Daniel Andler & Mehdi Khamassi pour le volume collectif TESaCo. version 3.1 21/10/2025 rev DA 30/10/2025

Chapitre de D. Andler & S. Ciranna (dir.), ***Technologies émergentes. Pour reprendre la main, la sagesse collective,***
Editions Odile Jacob, à paraître automne 2026

IA, quo vadis ? IA surpuissante et robots autonomes

Daniel Andler & Mehdi Khamassi

Introduction

L'intelligence artificielle est omniprésente dans l'actualité, et chacun a une idée plus ou moins précise de ce dont elle est capable. Ce n'était pas le cas il y a seulement peu d'années : l'IA a déboulé dans la conscience collective avec la soudaineté d'un orage d'été. Or, elle n'est pas née d'hier : ce qu'elle est aujourd'hui est l'aboutissement d'un cheminement théorique et technologique né il y sept décennies sinon plus[1]. Deux questions se posent : de quoi sont réellement capables les systèmes d'intelligence artificielle actuels, et jusqu'où iront-ils ? Pour pouvoir répondre à la première, il faut avoir une idée de la manière dont ils opèrent : qu'est-ce qui les fait marcher ? Pour pouvoir esquisser des réponses plausibles à la seconde, qui est l'objet du présent chapitre, il faut *en sus* comprendre ce que visent les concepteurs de ces systèmes.

Mais n'y a-t-il pas une première réponse toute simple à ces deux questions ? N'est-il pas clair que le but ultime poursuivi par l'industrie de l'IA est la construction de systèmes informatiques dotés d'intelligence, comme nous humains le sommes ? Que ce but n'est pas atteint aujourd'hui, mais que des progrès notables ont été faits, sous la forme de systèmes qui exécutent aussi bien sinon mieux que l'humain un certain nombre de tâches qui font appel à l'intelligence humaine. Et enfin, qu'à mesure que ces systèmes se perfectionnent et s'acquittent de tâches toujours plus difficiles et variées, l'IA se rapproche du but et finira tôt ou tard par l'atteindre ?

Tout cela serait en effet assez clair, mais à une condition qui n'est pas remplie : que nous sachions ce qu'est l'intelligence. Or ce dont nous disposons, c'est d'exemples, de manifestations d'intelligence[2]. Cela peut suffire tant qu'il ne s'agit que de constater qu'un ordinateur peut faire ceci ou cela, jouer aux échecs ou nous recommander un restaurant, organiser le ramassage des ordures à Naples, ou réserver les hôtels pour nos prochaines vacances, proposer une nouvelle molécule potentiellement efficace contre une certaine affection, etc. Mais si nous voulons savoir si certaines tâches que son intelligence permet à l'humain d'accomplir échappent et continueront d'échapper aux systèmes d'IA, nous devons disposer d'une définition de l'intelligence ; or nous n'en avons pas : nous avons une *idée* (ou plusieurs, selon le contexte) de ce qu'est l'intelligence, mais nos idées diffèrent et ne nous conduisent pas à un accord sur une définition ou du moins une caractérisation suffisante pour tracer un périmètre autour de ce qui est une manifestation ou une marque d'intelligence. Une chose est sûre néanmoins : il ne peut pas y avoir d'intelligence artificielle générale sans un certain degré d'*autonomie*, c'est-à-dire une capacité de se fixer ses propres buts, ses propres actions, et de les ordonner pour atteindre ces buts. On s'accorde généralement à y voir une condition nécessaire pour qu'un système d'IA soit capable non seulement de résoudre des problèmes dont la solution n'est pas encore connue, mais également d'identifier de nouveaux problèmes. Mais comment caractériser l'autonomie des systèmes d'IA actuels, et futurs ? Et comment évaluer les risques de cette autonomie ?

[1] La date de naissance officielle de l'IA est 1956, mais les idées fondamentales émergent de travaux des années 1930 et suivantes.

[2] Pour une analyse approfondie de l'intelligence humaine autant que de l'intelligence artificielle, leur comparaison, et l'évolution de leur compréhension par les scientifiques au fil des dernières décennies, voir Daniel Andler *(2024). Intelligence artificielle, intelligence humaine : la double énigme*, Gallimard.

