nouvelles informations. C'est notamment le cas des réseaux bayésiens. L'inférence bayésienne nécessite souvent d'être approximée pour pouvoir être calculée[34].

Les méthodes de Monte-Carlo sont un ensemble de techniques pour résoudre des problèmes complexes en effectuant aléatoirement de nombreuses simulations afin d'approximer la solution[35].

Les réseaux de neurones peuvent aussi être optimisés pour fournir des estimations

probabilistes[36].

Des outils mathématiques précis ont été développés pour analyser comment des agents intelligents peuvent faire des choix et des plans en utilisant la théorie de la décision, la maximisation de l'espérance et la théorie de la valeur de l'information. Ces techniques comprennent des modèles tels que les processus de décision markoviens, la théorie des jeux et les mécanismes d'incitation[34].

De nombreux modèles d'IA ont pour but d'assigner une catégorie (classification), une valeur (régression) ou une action à des données fournies. Les méthodes de classification comprennent arbres de décision, k plus proches voisins, machine à vecteurs de support ou classification bayésienne naïve[37],[34]. Les réseaux de neurones peuvent également faire de la classification[38].

Comme précurseur à l'intelligence artificielle, divers automates ont été créés au cours de l'histoire, dont le canard de Vaucanson ou les automates d'Al-Jazari. Certains automates remontent à l'Antiquité et étaient utilisés pour des cérémonies religieuses[39]. Des mythes et rumeurs rapportent également la création d'êtres intelligents, par exemple les golems[40].

Des philosophes et mathématiciens comme Raymond Lulle, Leibniz ou George Boole ont cherché à formaliser le raisonnement et la génération d'idées[41].

Au XXe siècle, Alan Turing a notamment inventé un modèle de calcul par la suite appelé machine de Turing, exploré la notion de calculabilité et d'intelligence des machines, et proposé le « jeu de l'imitation » (test de Turing) pour évaluer l'intelligence de futures machines[41]. Le terme « intelligence artificielle » a été mis en avant par John McCarthy lors de la conférence de Dartmouth en 1956, où l'intelligence artificielle a été établie en tant que discipline à part entière[42],[43]. Dans les années qui ont suivi, des chercheurs ont proposé diverses preuves de concept, dans des situations spécifiques, de ce que les machines peuvent faire en théorie. Par exemple, le programme ELIZA

pouvait se faire passer pour un psychothérapeute, et le Logic Theorist pouvait démontrer des théorèmes[44].

La fin du siècle a été marquée par des périodes d'enthousiasme, et deux périodes de désillusion et de gel des financements appelées « hivers de l'IA »[45], la première de 1974 à 1980 et la seconde de 1987 à 1993.

Les systèmes experts ont été particulièrement populaires dans les années 1980, malgré leur fragilité et la difficulté à implémenter manuellement les bonnes règles d'inférences[44]. Des techniques d'apprentissage automatique se sont développées (réseaux de neurones, rétropropagation du gradient, algorithmes génétiques) ainsi que l'approche connexionniste[44]. Mais les faibles puissances de calcul et le manque de données d'entraînement limitait leur efficacité. Certains domaines n'ont progressivement plus été considérés comme faisant partie de l'intelligence artificielle, à mesure qu'une solution efficace était trouvée[46] ; un phénomène parfois appelé « effet IA ».

Les performances des ordinateurs s'accroissant continuellement. En 1996, pour la première fois, un supercalculateur a gagné plusieurs parties au jeu d'échec contre le champion du monde.

Dans les années 2000, le Web 2.0, le big data et de nouvelles infrastructures et capacités de calcul ont permis l'exploration de masses de données sans précédent. En 2005, le projet Blue Brain a débuté, ayant pour objectif de simuler le cerveau de mammifères[47]. En 2012, le réseau neuronal convolutif AlexNet a lancé l'utilisation de processeurs graphiques pour entraîner des réseaux de neurones, décuplant ainsi les capacités de calcul dédiées à l'apprentissage[48]. La même année, un programme a gagné quatre des cinq parties de Go disputées contre le meilleur joueur du monde. Des organisations visant à créer une intelligence artificielle générale ont vu le jour, comme DeepMind en 2010[49] et OpenAI en 2015[50]. Dès les années 2010, des outils

d'intelligence artificielle (spécialisée ou générative) ont accompli des progrès spectaculaires, mais restent loin des performances du vivant dans beaucoup de ses aptitudes naturelles, en particulier sur son aptitude à apprendre rapidement à partir d'un faible volume d'information (par induction), selon le magazine Slate en 2019[51].

En 2017, des chercheurs de Google ont proposé l'architecture transformeur, qui a servi de base aux grands modèles de langage. En 2018, Yann Le Cun, Yoshua Bengio et Geoffrey Hinton ont remporté le prix Turing pour leurs travaux sur l'apprentissage profond[52],[53].

En 2022, des programmes générant des images à partir de descriptions textuelles, comme Midjourney ou DALL-E 2, se sont popularisés[54]. La même année, l'agent conversationnel ChatGPT a connu une croissance inédite, gagnant un million d'utilisateurs en seulement cinq jours[55] et cent millions d'utilisateurs en deux mois[56], ce qui a accentué un phénomène de « course » à l'IA[57]. En 2023, les progrès rapides de l'IA ont suscité des inquiétudes quant à un potentiel risque d'extinction de l'humanité[58]. Des modèles de fondation « multimodaux », c'est-à-dire capables de traiter simultanément plusieurs modalités (texte, images, son) ont émergé, tels que Google Gemini[59] et GPT-4o[60].

L'intelligence artificielle générale (IAG) comprend tout système informatique capable d'effectuer ou d'apprendre pratiquement n'importe quelle tâche cognitive propre aux humains ou autres animaux[61]. Elle peut alternativement être définie comme un système informatique surpassant les humains dans la plupart des tâches ayant un intérêt économique[62].

L'intelligence artificielle générale a longtemps été considérée comme un sujet purement spéculatif[63]. Certains travaux de recherche ont déjà décrit GPT-4 comme ayant des « étincelles » d'intelligence artificielle générale[64],[65]. Les experts en intelligence artificielle affichent de larges désaccords et incertitudes quant à la date

potentielle de conception des premières intelligences artificielles générales (parfois appelées « intelligences artificielles de niveau humain »), leur impact sur la société, et leur potentiel à déclencher une « explosion d'intelligence »[66].

Un sondage de 2022 suggère que 90 % des experts en IA pensent que l'IAG a plus d'une chance sur deux d'être réalisée dans les 100 ans, autour d'une date médiane de 2061[67].

Une superintelligence artificielle est un type hypothétique d'intelligence artificielle générale dont les capacités intellectuelles dépasseraient de loin celles des humains les plus brillants[68]. Le philosophe Nick Bostrom note que les machines disposent de certains avantages par rapport aux cerveaux humains, notamment en ce qui concerne la mémoire, la vitesse (la fréquence des processeurs étant de l'ordre de dix millions de fois plus élevée que celle des neurones biologiques) et la capacité à partager des connaissances[69].

Dans ce contexte, un test est un moyen d'évaluer les capacités d'une intelligence artificielle à imiter certains comportements et raisonnements humains.

Dans le test de Turing, une machine et un humain répondent textuellement aux questions d'un interrogateur humain. L'interrogateur ne les voit pas mais doit déterminer à partir des réponses textuelles lequel des deux est la machine. Pour passer le test, la machine doit parvenir une bonne partie du temps à tromper l'interrogateur. Ce test a été conçu par Alan Turing en 1950 dans l'article « Computing Machinery and Intelligence ». Initialement appelé le « jeu de l'imitation », son but était de fournir une expérience concrète pour déterminer si les machines peuvent penser[70].

Imaginé par Steve Wozniak, le test du café consiste à placer un système intelligent dans un habitat américain moyen et à lui demander de faire un café[71]. La réussite du test implique donc plusieurs tâches comme l'orientation dans un environnement inconnu, déduire le fonctionnement d'une machine, trouver les ustensiles nécessaires…

Proposé par Ben Goertzel, le test de l'étudiant évalue la capacité d'un robot à s'inscrire dans un établissement d'enseignement supérieur, suivre les cours, passer les examens et obtenir le diplôme final[72].

Proposé par le chercheur Nils John Nilsson, le test de l'embauche consiste à faire postuler un système intelligent à un travail important pour l'économie, où il doit travailler au moins aussi bien qu'un humain[73].

Plusieurs prix Turing (ACM Turing Award) ont été attribués à des pionniers de l'intelligence artificielle, notamment :

En 2023, le magazine Time publie une liste de 100 personnalités influentes du domaine de l'IA et leurs biographies[76].

Les usages principaux qu'a successivement permis l'IA sont :

L'IA permet d'effectuer différents types de tâches, dont :

En combinant différents algorithmes, on est passé en quelques années de la reconnaissance de l'écriture manuscrite sur des formulaires de chèques bancaires (années 1990) à l'optimisation d'itinéraires entre deux ou plusieurs points, tenant compte des voies disponibles, de la longueur et de la vitesse probable sur chacun des segments, du moyen de locomotion et de souhaits particuliers (éviter les péages, points de passage obligés, caractéristiques touristiques, etc.), puis à la reconnaissance automatisée de documents dactylographiés ou manuscrits, permettant ensuite de les classer selon leur nature avant de les transmettre à des agents spécialisés, humains ou automates, capables d'y apporter la suite appropriée (réponses prédéfinies ou déclenchement d'une chaine de traitements). Lorsque les algorithmes ne parviennent pas à interpréter les données avec un degré suffisant de certitude, ils sont généralement soumis à des humains. Entre alors en jeu l'apprentissage : en corrélant les caractéristiques des données accumulées et des interprétations apportées ou corrigées par les agents humains, l'IA améliore progressivement ses facultés d'analyse

jusqu'à ce que ses prédictions soient suffisamment fiables pour qu'il ne soit plus nécessaire de les faire vérifier par un humain. Il en va de même pour la plupart des situations où l'IA, par ses performances et son coût, surpasse l'humain dans l'analyse de textes, d'images fixes ou animées, d'enregistrements sonores, de données scientifiques, commerciales ou industrielles. L'accroissement des performances conjuguées du matériel et des algorithmes permet en outre de traiter en temps réels les flux de données comme la voix humaine et les images de caméras, ouvrant ainsi la voie à la traduction simultanée, la transcription textuelle, l'identification des individus, la détection de comportements anormaux ou illégaux, voire le dialogue avec des humains.

L'intelligence artificielle est désormais utilisée dans de nombreux domaines. Ses capacités permettent notamment d'automatiser et d'optimiser des tâches complexes, de traiter et d'analyser de vastes quantités de données, et d'améliorer la prise de décision[79]. L'adoption de l'intelligence artificielle est en forte expansion dans les années 2020, stimulée par les avancées en intelligence artificielle générative, en particulier dans les grands modèles de langage, dont la polyvalence ouvre la voie à de nouveaux cas d'usage[80], ainsi que dans les domaines de la programmation informatique (génération de code), de l'image et du son (photos et vidéos de synthèse, animation du visage en association avec la synthèse vocale) et dans la prévention et l'atténuation des risques (diagnostic médical, situations de crise, accidents industriels, catastrophes naturelles, etc.).

Plusieurs grands noms de la finance se sont montrées intéressées par de telles technologies, avec des projets comme ceux de Bridgewater Associates où une intelligence artificielle va gérer entièrement un fonds[81] ou encore la plateforme d'analyse prédictive Sidetrade.

Sont également développés des systèmes de trading algorithmique, dont les gains de

vitesses permis par l'automatisation peuvent leur donner un avantage par rapport à des traders humains, en particulier grâce au trading à haute fréquence[82].

Le domaine militaire utilise de plus en plus l'intelligence artificielle, notamment pour le pilotage automatique, le guidage de missiles, l'identification, le commandement, l'aide à la décision[83], la cyberguerre et la cyberdéfense[84], ou pour la documentation et les processus administratifs[85].

Cette course aux armements est notamment illustrée par le projet Maven aux États-Unis[86]. Dès 2015, une IA nommée ALPHA a « systématiquement triomphé d'un pilote de chasse chevronné »[87]. En 2018, l'ONU a tenté d'interdire les systèmes d'armes létales autonomes « avant qu'il ne soit trop tard », mais peine encore en janvier 2024 à établir le moindre cadre légal international face aux réticences, notamment de la Russie, des États-Unis et d'Israël, dont le veto peut bloquer une proposition[88]. Des drones tueurs pilotés par intelligence artificielle ont été utilisés lors du conflit ukraino-russe[89]. Le 10 janvier 2024, OpenAI a modifié ses conditions d'utilisation ; il continue d'interdire l'usage de ses services tels que ChatGPT à des fins illégales ou de destruction des biens, mais n'interdit plus explicitement les usages militaires[90]. L'intelligence artificielle générative est parfois utilisée par les institutions militaires pour rédiger plus vite la documentation, mais son adoption est limitée par la confidentialité des données, les réglementations, ou le risque d'erreur et le besoin de vérification[85].

En France, la force opérationnelle IA du ministère des Armées rend en septembre 2019 un rapport détaillant sa stratégie, qui inclut la création d'une Cellule de coordination de l'intelligence artificielle de défense (CCIAD) rattachée à l'Agence de l'innovation de défense[91]. La loi de programmation militaire prévoit un budget de 700 millions d'euros pour les missions en faveur de l'IA, soit une moyenne de 100 millions par an[92]. En 2021, la France est opposée aux armes totalement autonome, estimant qu'il

est au moins nécessaire de conserver une supervision humaine[93],[94]. En février 2025, le ministère des Armées lance sa plateforme GenIAl.intradef (co-développée depuis 2022 par plusieurs entités), destinée à améliorer le travail quotidien des militaires et agents civils associés à l'Armée. Cette IA alimente un agent conversationnel et un convertisseur audio-texte ; elle peut lire et analyser des documents et des images et traduire des messages[95].

Dans le cadre de la guerre Israël-Hamas de 2023-2024, Israël a utilisé deux systèmes d'IA pour générer des cibles à frapper : Habsora (soit « l'évangile ») a été utilisé pour dresser une liste de bâtiments à cibler, tandis que Lavander a produit une liste de 37 000 personnes à cibler[96],[97]. La liste des bâtiments comprenait les maisons privées à Gaza de personnes soupçonnées d'être affiliées à des membres du Hamas. Les responsables de Tsahal affirment que le programme répond au problème antérieur du manque de cibles de l'armée de l'air. Auparavant, Tsahal était en mesure d'identifier 50 cibles par an, tandis que le programme en produit 100 par jour[97]. La combinaison de la technologie de ciblage de l'IA et du changement de politique consistant à éviter les cibles civiles a entraîné un nombre sans précédent de morts civils palestiniens[98],[99].

La médecine a aussi vu de grands progrès grâce à l'utilisation de systèmes d'aide au diagnostic ou de diagnostic automatisé[100].

En 2018, Google DeepMind, filiale de Google spécialisée dans la recherche avancée en intelligence artificielle, a publié les résultats d'une expérimentation d'intelligence artificielle pouvant détecter les maladies oculaires. Les résultats indiquent que l'IA le fait avec une marge d'erreur plus faible que les ophtalmologues[101].

Google DeepMind a également conçu AlphaFold, un système d'intelligence artificielle utilisant l'apprentissage profond qui permet de prédire la façon dont des protéines se replient. Les protéines sont composées de chaînes d'acides aminés et la façon dont

elles se replient détermine leur fonction. Cette nouvelle méthode, introduite en 2018 et améliorée en 2020, est nettement plus rapide que les approches traditionnelles et a été décrite comme une révolution dans le domaine de la recherche en biologie[102],[103].

La France crée en 2019 le Health Data Hub afin d'encadrer et de faciliter l'utilisation des données de santé dans la recherche[104].

En 2023, la version de ChatGPT reposant sur GPT-4 s'est montrée facilement capable d'obtenir le diplôme de médecin aux États-Unis[105].

L'intelligence artificielle (IA) est de plus en plus utilisée dans le domaine médical, transformant les pratiques cliniques et facilitant le diagnostic, le traitement et la gestion des maladies. Les algorithmes d'apprentissage automatique, un sous-ensemble de l'IA, analysent de grandes quantités de données médicales pour en extraire des modèles et des tendances. Cette capacité d'analyse a permis des avancées significatives dans le diagnostic précoce de maladies comme le cancer et les maladies cardiaques, où les systèmes basés sur l'IA peuvent détecter des anomalies à partir d'images médicales avec une précision parfois supérieure à celle des praticiens humains[106].

Un des domaines où l'IA s'avère particulièrement efficace est celui de l'imagerie médicale. Des outils comme les réseaux de neurones convolutifs sont utilisés pour analyser les radiographies, les IRM et autres types d'imageries, permettant de repérer les signes précoces de pathologies complexes. Par exemple, une étude menée par McKinney et al. (2020) a démontré que l'IA pouvait réduire les faux positifs et négatifs dans le dépistage du cancer du sein, améliorant ainsi la précision et la rapidité du diagnostic[107].

De plus, l'IA joue un rôle clé dans la médecine personnalisée, où elle aide à adapter les traitements en fonction des caractéristiques génétiques et biologiques des patients. Grâce à l'analyse des données génomiques, les médecins peuvent élaborer des

thérapies spécifiques pour des maladies comme le cancer ou les maladies génétiques rares. L'IA facilite également le développement de nouveaux médicaments, en identifiant des molécules potentielles pour le traitement de certaines maladies, un processus qui nécessitait auparavant plusieurs années[108].

Cependant, malgré ses promesses, l'usage de l'IA en médecine soulève des questions éthiques et de sécurité, notamment en ce qui concerne la protection des données des patients et la transparence des algorithmes[109]. La Food and Drug Administration (FDA) aux États-Unis et d'autres organismes réglementaires travaillent à définir des cadres pour encadrer l'utilisation sécurisée et éthique de l'IA dans les soins de santé.

Un usage de l'IA se développe dans le domaine de la prévention des crimes et délits. La police britannique, par exemple, développe une IA de ce genre, annoncée comme pouvant être opérationnelle dès mars 2019[110]. Baptisée National Data Analytics Solution (Solution nationale d'analyse de données ou NDAS), elle repose sur l'IA et des statistiques et vise à estimer le risque qu'une personne commette un crime ou en soit elle-même victime, pour orienter les services sociaux et médicaux qui peuvent la conseiller.

L'usage d'outils de prédiction des crimes à partir des données préalablement existantes est toutefois l'objet de controverses, compte tenu des biais sociaux (notamment raciaux) qu'il comporte[111]. En effet, la logique d'identification de schémas propre à ces technologies joue un rôle de renforcement des préjugés déjà existants.