Ces difficultés à caractériser intelligence et autonomie ne sont pas les seules : l'IA pose nombre de problèmes que nous n'aborderons pas tous. Notre fil conducteur est indiqué dans le titre : où nous mène l'intelligence artificielle d'aujourd'hui ? Nous évoquerons pour commencer ses origines (§1) et les deux grandes phases de son histoire (§2). Nous expliquerons comment l'IA a renoué avec l'ambition de créer une « intelligence générale artificielle », ou IGA, après y avoir longtemps renoncé (§3), en raison des progrès de sa branche la plus avancée, l'IA générative, dont ChatGPT est l'exemple le plus connu (§4). Que l'IGA ne soit pas encore atteinte, et ne le sera peut-être jamais, n'empêche pas d'entrevoir l'avènement de formes d'intelligence artificielle encore plus puissantes qu'aujourd'hui (§5). En particulier, le rapprochement de l'IA et de la robotique, qui compte parmi les orientations les plus fécondes de la recherche actuelle, pose à nouveaux frais la question de l'autonomie, longtemps occultée dans l'histoire de l'IA (§6). Cette question conduit directement à l'éthique : dans quelles conditions un système d'IA peut-il être construit et déployé en conformité avec l'éthique ? Ces conditions sont-elles remplies aujourd'hui, peuvent-elles l'être demain ? (§7) La question de l'équilibre entre les coûts avérés de cette technologie, matériels et moraux, et ses bénéfices actuels et espérés est posée.

1. Le projet de l'IA : deux manières de le comprendre

Dès l'origine, comme en témoignent maintes déclarations de ses créateurs[3], l'IA[4] visait à rien de moins que la création d'un système informatique doté d'intelligence au sens plein du terme. Mais ce projet s'entendait de deux manières différentes, quoique compatibles, sous-tendue chacune par des conceptions distinctes de l'intelligence humaine.

Selon la première, l'intelligence est le produit de certains systèmes de traitement de l'information. L'exemple type d'un tel système est familier : c'est l'ordinateur, une machine qui en vertu de sa constitution physique se prête à une interprétation sémantique de sa fonction. Pour illustrer l'idée, prenons une machine beaucoup plus simple que l'ordinateur, l'humble calculette. C'est un système physique agencé en sorte que si l'on appuie successivement sur les touches <3>, <+>, <5>, il affiche le symbole <8>. Si on interprète ces marques physiques comme nombre 3, addition, nombre 5, nombre 11, on constate que la calculette a calculé la somme de 3 et de 5 : la sémantique donne au processus physique le sens que nous attendons. L'ordinateur généralise ce schéma de plusieurs façons, la plus fondamentale étant qu'on peut interpréter les symboles qu'il manipule d'une infinité de façons, selon qu'on s'en sert comme d'un joueur d'échecs, d'un calendrier, d'un metteur en page - typographe, d'un piano : le lecteur en fait l'expérience quotidiennement.

L'intelligence quant à elle est le produit de certains de ces systèmes de traitement de l'information, dont le paradigme est le système cognitif humain, sous sa double description physique (le système

[3] McCarthy, John, Minsky, Marvin, Rochester, Nathaniel et Shannon, Claude, « Artificial Intelligence. Project at Dartmouth College », AI Magazine, vol. 27, n° 4, 2006 [1955]. À ces noms, il faut associer ceux de Herbert Simon et d'Allen Newell. Voir aussi, sur cette question et sur l'ensemble des questions évoquées dans ce chapitre, Daniel Andler, *Intelligence artificielle, intelligence humaine : la double énigme*, Paris, Gallimard, 2024.

[4] L'expression « intelligence artificielle » et son abréviation courante « IA » s'entendent de trois manières : elles peuvent désigner la propriété qu'il s'agit de conférer à certains systèmes informatiques ; elles peuvent désigner le programme technoscientifique visant à construire de tels systèmes ; enfin, ces systèmes eux-mêmes sont souvent appelés « une IA », « des IA ». Nous parlerons quant à nous de « systèmes d'intelligence artificielle » (SIA en abrégé), réservant « IA » au programme de recherche et aux institutions (entreprises, think tanks, instituts de recherche, comités d'action…) qui le soutiennent, et « intelligence artificielle » en toute lettre à la propriété ou capacité visée.

nerveux central ou, pour faire bref, le cerveau) et informationnelle (l'esprit, *mens* en latin, qui donne l'adjectif « mental », ou encore la cognition)[5].