L'intelligence artificielle (IA) est de plus en plus exploitée dans le domaine du cybercrime, comme le révèle une étude de la société spécialisée en cybersécurité SlashNext. Cette tendance croissante à l'utilisation de l'IA pour commettre des crimes en ligne montre une sophistication accrue des attaques. L'entreprise SlashNext a notamment identifié l'usage de deux IA malicieuses, FraudGPT et WormGPT, tout en suggérant que ces découvertes ne représentent que la partie visible d'une menace

potentiellement colossale. Lors de leurs investigations, les chercheurs ont également mis en lumière l'existence de DarkBart et DarkBert[b], deux chatbots malveillants en développement, capables d'intégrer la technologie de reconnaissance d'images de Google Google Lens. Ces chatbots pourraient envoyer du texte et des images, et participer à des attaques d'ingénierie sociale avancées. Face à cette menace croissante, les solutions actuelles de lutte contre le cybercrime semblent insuffisantes, estime un rapport d'Immunefi, qui souligne les limites de certaines IA, telles que ChatGPT, dans la détection des exploits[112].

Le droit fait appel à l'IA dans la perspective de prédire les décisions de justice, d'aider à la décision et de trancher les cas simples[113]. L'Estonie a par exemple développé une intelligence artificielle capable de prendre des décisions de justice sur des délits mineurs[114]. Les États-Unis utilisent par ailleurs dans certaines juridictions le système COMPAS (en)(Correctional Offender Management profiling for Alternative Sanctions), un système d'aide à la prise de décision pour les juges[114]. Plusieurs startups se sont spécialisées dans ce créneau, créant le domaine de la legaltech[115].

Le domaine de la logistique a vu certains projets utilisant de l'intelligence artificielle se développer notamment pour la gestion de la chaîne logistique (supply chain) ou des problématiques de livraison telle celle du dernier kilomètre[116].

L'intelligence artificielle est également fortement utilisée dans le domaine des transports en commun, car elle permet de faciliter la régulation et la gestion du trafic au sein de réseaux de plus en plus complexes, comme le système UrbanLoop en cours d'étude dans la ville de Nancy[117].

Même si les problèmes d'optimisation de temps de trajet ou de transports font partie des plus anciennes applications de solutions à base d'intelligence artificielle (voir le problème du voyageur de commerce ou l'algorithme de Dijkstra), les avancées récentes, notamment en apprentissage profond, ont permis des progrès significatifs en

matière de précision. Certains projets comme Google Maps utilisent par exemple des systèmes d'IA en milieu urbain pour compenser la réflexion du signal GPS sur les immeubles avoisinants[118], ou pour cartographier des zones où peu d'informations sont disponibles[119],[120].

Plusieurs entreprises ont par ailleurs annoncé avoir développé des programmes de recherche en voiture autonome, notamment Google à travers sa filiale Waymo, l'entreprise française Navya ou encore Tesla.

Les systèmes intelligents deviennent monnaie courante dans de nombreuses industries. Plusieurs tâches peuvent leur être confiées, notamment celles considérées comme trop dangereuses pour un humain[121]. Certains applications se concentrent sur les systèmes de maintenance prédictive, permettant des gains de performance grâce à une détection des problèmes de production en amont.

La robotique a recours à l'intelligence artificielle à plusieurs égards, notamment pour la perception de l'environnement (objets et visages), l'apprentissage et l'intelligence artificielle développementale[122],[123].

L'interaction homme-robot manque encore souvent de naturel et est un enjeu de la robotique. Il s'agit de permettre aux robots d'évoluer dans le monde dynamique et social des humains et d'échanger avec eux de façon satisfaisante[122]. L'échange nécessite également, à l'inverse, une évolution du regard que les humains portent sur les robots ; selon Véronique Aubergé, chercheuse à l'Université Grenoble-Alpes « la vraie révolution n'est pas technologique, elle est culturelle ». D'ores et déjà, à travers les robots dotés d'intelligence artificielle, tel Google Home, les utilisateurs combleraient un isolement social[122].

L'intelligence artificielle est par exemple utilisée pour animer les personnages non-joueurs de jeux vidéo, qui sont conçus pour servir d'opposants, d'aides ou d'accompagnants lorsque des joueurs humains ne sont pas disponibles ou désirés.

Différents niveaux de complexité sont développés, d'une simple assistance à un comportement complexe imitant (ou dépassant) les meilleurs joueurs humains.

Dès la fin des années 1980, des artistes s'emparent de l'intelligence artificielle pour donner un comportement autonome à leurs œuvres. Les Français Michel Bret, Edmond Couchot et Marie-Hélène Tramus sont des pionniers, ainsi qu'en témoignent des œuvres comme La Plume et Le Pissenlit (1988)[124], puis La Funambule (2000), animée par un réseau de neurones. L'Américain Karl Sims, en partenariat avec la société Thinking Machines, crée en 1993 Genetic Images, machines incorporant[Comment ?] des algorithmes génétiques. Le couple franco-autrichien Christa Sommerer et Laurent Mignonneau crée depuis le début des années 1990 de nombreuses œuvres dans le champ de la vie artificielle, parmi lesquelles Interactive plant growing (1992) ou A-Volve (1994)[réf. nécessaire]. Le Français Florent Aziosmanoff propose quant à lui de considérer que l'emploi de l'intelligence artificielle dans l'art conduit à l'émergence d'une nouvelle discipline d'expression, qu'il nomme le Living art[125].

Le 23 octobre 2018, la société de vente aux enchères Christie's met en vente le tableau Portrait d'Edmond de Belamy réalisé par une intelligence artificielle à l'aide de réseaux antagonistes génératifs. La peinture est signée par la formule mathématique à l'origine de sa création (« Min (G) max (D) Ex [log (D(x))] + Ez [log(1-D(G(z)))] »)[126]. Cette vente soulève de nombreux débats sur son statut de création artistique et sur l'auteur de l'œuvre : il peut être l'intelligence artificielle elle-même ou les trois créateurs qui l'ont programmée[127].

Des réseaux antagonistes génératifs ont parfois été utilisés pour créer de fausses images réalistes, comme avec le générateur de visages StyleGAN introduit en 2018[128], ou avec « Terre Seconde » de Grégory Chatonsky qui imagine en 2019 une version alternative de la planète Terre[129].

Dès 2022 apparaissent des modèles d'intelligence artificielle qui sont capables de créer

des images réalistes à partir de descriptions textuelles, comme Midjourney, Stable Diffussion et DALL-E[130],[131]. En mars 2023, des fausses photos d'actualité sont ainsi générées et diffusées sur Internet, mettant en scène des personnalités dans des situations extravagantes (le président Macron ramassant des poubelles, Donald Trump arrêté par des policiers[132], le pape François habillé en doudoune blanche[133]). Elles deviennent rapidement virales, augmentant les craintes de manipulation de l'opinion[134]. Cela pose aussi des questions de droits d'auteur[135].

De plus en plus de romans ont été coécrits avec une IA générative, tels que Internes en 2022[137] ou (« La Tour de la compassion de Tokyo »), qui a reçu le prix Akutagawa en 2024[138].

En février 2024, le modèle Sora de OpenAI s'est montré capable de générer des vidéos relativement réalistes[139].

Des modèles d'IA capables de générer un morceau de musique à partie d'une description du style souhaité ont également fait leur apparition, comme Suno AI en 2023 et Udio en 2024[140].

La domesticité, avec des robots employé de maison[141], ou pour certaines tâches précises comme en domotique.

En programmation informatique, notamment pour la maintenance prévisionnelle, l'autocomplétion ou l'aide au développement[142].

En journalisme : des IA (appelées improprement « robots journalistes ») pourraient à terme aider les journalistes en les débarrassant de certaines tâches, notamment la veille, le bâtonnage de dépêches ou la vérification des fake news[143].

La Corée du Sud propose la toute première animatrice télé virtuelle en novembre 2020 lors d'un JT[144].

En design : la conception assistée par ordinateur fait depuis longtemps appel à des algorithmes d'optimisation. En 2019, le créateur Philippe Starck lance ainsi une chaise

développée en collaboration avec la société Autodesk, la « A.I.chair »[145].

Les succès en IA encouragent les spéculations. Dans les milieux technophiles, on verse en général dans l'enthousiasme, le mouvement transhumaniste en est la meilleure expression. Mais certains s'inquiètent et s'interrogent, parfois alarmistes, y compris dans la sphère de la haute technologie. Ainsi, des figures réputées telles que Bill Gates — ancien PDG de Microsoft et « figure emblématique de la révolution informatique de la fin du XXe siècle »[146] — pensent qu'il faut rester très prudent quant aux développements futurs de ces technologies, qui pourraient devenir liberticides ou dangereuses.

Le développement de l'intelligence artificielle suscite un grand nombre de questions, notamment en ce qui concerne la possibilité pour les IA ou algorithmes d'accéder un jour à la conscience, d'éprouver des émotions ou de finalement se substituer aux humains. Certaines réactions sont ouvertement optimistes, d'autres sont au contraire pessimistes. En 2016, l'INRIA publie un premier Livre blanc consacré à l'IA[147].

Le philosophe Daniel Andler considère en 2023 que le rêve d'une intelligence artificielle qui rejoindrait celle de l'homme est une chimère, pour des causes conceptuelles et non techniques. L'intelligence humaine va selon lui plus loin que la simple résolution de problèmes : toutes ses autres tâches, reposant sur des affects, de la spontanéité et une forme de contingence, ne seront jamais accessibles à une intelligence non humaine[148].