Dans ce cadre, les termes « intelligence » et « cognition » sont presque synonymes. L'Intelligence est le produit du système cognitif, de la même manière que la digestion est produite par le système digestif, la reproduction par le système reproducteur et ainsi de suite. La différence est que la cognition renvoie au répertoire des fonctions cognitives (aussi diverses que la compréhension et la production du langage, la vision et autres modalités perceptives, la coordination motrice, le raisonnement…, dont chacune est composée de fonctions subordonnées), tandis que l'intelligence désigne plutôt l'effet émergent, global, du déploiement de ces fonctions. On aurait pu disposer de deux termes analogues, « cognition artificielle » et « intelligence artificielle ». L'usage ne les a pas consacrés, ce qui est dommage, en sorte que « intelligence artificielle » désigne tantôt un système qui a une certaine capacité, contrepartie *in silico* de la cognition, et tantôt ce que cette capacité produit, contrepartie de l'intelligence. Mais, toujours selon la première conception de l'IA, l'intelligence artificielle visée est de la même étoffe que l'intelligence humaine : elle est une version synthétique du même produit[6], la différence résidant dans le mode de production (par le cerveau, par un ordinateur ou par tout autre système artificiel)[7].

La seconde conception laisse de côté la nature de l'intelligence et ne porte que sur sa fonction principale, à savoir produire un comportement adapté, c'est-à-dire, toujours dans cette perspective, à accomplir des tâches ou encore à résoudre des problèmes. Selon cette seconde conception, la manière dont le système cognitif humain remplit sa fonction principale est sans pertinence directe pour le projet de l'IA, qui peut cependant s'inspirer de tel ou tel mécanisme naturel que suggèrent l'observation ou l'introspection. La première conception est beaucoup plus exigeante que la seconde : elle n'autorise aucun écart entre les opérations de l'intelligence humaine et celles de l'intelligence artificielle. Selon la seconde conception, on ne vise qu'à construire une intelligence de substitution, un *Ersatz* d'intelligence humaine, qui fait l'affaire en pratique et peut s'écarter de son modèle dans certains cas non prévus ou sans importance, voire le dépasser (par exemple au jeu d'échecs ou au jeu de go).

L'intelligence de synthèse visée par le programme ambitieux étant comprise comme indiscernable de l'intelligence humaine, elle serait par principe dotée de toutes ses capacités. Or ces capacités et la manière dont elles sont réalisées dans l'espèce humaine étaient loin d'être comprises lors de la création de l'IA (elles ne le sont que partiellement aujourd'hui). Les chercheurs étaient placés devant un double défi : *comprendre* les capacités chez l'humain et *doter* la machine de ces capacités. C'est pourquoi, dans la conception ambitieuse prônée par ses fondateurs, le projet d'intelligence artificielle et ce qui allait s'appeler, deux décennies plus tard, les sciences cognitives[8] constituaient les deux faces d'un même volet.

---

[5] Ces explications laisseront peut-être perplexe un lecteur non averti. Il pourra en faire abstraction et se reposer sur sa compréhension intuitive des processus mentaux d'un côté, de l'informatique de l'autre. S'il ne s'en contente pas, il pourra consulter le premier texte indiqué à la note 8 *infra,* ou d'autres textes introductifs figurant sur le site de Daniel Andler, https://andler.ens.psl.eu. Les fondements philosophiques sont dus non bien sûr à D. Andler mais aux autres auteurs cités dans la note 8.

[6] En chimie, en pharmacologie, en sciences des matériaux, en science des aliments, une substance est une forme synthétique d'une substance naturelle S si elle est produite par des procédés physico-chimiques inventés par l'homme mais est (en principe) indiscernable de S dans sa composition, et par conséquent dans ses propriétés. Un *Ersatz* de S (qui peut être naturel ou artificiel) conserve certaines propriétés utiles de S sans avoir la même composition.

[7] On peut y ajouter la cognition et l'intelligence animales, produites par des systèmes biologiques, qui peuvent être vues comme des variantes, différant d'une espèce à l'autre. Voir par exemple *The Cambridge Handbook of Animal Cognition* (2021), ou encore Frans De Waal *Sommes-nous trop « bêtes » pour comprendre l'intelligence des animaux ?,* Paris, Les Liens qui Libèrent, 2016. Mais on peut les laisser de côté dans le présent chapitre.