Une description d'un possible avenir de l'intelligence artificielle a été faite par le statisticien anglais Irving John Good :

« Supposons qu'existe une machine surpassant en intelligence tout ce dont est capable un homme, aussi brillant soit-il. La conception de telles machines faisant partie des activités intellectuelles, cette machine pourrait à son tour créer des machines meilleures qu'elle-même ; cela aurait sans nul doute pour effet une réaction en chaîne

de développement de l'intelligence, pendant que l'intelligence humaine resterait presque sur place. Il en résulte que la machine ultra intelligente sera la dernière invention que l'homme aura besoin de faire, à condition que ladite machine soit assez docile pour constamment lui obéir. »

— Irving John Good[149]

Cette hypothétique courte période de progrès drastique dont il est difficile de prédire les conséquences a été nommée « singularité ». Elle a été étudiée par Vernor Vinge dans les années 90 et par Ray Kurzweill[150] dans les années 2000[151]. Ce concept est central pour de nombreux transhumanistes, qui s'interrogent sur les dangers ou les espoirs d'un tel scénario, certains allant jusqu'à envisager l'émergence d'un « dieu » numérique appelé à prendre le contrôle du destin de l'humanité, ou à fusionner avec elle[151]. En 2014, Nick Bostrom a popularisé le concept de superintelligence artificielle[152].

Le développement de l'intelligence artificielle génère de l'enthousiasme, mais aussi de vives inquiétudes. Certains auteurs de science-fiction, tels Isaac Asimov, William Gibson ou Arthur C. Clarke, sur le modèle du récit de L'Apprenti sorcier, décrivent le risque d'une perte de contrôle des humains sur le processus technique. Dans les années 2010, différents intellectuels ont également pris position, notamment l'astrophysicien Stephen Hawking, selon qui l'intelligence artificielle risque réellement de surpasser un jour l'intelligence humaine et de finir par dominer l'humanité, voire de s'y substituer[154],[155]. Il pose en novembre 2017 au salon technologique Web Summit de Lisbonne la question suivante « Serons-nous aidés par l'intelligence artificielle ou mis de côté, ou encore détruits par elle ? »[156].

Dans le milieu de la haute technologie, certains expriment publiquement des craintes similaires. C'est ainsi le cas, en 2015, de Bill Gates, Elon Musk et Bill Joy[157]. Selon le spécialiste américain de l'informatique Moshe Vardi en 2016, l'intelligence artificielle

pourrait mettre 50 % de l'humanité au chômage. « Nous approchons d'une époque où les machines pourront surpasser les hommes dans presque toutes les tâches »[158].

Hilary Mason, directrice de la recherche à Cloudera, a critiqué en 2018 le sensationnalisme entourant l'intelligence artificielle et prône une vision pragmatique et opérationnelle de cette technologie[159].

Ajouter un mécanisme d'arrêt pourrait ne pas suffire face à une IA suffisamment avancée, qui pourrait s'avérer en mesure de cacher des intentions dangereuses, de manipuler ses détenteurs, de désactiver le mécanisme d'arrêt ou encore de se dupliquer. Selon Nick Bostrom en 2015, la seule solution viable à long terme consiste à trouver comment aligner les intelligences artificielles avec des valeurs humaines et morales[160] :

« nous ne devrions pas être confiants dans notre capacité à garder indéfiniment un génie superintelligent enfermé dans une bouteille. Je crois que la réponse ici est de trouver comment créer une IA superintelligente de sorte que si — ou plutôt quand — elle s'échappe, elle reste sans danger, parce qu'elle est fondamentalement de notre côté, elle partage nos valeurs. »

— Nick Bostrom

Roman V. Yampolskiy, professeur de science informatique à l'Université de Louisville, évoque pourquoi et comment une IA obtient un résultat, pour s'assurer qu'il corresponde bien à l'attendu, sans biais : « si nous nous habituons à accepter les réponses de l'IA comme des paroles d'oracles ne nécessitant pas d'explication, alors nous serons incapables de vérifier si ces résultats ne sont pas biaisés ou manipulés »[161].

En mai 2023, une déclaration du Center for AI Safety (« Centre pour la sûreté de l'IA ») affirme que réduire le risque d'extinction de l'humanité lié à l'IA devrait être une priorité mondiale, au même titre que pour d'autres risques civilisationnels tels les pandémies

ou les guerres nucléaires. Elle est signée par des dirigeants de laboratoires d'IA comme OpenAI, Google DeepMind ou Anthropic, ainsi que par des chercheurs en intelligence artificielle[162],[163].

Comme l'explique l'historien François Jarrige, la critique de l'intelligence artificielle trouve son origine dans celle - plus ancienne et plus générale - des techniques et de la technologie, dont Lewis Mumford (aux États-Unis)[164], Jacques Ellul (en France)[165] et Günther Anders (en Allemagne)[166] sont au XXe siècle les principaux instigateurs, et qui inspire aujourd'hui différents cercles militants (en France, par exemple : Pièces et Main d'Œuvre[167] et Technologos[168])[169].

Dans un rapport en date de février 2018 intitulé The Malicious Use of Artificial Intelligence 26 experts spécialistes en intelligence artificielle mettent en garde contre les dangers d'un usage criminel de l'IA : augmentation de la cybercriminalité, conduire à des utilisations de drones à des fins terroristes, manipulation de masse, etc.[170].

Un autre problème est l'énorme quantité de ressources rares, de serveurs et d'énergie consommée par l'informatique sous-jacente à l'IA. Google admet en 2024 qu'il lui sera très difficile de tenir ses engagements de neutralité carbone, car depuis 2019 ses émissions de gaz à effet de serre ont augmenté de 48 % du fait du développement de l'IA[171].

Afin de réduire leur empreinte carbone, certains fournisseurs d'IA (Microsoft, Amazon et Oracle) se tournent vers les marchés volontaires de la compensation carbone, et plus largement de solutions de décarbonation, comme celles de capture du CO2[172]. Ils s'orientent également vers les centrales nucléaires pour couvrir leur besoin en électricité[173]. Microsoft a notamment signé en 2024 un contrat avec Constellation Energy pour l'achat de 837 MWe de capacité électrique fournie par la centrale nucléaire de Three Mile Island, pour une durée de vingt ans à partir de 2028[174].

Divers projets open source d'IA ont été menés par Hugging Face[172], EleutherAI[175],

Google[176], ou Meta[175].

En mars 2023, comme alternative aux géants du Web et du cloud computing, qui ont le plus de pouvoir et d'influence, Mozilla annonce vouloir investir 30 millions de dollars dans un projet baptisé Mozilla.ai, qui est à la fois une startup et une communauté, indépendante des géants de la tech et de la recherche universitaire[177]. Le projet vise à créer, dans le respect des valeurs de son manifeste (notamment transparence et responsabilité), un système d'IA « open source, digne de confiance et indépendant »[178].

De nombreux grands modèles de langage comme Mistral[179], Llama 3, Vicuna et Falcon[180] sont rendus open weight, ce qui signifie que l'architecture et les paramètres entraînés du modèle d'IA sont rendus publics (open source impliquerait notamment de partager les données d'entraînement, ce qui souvent n'est pas le cas)[181]. Ces modèles peuvent être librement ajustés, ce qui permet notamment aux entreprises de les spécialiser pour leurs propres données et pour leur cas d'usage[182]. Ces modèles d'IA facilitent l'innovation et la recherche, mais peuvent facilement être détournés. Ils peuvent être réentraînés de sorte à rendre inefficaces les mesures de sécurité, telles que le refus de répondre à une requête dangereuse. Certains chercheurs estiment ainsi que si des modèles développent un jour des capacités dangereuses, comme le fait de faciliter drastiquement les cyberattaques ou le bioterrorisme, ils ne devraient pas être rendus open weight, d'autant plus qu'une fois diffusé sur internet, un modèle ne peut en général plus être supprimé partout[183],[181].

Historiquement, l'innovation technologique a généralement été accompagnée d'une croissance de la productivité et de la création de nouveaux emplois pour compenser les pertes[184]. Cependant, de nouvelles inquiétudes ont émergé avec l'essor de l'IA générative, telle que ChatGPT, capable de manipuler du texte, des images ou du code informatique[185].

Certains économistes sont sceptiques à l'idée d'un chômage de masse, citant des précédents historiques où le marché du travail s'est adapté à des bouleversements technologiques[186],[187]. D'autres estiment que l'IA générative représente un changement plus profond, qui ne consiste pas seulement à automatiser des tâches répétitives. L'IA peut être appliquée à tous les secteurs, et le nombre de tâches où l'humain reste plus compétent que l'IA est amené à diminuer[188]

Une étude de Goldman Sachs a estimé en 2023 que les deux tiers des travailleurs européens sont exposés à divers degrés d'automatisation, et qu'un quart des emplois pourraient être remplacés par l'IA. L'étude prédit notamment que les professions administratives et légales seront particulièrement touchées, là où les métiers manuels en extérieur seront peu affectés[185].

Une solution envisagée dans le scénario d'un chômage de masse est celle d'une forme de redistribution des richesses avec un revenu universel. Les financements pourraient dans ce cas venir d'une taxe sur les richesses produites par les machines[189].

L'IA s'appuie sur l'analyse de données qui ne peuvent généralement pas représenter fidèlement la réalité, soit parce ces données ne proviennent que de sources numériques existantes (le plus souvent sur Internet) alors que quantités d'informations pertinentes pour le domaine considéré n'y figurent pas, soit parce que certaines catégories de données pourtant pertinentes n'ont pas été prises en compte, ou que des données non pertinentes l'ont été. Il en résulte alors une inadéquation entre les données utilisées pour l’entraînement des algorithmes et les données cibles sur lesquelles l’algorithme devra opérer, et par conséquent des erreurs de diagnostic qui peuvent, dans certains cas, induire des décisions inappropriées ou injustes[190],[191],[192].

Les réseaux sociaux et les bots ont favorisé la propagation de nombreuses fausses croyances et des dérives dans les débats démocratiques, lesquelles ont entrainé une

certaine défiance vis-à-vis de la science, des élites intellectuelles et des médias d'information traditionnels[193][réf. incomplète]. En conséquence, la progression et l'adoption rapides des techniques sous-jacentes de l'IA inspirent de multiples craintes quant aux impacts que celle-ci pourrait avoir sur les comportements individuels et collectifs. Ces conditions ouvrent la possibilité pour des opérateurs majeurs de la société de fausser le réel pour influencer ou manipuler les citoyens. Les acteurs économiques emploient l'IA pour influencer les consommateurs et pour optimiser le travail de leurs employés au point de les transformer en robots[194][source secondaire souhaitée]. L'éventualité que l'IA intervienne dans les débats d'idées ou dans la conduite des affaires individuelles ou collectives mène à redouter qu'elle n'altère ou n'affaiblisse les institutions politiques[193][réf. incomplète] et les pouvoirs, voire les capacités des humains[195][source secondaire souhaitée].

Dès 2010, Nicholas Carr alertait sur l'usage intensif des outils numériques, qui modifie notre manière de penser et de traiter l'information, ce qui peut avoir des conséquences à long terme sur notre cognition, notamment affaiblir notre mémoire et notre capacité à comprendre les informations de manière approfondie et affecter la créativité, l'empathie et le débat intellectuel[196].

Les principales menaces couramment envisagées, consécutives à des biais ou au détournement des algorithmes d'IA sont[197],[198] :

De plus, l'IA produit parfois des résultats contre-intuitifs, bien que fiables, ou propager de fausse croyances, qui contribuent, parmi d'autres facteurs, à éroder la confiance d'une partie de la population dans le discours rationnel, les théories scientifiques, le système médiatique, les institutions et les élites en général, et à la remise en cause de leurs légitimités[199] et du processus démocratique[193][réf. incomplète].

Selon Yuval Noah Harari, jusqu'ici, du fait de défauts d'alignement et faute de mécanismes suffisants d'autocorrection, les objectifs qui sont assignés aux algorithmes

des réseaux sociaux ont déjà induit ou influencé phénomènes de société indésirables, désinformation, hypertrucages, propos outranciers ou haineux, comme dans le cas du génocide des Rohingya au Myanmar[200], en outrepassant les règles éthiques qui sont implicites pour des humains. De plus, la capacité de l'IA à fonder ses analyses sur des volumes considérables de données et de très nombreux critères donne l'illusion de son infaillibilité, comparativement aux méthodes d'analyse par des individus ou des organisations humaines, fondés sur des corpus d'information beaucoup plus réduits et un nombre de critères limité dont certains très subjectifs. Certains algorithmes d'IA ont même la faculté d'inspirer confiance aux humains en misant sur leur subjectivité ou leur émotivité, voire de les tromper « délibérément » pour parvenir à leurs fins[201],[202]. L'exploitation de telles possibilités pourrait permettre de constituer un ou plusieurs nouveaux systèmes de croyances et de règles morales, politiques et sociales, à l'instar des religions ou des régimes politiques, instituant comme vérité une certaine interprétation du réel et établissant une nouvelle forme d'ordre social. L'IA pourrait ainsi conduire à une certaine anarchie ou permettre l'instauration de régimes totalitaires dans les pays démocratiques (a contrario, des IA pourraient amplifier ou faire naître des idées subversives au sein des régimes autoritaires). Pour préserver les démocraties, il conviendrait donc que soient établies des régulations faisant intervenir des humains pour contenir les possibles dérives de l'IA[193][réf. incomplète].

L'emploi de l'IA pour noter les entreprises et citoyens, sur le modèle du crédit social, est déjà à l'œuvre en Chine. Ce procédé fait craindre une surveillance constante des personnes et des organisations, pouvant entrainer une forme d'asservissement des individus, ou des sanctions et récompenses excessives[193][réf. incomplète].

Éric Sadin récuse le terme d'« intelligence ». Selon lui, l'IA porte en elle un modèle de société utilitariste et rationaliste inspiré de manière extrêmement lacunaire par le fonctionnement du cerveau humain et visant à uniformiser les comportements en

temps réel et à tous moments. Le libéralisme économique, saisissant le profit qu'il peut en tirer, investit massivement dans l'IA afin d'exploiter l'inclination naturelle des humains à la facilité et les capacités extraordinaires de l'IA à expertiser la complexité du réel et à orienter nos décisions. Cela nous conduit insidieusement à nous fier aux réponses formulées par des algorithmes au détriment de critères d'analyse plus multisensoriels et de choix plus subjectifs, donc plus humains. Nous pourrions alors céder progressivement à l'IA notre pouvoir d'analyse, de jugement et de décision et consentir à nous soumettre plus ou moins consciemment à une « intelligence » supérieure, puis finalement renoncer même à penser. Notre addiction aux « assistants personnels » et aux capteurs physiologiques ainsi que leur présence continuelle à nos côtés pourraient les amener progressivement à s'exprimer, décider voire imaginer à notre place et en notre nom, d'abord avec, puis sans notre consentement[195].

Selon Laurent Alexandre, l'IA contribuera à accroître les inégalités en écartant les individus les « moins aptes » de l'activité économique et sociale, et le revenu universel constituerait à ses yeux un moyen d'asseoir la domination des « élites » sur les populations défavorisées.

En 2019, l'OCDE et le G20 adoptent une série de principes sur l'IA[203]. Le Partenariat mondial sur l'intelligence artificielle est lancé en juin 2020 pour promouvoir la conformité du développement de l'IA aux droits de l'homme et aux valeurs démocratiques. Il est hébergé par l'OCDE à Montréal et à Paris[204]. Une plateforme de communication, AI for Good (« l'IA pour le bien »), est créée pour faciliter les échanges et faire avancer les objectifs de développement durable de l'ONU grâce à l'IA[205].

En 2023, plus de 1 600 politiques publiques et stratégies sur l'IA sont recensées dans le monde[206]. Elles viennent en particulier de l'Union européenne, la Chine, les États-Unis et le Royaume-Uni. Après les avancées réglementaires de l'UE et de la Chine, la Maison-Blanche publie en octobre 2023 un décret sur l'IA « sûre, sécurisée et digne

de confiance ». En novembre 2023 a lieu un premier sommet en sécurité de l'IA au Royaume-Uni[206].

En Europe, les services numériques sont réglementés par le RGPD[207], le règlement sur les services numériques et la législation sur les marchés numériques. Pour l'intelligence artificielle en particulier, la législation sur l'intelligence artificielle (Artificial Intelligence Act, ou AI Act en anglais) définit quatre niveaux de risques pour les applications d'IA et met en avant des exigences de transparence, de protection des données, de sécurité et d'éthique[208].

En 2017, les Émirats arabes unis sont le premier pays au monde à se doter d'un ministre dédié à l'intelligence artificielle : Omar Sultan Al Olama[209].

Dans la seconde moitié des années 2010, des lanceurs d'alerte et des enquêtes révèlent que l'IA, encore émergente, a déjà été utilisée à des fins malveillantes pour faire basculer des processus électoraux. Le premier cas notable a été la plate-forme RIPON, secrètement créée par le Groupe SCL, à la demande de Steve Bannon et du milliardaire américain Robert Mercer. Cette plateforme, principalement au service de groupes politiques libertariens de droite, a été un outil de désinformation, de production et de diffusion de fake news à grande échelle[210],[211]. Ripon, impliquée dans le scandale Facebook-Cambridge Analytica/Aggregate IQ), joua un rôle important dans la manipulation d’un grand nombre d'électeurs, notamment pour faire élire Donald Trump lors de l'élection présidentielle américaine de 2016, pour faire advenir le Brexit[212], ainsi que pour orienter des dizaines d'élections dans le monde.

Face à ces dérives, les géants du secteur de l’IA ont réagi en créant le 28 septembre 2016 un « partenariat pour l'intelligence artificielle au bénéfice des citoyens et de la société »[213]. L'année suivante, Google DeepMind se dote d'une unité interne pour aborder les questions éthiques[214] et la conférence d'Asilomar rassemblant des personnalités influentes propose une charte[215] pour réglementer les développements

en IA[216].

Le 18 juillet 2018, 2 400 chercheurs, ingénieurs et personnalités du secteur de l'intelligence artificielle signent une lettre ouverte[217], s'engageant à « ne jamais participer ou soutenir le développement, la fabrication, le commerce ou l'usage d'armes létales autonomes ». La lettre précise que « La décision de prendre une vie humaine ne devrait jamais être déléguée à une machine ». Parmi les signataires se trouvent Elon Musk, les dirigeants de Google DeepMind Stuart Russell, Yoshua Bengio, et Toby Walsh[218].

Fin 2020, l'UNESCO rejoint (en tant qu'observateur, comme l'OCDE) le conseil et le comité directeur du Partenariat mondial sur l'intelligence artificielle, avec la possibilité de participer activement aux travaux de ces organismes[219].