[8] Pour un rapide parcours historique, voir Daniel Andler, "« Les sciences cognitives : un tour d'horizon », in T. Collins, D. Andler et C. Tallon-Baudry (dir.), La cognition, du neurone à la culture, Paris, Gallimard, 2018, p. 15-70. Le projet des sciences cognitives a été formulé par Hilary Putnam (*Mind, Language and Reality, Philosophical Papers vol.2*, Cambridge University Press, 1975), Noam Chomsky (*Rules and Representations*, Blackwell, 1980 ; trad. fr. *Règles et représentations*, Flammarion, 1988) ; Jerry Fodor (*Representations*, MIT Press, 1981 et *The Modularity of Mind : An Essay on Faculty*

Comprendre en quoi consiste une capacité telle que la localisation d'une source sonore ou l'accès au lexique dans la production d'une phrase, et comment le cerveau réalise l'une et l'autre de ces fonctions, relève des sciences cognitives. Programmer un ordinateur en sorte qu'il en fasse autant relève de l'IA, mais dans la conception ambitieuse de l'IA les deux tâches sont par principe étroitement liées, sinon rigoureusement identiques. C'est pourquoi un Herbert Simon est célébré à la fois comme l'un des fondateurs des sciences cognitives (notamment couronné par le prix Nobel d'économie) et comme l'un des fondateurs de l'IA (notamment couronné par le prix Turing). Il est intéressant de noter également qu'un autre père fondateur de l'IA, John McCarthy, a exprimé au XXI^{e} siècle le regret de ne pas avoir suffisamment exploré les fonctions cognitives naturelles.

Le lien avec les sciences cognitives, quoique moins contraignant dans la seconde conception du projet que dans la première, reste néanmoins important, parce qu'elles indiquent au programme de l'IA certaines des pistes à suivre : la vision artificielle que l'on cherche à créer, même si elle peut fortement différer de la vision naturelle, a néanmoins de bonnes chances de mettre en jeu des composantes analogues à celles que les sciences cognitives mettent au jour. Et c'est ainsi qu'une vague inspiration des neurones biologiques a fini par donner naissance aux « réseaux de neurones », ingrédient clef des méthodes actuelles d'apprentissage statistique comme le *deep learning* (parfois appelé en français *apprentissage profond*).

2. Les hauts et les bas du projet : du paradigme symbolique au connexionnisme et au *deep learning*

Cette alliance s'est assez rapidement dénouée. La vision biface initiale fut d'abord portée par le programme dit *symbolique*, qui voit dans la cognition une sorte de logique générale portant sur des représentations, quelque chose comme des phrases écrites dans un langage mental intérieur analogue au langage naturel. L'intelligence humaine, dans ce cadre, consiste à acquérir de nouvelles connaissances par des calculs logiques opérant sur des données ou informations fournies par les sens, notamment lorsqu'ils traitent des stimuli linguistiques (oraux, écrits, gestuels…) et par la mémoire.

Le programme symbolique a occupé le devant de la science tant dans les sciences cognitives que dans l'IA, mais après une phase ascensionnelle s'est enlisé : les sciences cognitives progressaient moins rapidement que prévu, et l'IA s'en détournait, sans pour autant parvenir à des résultats convaincants. Ce jugement mérite d'être nuancé, et de fait le programme symbolique continue de jouer un certain rôle en IA, dont certains estiment même qu'il ira croissant.

Cependant un nouveau départ a été pris au milieu des années 1980, lorsque a été pleinement développée l'approche dite « connexionniste », connue aujourd'hui sous le nom de « deep learning » [9]. Pour ce courant, la cognition est une affaire de perception et non de logique : c'est en repérant des « formes », des « patterns » dans la masse de données auxquelles il est exposé qu'un système cognitif, qu'il soit naturel ou artificiel, construit une certaine représentation du monde qui lui permet d'y naviguer à son avantage, en complétant ces formes de la façon la plus probable.

Une deuxième différence importante sépare le nouveau programme de l'ancien. Le programme symbolique n'accordait aux neurosciences qu'un rôle restreint : pas plus qu'on ne peut comprendre ce

Psychology, MIT Press, 1983; trad. fr. par Abel Gerschenfeld, *La modularité de l'esprit : Essai sur la psychologie des facultés*, Minuit, 1986) ; Daniel Dennett (*Brainstorms. Philosophical Essays on Mind and Psychology,* MIT Press, 1978) et quelques autres.

[9] L'idée clé du connexionnisme était présente dès l'aube de l'IA, mais elle n'a cheminé que souterrainement pendant des décennies. L'ouvrage auquel est principalement dû le tournant connexionniste tant dans les sciences cognitives qu'en IA est Rumelhart, David E., McClelland, James L., The PDP Research Group, *Parallel Distributed Processing. The microstructure of cognition (vol. 1 & 2)*, MIT Press, 1986. On parle aussi aujourd'hui à son propos de « machine learning », une appellation impropre dans la mesure où cette notion est plus large, abritant d'autres approches que le deep learning.

que « fait » un ordinateur en disséquant les processus micro-électroniques qui se déroulent dans ses composants, on ne peut comprendre ce qui se déroule dans notre esprit lorsque nous jouons aux échecs ou que nous récitons un poème en examinant la manière dont les neurones échangent des signaux électriques et chimiques. Telle était du moins la prescription méthodologique du programme symbolique ; il ne contestait pas que le cerveau *produise* la pensée, mais niait que ce soit au niveau des neurones qu'il faut, du moins dans un premier temps, détecter les effets émergents, sous forme de pensées, de l'activité neuronale. L'étude du cerveau n'intervenait que pour garantir la faisabilité physique des processus mentaux postulés par l'analyse psychologique. Le connexionnisme revient sur ce choix : il postule que la structure fine du cerveau, son organisation en réseaux de neurones obéissant à certaines lois simples, fournit un modèle fécond tant pour l'étude de la cognition que pour la construction de systèmes d'IA.

On se serait attendu à ce que cette idée scelle l'alliance entre les deux disciplines. Il n'en a rien été. Les sciences cognitives sont restées, dans l'ensemble, indifférentes aux développements de l'IA, et celle-ci, particulièrement dans la branche de l'IA générative dont les modèles, tels ChatGPT, occupent la une des journaux, progresse à marche forcée sans rechercher dans les sciences cognitives des remèdes éventuels à ses difficultés conceptuelles ou pratiques. La course pour la suprématie dans laquelle sont engagés les grands acteurs de l'IA les rabat sur une conception purement techniciste de son devenir.

Cependant, si les liens entre les deux disciplines ont toujours été ténus, ils sont depuis l'origine entretenus par une minorité de chercheurs dont la voix est entendue et relayée depuis quelque temps[10]. Les sciences cognitives peuvent mettre au jour des capacités naturelles aujourd'hui absentes, même sous forme embryonnaire, dans les systèmes d'intelligence artificielle (nous emploierons le sigle SIA)[11], ce qui peut inciter les chercheurs en IA à tenter d'en doter les systèmes de demain ; ou inversement, mettre en garde la profession et les utilisateurs contre la tentation de croire que la marche de l'IA vers la réalisation de son objectif initial est irrésistible.

Inversement, les difficultés souvent inattendues rencontrées par l'IA attirent notre attention sur des aspects inaperçus de la cognition humaine. Enfin, certaines propriétés des modèles d'IA générative en font des terrains d'expérience pour les neurosciences — toutes questions que nous laisserons de côté.

Ces trajectoires (très simplifiées ici) nous rabattent sur la question fondamentale qui continue de commander d'une part l'évolution effective du domaine, d'autre part l'appréciation qu'en portent les acteurs et les observateurs et leurs attentes. La question est double : que vise la recherche en IA, et qu'est-ce qui constitue un progrès vers son objectif ?

3. Des replis féconds

Les premiers chercheurs n'ont pas tardé à mesurer l'immensité de leur tâche, et ont très tôt opéré une série de replis considérés par certains comme tactiques ou provisoires, d'autres comme stratégiques, d'autres enfin comme définitifs, impliquant une révision substantielle de l'objectif final. Les débats sur l'état présent de l'IA, sur son devenir proche et sur sa trajectoire ne se comprennent qu'à la lumière de l'histoire de ses replis, sur lesquels certains veulent désormais revenir.

Le premier repli opéré par l'IA consistait à mettre entre parenthèses tout ce qui, dans les phénomènes mentaux, relève d'une forme ou d'une autre d'intériorité, c'est-à-dire qui ne semble pas

[10] Parmi les plus connus, mentionnons Yann LeCun et Gary Marcus (qui ne s'accordent que sur ce point), mais aussi Joshua Tenenbaum et Joscha Bach.

[11] La locution "intelligence artificielle" et l'abréviation IA désignent tantôt le projet général dont il a été question jusqu'ici, tantôt la propriété que ce projet vise à conférer à certains machines, tantôt encore les systèmes qui sont les réalisations concrètes du projet au cours de son développement, à savoir des algorithmes ou des robots : ce sont ces derniers que nous regroupons sous le terme de système d'intelligence artificielle. On dit aussi souvent "une, des IA".

pouvoir être identifié ou compris sans référence à un état mental subjectif ou à un événement vécu par un être humain. C'est ainsi que la conscience elle-même, la compréhension, les émotions, l'agentivité (le fait de se percevoir, au présent, au passé ou au futur comme l'auteur d'une action qu'on accomplit) … ont été provisoirement exclues du périmètre de l'IA.