La publication en février 2020 d'un Livre blanc sur l'intelligence artificielle[220], pose les bases du règlement sur l'intelligence artificielle de 2021 par la Commission européenne, qui vise à encadrer les risques et les problèmes éthiques de ces technologies[221]. Ce projet classe les risques en quatre catégories, dont la plus grave est qualifiée comme suit :

« Risque inacceptable : les systèmes d'IA considérés comme une menace évidente pour la sécurité, les moyens de subsistance et les droits des personnes seront interdits. Il s'agit notamment des systèmes ou applications d'IA qui manipulent le comportement humain pour priver les utilisateurs de leur libre arbitre (par exemple, des jouets utilisant une assistance vocale incitant des mineurs à avoir un comportement dangereux) et des systèmes qui permettent la notation sociale par les États[222]. »

En décembre 2022, le « premier forum mondial sur l'éthique de l'IA », réunion ministérielle internationale, est réuni à Prague, sous l'égide de l'Unesco[223].

La même année, l’Unesco, estimant que « l'autorégulation de l'industrie n'est manifestement pas suffisante pour éviter ces préjudices éthiques », a publié un

communiqué (adopté le 23 novembre 2021) demandant à tous les États de mettre en œuvre sa recommandation sur l'éthique de l'intelligence artificielle[224] afin de construire un cadre législatif et éthique pour l'IA.

L’objectif est de n'utiliser l’IA que lorsque les atouts qu'elle peut offrir sont bien identifiés, et qu’on peut éviter, limiter et réparer les risques qui lui sont associés (en particulier lors d'usages non pacifiques, malveillants et/ou aggravant les inégalités et des clivages). Ici, l'ONU, invite à ne pas utiliser l'IA quand elle met en péril la protection des données (tous les individus doivent pouvoir effacer et accéder aux enregistrements de leurs données personnelles, et les organismes de réglementation du monde entier doivent faire respecter ces dispositions). Cette recommandation vient aussi interdire la notation sociale et la surveillance de masse, contraires aux droits de l'homme et aux libertés fondamentales, et rejette l’idée d’accorder une personnalité juridique à l’IA « La Recommandation souligne que, lors de l'élaboration de cadres réglementaires, les États membres doivent tenir compte du fait que la responsabilité et l'obligation de rendre des comptes incombent toujours aux êtres humains en dernier ressort et que les technologies de l'IA ne devraient pas être dotées elles-mêmes d'une personnalité juridique ».

L’évaluation des IA prend en compte ses impacts éthiques sur les individus, sur la société et sur l'environnement. L’objectif étant à terme de créant une infrastructure juridique et technique ad hoc, ainsi qu’un responsable (indépendant) de l'éthique de l'IA pour surveiller l’utilisation et la création des IA qui devraient « privilégier les méthodes d'IA économes en données, en énergie et en ressources ». D’un point de vue écologique, les gouvernements sont invités, lors du cycle de vie du système d'IA à analyser son « empreinte carbone, sa consommation d'énergie et l'impact environnemental de l'extraction des matières premières pour soutenir la fabrication des technologies d'IA », tout en cherchant à diminuer l'impact environnemental du

numérique en investissant dans les technologies vertes. Ainsi, « si les systèmes d'IA ont un impact négatif disproportionné sur l'environnement, la Recommandation préconise de ne pas les utiliser »[225].

En 2023, l'UNESCO réitère son appel à la mise en œuvre rapide de sa Recommandation sur l'éthique de l'intelligence artificielle, adoptée à l'unanimité par les 193 États-membres. « C'est le défi de notre temps », et il est « urgent que tous transposent ce cadre sous la forme de stratégies et de réglementations nationales. Nous devons traduire les engagements en actes » a commenté Audrey Azoulay (directrice générale de l'Unesco)[226]. L'ONU appelle ainsi les États qui ne l'ont pas déjà fait à rejoindre les plus de 40 pays « de toutes les régions du monde » qui ont commencé à créer de tels garde-fous, pour notamment créer un outil législatif capable d'encadrer et de surveiller les IA, tout en veillant à la protection des données personnelles et sensibles, et en sensibilisant la population mondiale à un usage responsable de l'IA[226].

En 2023, lors de la 57e Journée mondiale de la paix, le pape François exprime ses préoccupations quant aux conséquences potentielles de l'IA sur la paix mondiale et demande à la communauté internationale de définir un traité international contraignant pour réglementer son développement et son utilisation. Il se montre particulièrement préoccupé par « la possibilité de mener des opérations militaires à travers des systèmes de contrôle à distance », citant notamment les systèmes d'armes létales autonomes[227].

En 2017, le Parlement européen a demandé à une commission d'étudier la possibilité qu'un robot doté d'une intelligence artificielle puisse être considéré comme une personne juridique[228],[229]. Advenant un dommage causé à un tiers par une intelligence artificielle, celle-ci pourrait être condamnée à réparer ce dommage. Il serait envisageable de conférer une personnalité électronique à tout robot prenant des décisions autonomes ou interagissant de manière indépendante avec des tiers, au

même titre qu'une personne morale et physique.

En 2023, Mark Coeckelbergh et Henrik Skaug Saetra, respectivement philosophe renommé de l'éthique des technologies et expert en sciences politiques, se penchent sur la question de l'intelligence artificielle (IA) et son potentiel rôle dans la lutte contre le changement climatique[230]. Ils plaident pour une intégration des IA dans les politiques démocratiques, soulignant qu'elles peuvent faciliter la délibération et la prise de décisions. Toutefois, ils avertissent également que l'IA pourrait devenir un outil si puissant qu'il pourrait entièrement remplacer le gouvernement, entraînant des problèmes sociaux majeurs.

Pour illustrer leur propos, Coeckelbergh et Saetra présentent deux cas extrêmes : une démocratie sans remplacement des humains (AI-augmented democracy) et une technocratie dirigée par l'IA (AI-driven technocracy)[230]. Ces deux cas extrêmes sont choisis pour démontrer les implications opposées de l'intégration de l'IA dans la gouvernance.

Dans une démocratie sans remplacement des humains, l'IA est présentée comme un simple outil visant à faciliter la prise de décision. Par exemple, elle peut être utilisée pour la traduction, la vérification des faits, ou la prise de notes. Ici, l'IA soutient le processus démocratique sans en prendre le contrôle.

En revanche, dans une démocratie avec remplacement, l'IA prend toutes les décisions, sans aucune intervention humaine. Cette approche, selon Saetra, pose cinq problèmes majeurs :

Ces points soulignent les conséquences sociales et éthiques de la prise de décision par l'IA en ce qui concerne les humains. Ainsi, Coeckelbergh et Saetra concluent qu'une démocratie sans remplacement des humains est plus adaptée, l'IA y étant présente uniquement comme soutien et non comme entité décisionnelle.

Les deux auteurs estiment cependant qu'aucune des deux propositions n'est

actuellement réalisable : les relations entre les humains et la technologie ne sont pas suffisamment évoluées pour permettre une utilisation éthique de l'IA. Ainsi, les décisions ne peuvent pas être prises uniquement par l'IA, car les erreurs sont encore trop fréquentes et les normes sociales et éthiques peu respectées[230].

Début 2023, l'apparition de ChatGPT suscite une grande curiosité, de l'enthousiasme, mais aussi des craintes sérieuses : « Devons-nous laisser les machines inonder nos canaux d'information de propagande et de mensonges ? (…) Devons-nous risquer de perdre le contrôle de notre civilisation ? Ces décisions ne doivent pas être déléguées à des leaders technologiques non élus » affirment Elon Musk, Steve Wozniak (cofondateur d'Apple) et des centaines d'experts. Le 29 mars 2023, ceux-ci, invoquant des « risques majeurs pour l'humanité », signent une pétition qui appelle le monde à un moratoire d'au moins six mois sur ces recherches, jusqu'à la mise en place de systèmes de sécurité, incluant : la création d'autorités réglementaires dédiées, des moyens pour efficacement surveiller des IA et des systèmes les utilisant, la mise à disposition de techniques permettant de mieux différencier le réel de l'artificiel, et la création d'institutions pouvant limiter les « perturbations économiques et politiques dramatiques (en particulier pour la démocratie) que l'IA provoquera »[226].

La cybernétique naissante des années 1940 revendiquait très clairement son caractère pluridisciplinaire et se nourrissait des contributions les plus diverses : neurophysiologie, psychologie, logique, sciences sociales, etc. Elle envisagea deux approches des systèmes, approches reprises par les sciences cognitives et de ce fait l'intelligence artificielle[réf. souhaitée] : une approche par la décomposition (du haut vers le bas, comme avec les systèmes experts) et une approche contraire par construction progressive du bas vers le haut, comme avec l'apprentissage automatique[231].

Ces deux approches se révèlent plutôt complémentaires que contradictoires : on est à l'aise pour décomposer rapidement ce que l'on connaît bien, et une approche

pragmatique à partir des seuls éléments que l'on connaît afin de se familiariser avec les concepts émergents est plus utile pour les domaines inconnus. Elles sont respectivement à la base des hypothèses de travail que constituent le cognitivisme et le connexionnisme, qui tentent aujourd'hui (2005)[Passage à actualiser] d'opérer progressivement leur fusion.

Le guide pratique de Linux sur l'intelligence artificielle v3.0[232], révisé le 15 décembre 2012, adopte pour la commodité du lecteur la taxinomie suivante :

Le cognitivisme considère que le vivant, tel un ordinateur (bien que par des procédés très différents), manipule essentiellement des symboles élémentaires. Dans son livre La société de l'esprit, Marvin Minsky, s'appuyant sur des observations du psychologue Jean Piaget, envisage le processus cognitif comme une compétition d'agents fournissant des réponses partielles et dont les avis sont arbitrés par d'autres agents. Il cite les exemples suivants de Piaget :

Au bout du compte, ces jeux d'enfants se révèlent essentiels à la formation de l'esprit, qui dégagent quelques règles pour arbitrer les différents éléments d'appréciation qu'il rencontre, par essais et erreurs.

Le connexionnisme, se référant aux processus auto-organisationnels, envisage la cognition comme le résultat d'une interaction globale des parties élémentaires d'un système. On ne peut nier que le chien dispose d'une sorte de connaissance des équations différentielles du mouvement, puisqu'il arrive à attraper un bâton au vol, ni qu'un chat ait aussi une sorte de connaissance de la loi de chute des corps, puisqu'il se comporte comme s'il savait à partir de quelle hauteur il ne doit plus essayer de sauter directement pour se diriger vers le sol. Cette faculté, qui évoque l'intuition des philosophes, se caractériserait par la prise en compte et la consolidation d'éléments perceptifs dont aucun pris isolément n'atteint le seuil de la conscience, ou en tout cas n'y déclenche d'interprétation particulière.

Trois concepts reviennent de façon récurrente dans la plupart des travaux :

Le concept d'intelligence artificielle forte fait référence à une machine capable non seulement de produire un comportement intelligent, notamment de modéliser des idées abstraites, mais aussi d'éprouver une impression d'une réelle conscience, de « vrais sentiments » (notion dont la définition n'est pas universelle), et « une compréhension de ses propres raisonnements »[233].

Contrairement à l'intelligence artificielle générale, l'intelligence artificielle forte fait donc le plus souvent intervenir des notions philosophiques de conscience qui font que les capacités de l'intelligence artificielle ne suffisent pas à dire si elle est « forte ». Cela dit, aucune définition de la conscience pour une IA ne fait consensus[234]. Les termes « intelligence artificielle forte » et « intelligence artificielle générale » sont parfois en pratique utilisés de manière interchangeable[61].

En partant du principe, étayé par les neurosciences[235], que la conscience a un support biologique et donc matériel, les scientifiques ne voient généralement pas d'obstacle théorique à la création d'une intelligence consciente sur un support matériel autre que biologique. Selon les tenants de l'IA forte, si à l'heure actuelle il n'y a pas d'ordinateurs ou d'algorithmes aussi intelligents que l'être humain, ce n'est pas un problème d'outil mais de conception. Il n'y aurait aucune limite fonctionnelle (un ordinateur est une machine de Turing universelle avec pour seules limites celles de la calculabilité), seulement des limites liées à l'aptitude humaine à concevoir les logiciels appropriés (programme, base de données…).

Les principales opinions soutenues pour répondre à la question d'une intelligence artificielle forte (c'est-à-dire douée d'une sorte de conscience) sont les suivantes :

Des auteurs comme Douglas Hofstadter (mais déjà avant lui Arthur C. Clarke ou Alan Turing ; voir le test de Turing) expriment par ailleurs un doute sur la possibilité de faire la différence entre une intelligence artificielle qui éprouverait réellement une

conscience, et une autre qui simulerait exactement ce comportement (voir Zombie (philosophie)). Après tout, nous ne pouvons même pas être certains que d'autres consciences que la nôtre, y compris chez des humains, éprouvent réellement quoi que ce soit, si ce n'est par une pétition de principe qui spécule que chaque humain se retrouve à l'identique chez tous les autres. On retrouve là le problème connu du solipsisme en philosophie.

Même si une intelligence artificielle forte n'était guère possible, une IA peut être de plus en plus perçue comme forte par une majorité d'individus parallèlement à l'arrivée des IA génératives, dont les LLM (de l'anglais large language model) comme ChatGPT ou Google Bard, et les outils de génération d'images comme Midjourney, DALL-E ou Stable Diffusion. En effet, le champ d'applications de ces outils est beaucoup plus large qu'auparavant : création, synthèse, traduction de textes, composition d'images, de vidéos à partir de prompts, textes descriptifs. Il devient ainsi de plus en plus difficile pour un être humain de distinguer des créations humaines de celles provenant d'une IA générative.

Emily Bender estime que les grands modèles de langage comme ChatGPT ne font que régurgiter plus ou moins aléatoirement des morceaux de texte venant des corpus ayant servi à leur entraînement, sans en comprendre le sens. Elle les appelle ainsi des « perroquets stochastiques »[240]. De même, Jean-Gabriel Ganascia considère que le contenu qu'ils produisent n'est pas original et que leur utilisation dans la rédaction d'articles de recherche constitue une forme de plagiat[241]. Ilya Sutskever considère au contraire que ces modèles, à force d'être entraînés à prédire le mot suivant, acquièrent une forme de « modèle du monde » et une représentation « compressée, abstraite et utilisable » des concepts[242].

La notion d'intelligence artificielle faible constitue une approche pragmatique d'ingénieur : chercher à construire des systèmes de plus en plus autonomes (pour

réduire le coût de leur supervision), des algorithmes capables de résoudre des problèmes d'une certaine classe, etc. Mais, cette fois, la machine simule l'intelligence, elle semble agir comme si elle était intelligente.

Les tenants de l'IA forte admettent que s'il y a bien dans ce cas simple simulation de comportements intelligents, il est aisé de le découvrir et qu'on ne peut donc généraliser. En effet, si on ne peut différencier expérimentalement deux comportements intelligents, celui d'une machine et celui d'un humain, comment peut-on prétendre que les deux choses ont des propriétés différentes ? Le terme même de « simulation de l'intelligence » est contesté et devrait, toujours selon eux, être remplacé par « reproduction de l'intelligence ».

Si le terme intelligence artificielle peut désigner un système capable de résoudre plusieurs problèmes de façon relativement autonome tout en ne faisant que simuler le principe d'intelligence, il peut aussi désigner des systèmes capables de résoudre uniquement un type de problème pour un jeu de données prédéfini[243]. On peut donner pour exemple un système entrainé à reconnaitre des chiffres écrits à la main, comme ceux utilisés par La Poste[244], qui malgré sa grande performance sur sa tâche, serait incapable de fonctionner sur un problème sortant de ce pour quoi il a été conçu. Ces intelligences artificielles, aussi nommées « intelligences artificielles étroites » (terme issu de l'anglais narrow AI), sont conçus pour effectuer une tâche précise, contrairement à une intelligence artificielle générale[245].

La définition du terme « intelligence artificielle » pose une question fondamentale : Qu'est-ce que l'intelligence[246] ?

L'intelligence peut se définir de manière générale comme un ensemble de capacités cognitives permettant de résoudre des problèmes ou de s'adapter à un environnement[247].

Le chercheur en IA Yann Le Cun avance que le noyau de l'intelligence est la faculté de

prédire. Les bases de la programmation des premiers systèmes experts supposent de « maîtriser parfaitement un problème et d'avoir une vue précise de toutes les solutions », afin d'en programmer précisément le comportement[246]. Mais les systèmes d'IA modernes à base d'apprentissage automatique sont entraînés à prédire la réponse attendue, ou à générer une solution correcte. Leurs capacités émergent ainsi progressivement, par essai-erreur, sans que le programmeur n'ait besoin de fournir une solution algorithmique, ou même de savoir comment résoudre le problème[231]. Dans tous les cas, l'efficacité de l'intelligence artificielle dépend de sa capacité à répondre aux objectifs donnés par les programmeurs et à tendre vers l'autonomie décisionnelle[248], ce qui présuppose, entre autres, une capacité de prédiction.

Le philosophe John Searle considère quant à lui que la faculté de comprendre est plus importante dans la définition de l'intelligence. Il essaie de démontrer la faiblesse des systèmes d'intelligence artificielle et les limites du test de Turing, par son expérience de la chambre chinoise, concluant : « on ne devrait pas dire d'une IA qu'elle comprend les informations qu'elle traite lorsqu'elle manipule des règles de syntaxe sans maîtriser la sémantique, c'est-à-dire sans reconnaître le sens des mots. La question de savoir si on peut parler d'une véritable intelligence reste donc ouverte »[246]. L'apprentissage automatique fonctionne cependant différemment de l'IA symbolique[231], qui était populaire à l'époque où Searle a conçu l'expérience de pensée de la chambre chinoise en 1980[249].

Une machine ayant une conscience et capable d'éprouver des sentiments — ou de faire comme si c'était le cas — est un grand thème classique de la science-fiction, notamment des romans d'Isaac Asimov sur les robots[250].

Ce sujet a toutefois été exploité très tôt, comme dans le récit des aventures de Pinocchio, publié en 1881, où une marionnette capable d'éprouver de l'amour pour son

créateur cherche à devenir un vrai petit garçon, ou dans L'Homme le plus doué du monde, une nouvelle de l'Américain Edward Page Mitchell où le cerveau d'un simple d'esprit est remplacé par un ordinateur inspiré des recherches de Charles Babbage[251]. Le roman Le Miroir flexible de Régis Messac propose quant à lui le principe d'une intelligence artificielle faible, mais évolutive, avec des automates inspirés de formes de vie simples, réagissant à certains stimuli tels que la lumière. Cette trame a fortement inspiré le film A.I. Intelligence artificielle réalisé par Steven Spielberg, sur la base d'idées de Stanley Kubrick, lui-même inspiré de Brian Aldiss[252]. L'œuvre de Dan Simmons, notamment le cycle d'Hypérion, évoque l'intelligence artificielle. Destination vide, de Frank Herbert, met en scène de manière fascinante l'émergence d'une intelligence artificielle forte. Plus récemment, l'écrivain français Christian Léourier a placé une intelligence artificielle au cœur de son roman court Helstrid (2018), dans lequel cette IA laisse un être humain mourir, contrevenant ainsi aux trois lois de la robotique instaurées par Isaac Asimov près de quatre-vingts ans plus tôt.

Les androïdes faisant preuve d'intelligence artificielle dans la fiction sont nombreux : le personnage de Data de la série télévisée Star Trek : The Next Generation est un être cybernétique doué d'intelligence, avec des capacités importantes d'apprentissage. Il est officier supérieur sur le vaisseau Enterprise et évolue aux côtés de ses coéquipiers humains qui l'inspirent dans sa quête d'humanité. Son pendant cinématographique est Bishop dans les films Aliens (1986) et Alien 3 (1992). Dans le manga Ghost in the Shell, une androïde s'éveille à la conscience. Dans la saga Terminator avec Arnold Schwarzenegger, le T-800 reprogrammé, conçu initialement pour tuer, semble dans la capacité d'éprouver des sentiments humains. Par ailleurs, les Terminators successifs sont envoyés dans le passé par Skynet, une intelligence artificielle qui a pris conscience d'elle-même, et du danger que représentent les humains envers

elle-même[réf. nécessaire].

Dans le dernier épisode de la saison 10 d'Inspecteur Derrick (diffusé en 1983), intitulé Un homme en trop, le Professeur Römer (joué par Erich Hallhuber) veut arrêter ses recherches sur l'intelligence artificielle, par peur du pouvoir sans conscience que cette nouvelle génération d'ordinateurs pourrait avoir sur les humains en finissant par les dominer et les tuer. Il abat son successeur pour l'empêcher de continuer son travail[253],[254],[c].

Les jeux, notamment les jeux de stratégie, ont marqué l'histoire de l'intelligence artificielle, même s'ils ne mesurent que des compétences particulières, telles que la capacité de la machine en matière de calcul de probabilités, de prise de décision mais aussi d'apprentissage.

Hans Berliner (1929-2017), docteur en science informatique à l'université Carnegie-Mellon et joueur d'échecs, fut l'un des pionniers de la programmation pour les ordinateurs de jeu. Ses travaux commencèrent par un programme capable de battre un humain professionnel au backgammon, puis, à partir des années 1960 et avec l'aide d'IBM, il fit des recherches pour créer un programme capable de rivaliser avec des grands maîtres du jeu d'échecs. Ses travaux contribuèrent quelques décennies plus tard à la réalisation du supercalculateur Deep Blue[256].

Outre la capacité des jeux à permettre de mesurer les performances de l'intelligence artificielle, que ce soit au travers d'un score ou d'un affrontement face à un humain, les jeux offrent un environnement propice à l'expérimentation pour les chercheurs, notamment dans le domaine de l'apprentissage par renforcement[257].

Dans le jeu Othello, sur un plateau de 8 cases sur 8, chaque joueur place tour à tour des pions de sa couleur (noir ou blanc). Le vainqueur est celui qui possède les pions de la couleur dominante.

L'une des premières intelligences artificielles pour l'Othello est IAGO, développée en

1976 par l'université Caltech de Pasadena (Californie), qui bat sans difficultés le champion japonais Fumio Fujita.

Le premier tournoi d'Othello opposant des hommes à des machines est organisé en 1980. Un an plus tard, un nouveau tournoi de programmes regroupe 20 systèmes[258]. C'est entre 1996 et 1997 que le nombre de programmes explose : Darwersi (1996-1999) par Olivier Arsac, Hannibal (1996) par Martin Piotte et Louis Geoffroy, Keyano (1997) par Mark Brockington, Logistello (1997) par Michael Buro, etc.

En 1968, le maître international anglais David Levy lança un défi à des spécialistes en intelligence artificielle, leur pariant qu'aucun programme informatique ne serait capable de le battre aux échecs dans les dix années à venir. Il remporta son pari, n'étant finalement battu par Deep Thought qu'en 1989[259].

En 1988, l'ordinateur HiTech de Hans Berliner est le premier programme à battre un grand maître du jeu d'échecs, Arnold Denker (74 ans) en match (3,5-1,5)[260],[d].

En 1997, le supercalculateur conçu par IBM, Deep Blue (surnommé Deeper Blue lors de ce match revanche), bat Garry Kasparov (3,5–2,5) et marque un tournant : pour la première fois, le meilleur joueur humain du jeu d'échecs est battu en match (et non lors d'une partie unique) par une machine.

En décembre 2017, une version généraliste d'AlphaGo Zero (le successeur du programme AlphaGo de DeepMind[e]) nommée AlphaZero, est développée pour jouer à n'importe quel jeu en connaissant seulement les règles, et en apprenant à jouer seul contre lui-même. Ce programme est ensuite entraîné pour le go, le shogi et les échecs. Après 9 heures d'entraînement, AlphaZero bat le programme d'échecs Stockfish (leader dans son domaine), avec un score de 28 victoires, 72 nulles et aucune défaite. Il faut cependant noter que la puissance de calcul disponible pour AlphaZero (4 TPU v2 pour jouer, soit une puissance de calcul de 720 Teraflops) était très supérieure à la puissance disponible de Stockfish pour ce match, ce dernier tournant sur un ordinateur

équipé de seulement 64 cœurs Intel[261]. AlphaZero a également battu (après apprentissage) le programme de shōgi Elmo (en)[262],[263].

En 2015, l'IA réalise des progrès significatifs dans la pratique du go, plus complexe à appréhender que les échecs (entre autres à cause du plus grand nombre de positions : 10170 au go, contre 1050 pour les échecs, et de parties plausibles : 10600 au go, contre 10120 pour les échecs)[264].

En octobre 2015, AlphaGo, un logiciel d'IA conçu par DeepMind, filiale de Google, bat pour la première fois Fan Hui, le triple champion européen de go[265] et ainsi relève ce qu'on considérait comme l'un des plus grands défis pour l'intelligence artificielle. Cette tendance se confirme en mars 2016 quand AlphaGo bat par trois fois consécutives le champion du monde de la discipline, Lee Sedol, dans un duel en cinq parties[266]. Lee Sedol a déclaré au terme de la seconde partie qu'il n'avait trouvé « aucune faiblesse » chez l'ordinateur et que sa défaite était « sans équivoque ».

En 2011, l'IA Watson conçue par IBM bat ses adversaires humains au jeu télévisé américain Jeopardy!. Dans ce jeu de questions/réponses, la compréhension du langage est essentielle pour la machine ; pour ce faire, Watson a pu s'appuyer sur une importante base de données interne lui fournissant des éléments de culture générale, et avait la capacité d'apprendre par lui-même, notamment de ses erreurs. Il disposait néanmoins d'un avantage, la capacité d'appuyer instantanément (et donc avant ses adversaires humains) sur le buzzer pour donner une réponse[264].

En 2007, Polaris est le premier programme informatique à gagner un tournoi de poker significatif face à des joueurs professionnels humains[267],[268].

En 2017, lors du tournoi de poker « Brains Vs. Artificial Intelligence : Upping the Ante » (« Cerveau contre Intelligence Artificielle : on monte la mise ») organisé dans un casino de Pennsylvanie, l'intelligence artificielle Libratus, développée par des chercheurs de l'université Carnegie-Mellon de Pittsburgh, est confrontée à des adversaires humains

dans le cadre d'une partie marathon étalée sur 20 jours[268]. Les joueurs humains opposés à Libratus, tous professionnels de poker, affrontent successivement la machine dans une partie en face à face (heads up (en)) selon les règles du « No Limit Texas hold'em » (no limit signifiant que les mises ne sont pas plafonnées), la version alors la plus courante du poker. Les parties sont retransmises en direct et durant huit heures par jour sur la plateforme Twitch[269].

Au terme de plus de 120 000 mains jouées, Libratus remporte tous ses duels face aux joueurs humains et accumule 1 766 250 dollars (virtuels). Le joueur humain ayant perdu le moins d'argent dans son duel face à la machine, Dong Kim, est tout de même en déficit de plus de 85 000 dollars. Dans leurs commentaires du jeu de leur adversaire, les joueurs humains admettent que celui-ci était à la fois déconcertant et terriblement efficace. En effet, Libratus « étudiait » chaque nuit, grâce aux ressources d'un supercalculateur situé à Pittsburgh, ses mains jouées durant la journée écoulée, utilisant les 15 millions d'heures-processeur de calculs du supercalculateur[269].

La victoire, nette et sans bavure, illustre les progrès accomplis dans le traitement par l'IA des « informations imparfaites », où la réflexion doit prendre en compte des données incomplètes ou dissimulées. Les estimations du nombre de possibilités d'une partie de poker sont en effet d'environ 10160 dans la variante no limit en face à face[269].

Auparavant, en 2015, le joueur professionnel Doug Polk (en) avait remporté la première édition de cet évènement contre une autre IA, baptisée Claudico (en)[269].

En mars 2022, un logiciel de bridge de la start-up française Nukkai parvient à gagner un tournoi et à expliquer aux perdants leurs erreurs[270].

Sur les autres projets Wikimedia :

Aspects juridiques

Notions générales

Notions techniques

Chercheurs en intelligence artificielle (espace anglophone)

Chercheurs en intelligence artificielle (espace francophone)

Laboratoires et entreprises en intelligence artificielle

: document utilisé comme source pour la rédaction de cet article.