Or ces capacités semblent être des dimensions constitutives de la cognition humaine : on conçoit avec la plus grande difficulté une cognition privée de conscience, de compréhension, etc. Les mettre de côté, c'était effectuer un deuxième repli : l'IA renonçait à produire une intelligence de synthèse, puisqu'une telle intelligence possèderait, comme son modèle, les différentes facettes de l'intériorité.

Le troisième repli, le plus important en pratique pour les chercheurs et ingénieurs, consistait à diviser leur objectif en autant de tâches particulières qu'il en existe dans l'activité humaine, et pour chacune construire un programme, un algorithme permettant à un ordinateur de l'accomplir : une tâche, un programme. Le non-spécialiste ne voit peut-être pas immédiatement à quelle ambition l'IA renonçait ainsi. C'est celle à laquelle Herbert Simon et Alan Newell pensaient contribuer en proposant, très tôt dans l'histoire du domaine, leur célèbre concept du « General Problem Solver » (GPS), capable non de résoudre *un* problème particulier, mais *tout* problème ou du moins une grande variété de problèmes, conformément à l'exemple humain. L'idée, à titre d'un horizon lointain, n'avait jamais complètement disparu de l'imaginaire de certains chercheurs, notamment en robotique ; mais elle a été longtemps absente de l'ordre du jour de l'IA.

Ces replis n'ont pas constitué une capitulation. Ils ont permis aux chercheurs[12] de s'attaquer à des problèmes qui se sont révélés solubles, au moins en partie. Tandis qu'une partie de la profession se perdait dans des tentatives infructueuses pour se rapprocher de l'objectif d'une intelligence semblable à l'intelligence humaine, la plupart des chercheurs constituaient un savoir-faire et construisaient des systèmes qui marchent et dont certains sont incontestablement utiles. Il importe de le préciser et d'en donner quelques exemples, avant de développer l'objet du chapitre, à savoir le sens et la portée des tentatives en cours pour réaliser l'ambition des pionniers, une intelligence artificielle à l'égal de l'intelligence humaine.

On entend dire que l'IA est désormais partout, qu'elle pénètre tous les secteurs d'activité, toutes les professions et qu'elle est sur le point de les transformer totalement, si ce n'est déjà fait, et même d'en expulser les humains. Nous sommes réservés sur ce diagnostic, mais laisserons la question de côté. Indiquons seulement que l'exemple donné le plus fréquemment, celui de la radiologie ou plus généralement de l'imagerie médicale, autant l'aide apportée par l'IA est incontestable, autant la prédiction d'une disparition des radiologues semble erronée. L'IA fournit dans ce domaine, comme dans l'ensemble de la médecine et nombre d'autres professions, des outils précieux aux professionnels ; elle ne fournit pas des professionnels.

Pour illustrer de manière éclatante la contribution de l'IA à certains secteurs, nous choisissons la recherche scientifique. La percée scientifique la plus significative concerne la prédiction du repliement tri-dimensionnel des protéines par l'algorithme Alphafold de la société Deepmind, qui a valu un récent prix Nobel à certains de ses auteurs. AlphaFold3 étend cette capacité de prédiction à d'autres structures moléculaires. Des méthodes analogues permettent de concevoir d'énormes quantités de nouvelles molécules potentiellement utiles en pharmacologie ou pour de nouveaux matériaux. La mise au point de vaccins à ARN a été facilitée par l'IA.

Les exemples se multiplient dans la plupart des disciplines des sciences dites dures. C'est ainsi que des techniques d'IA ont récemment permis d'améliorer l'image de M87*, un trou noir supermassif. L'IA semble avoir contribué à la modélisation de la structure de futures centrales de fusion nucléaire, qui permettraient de répondre aux besoins énergétiques croissants des systèmes d'IA sans contribuer

[12] Nous ne distinguons pas ici les chercheurs des ingénieurs. La différence existe, mais elle est de degré, et peut être vue comme un continuum entre deux pôles.

significativement au réchauffement climatique. L'IA permet d'améliorer l'étude de l'évolution du climat en améliorant les modèles et en réduisant leur coût énergétique ; elle fournit aussi des prédictions météorologiques très précises et très rapides avec beaucoup moins de calculs que les modèles classiques.

Les sciences empiriques ne sont pas les seules bénéficiaires : les applications en mathématiques se multiplient, depuis deux exemples spectaculaires — une découverte en théorie des nœuds et une piste pour la résolution d'une conjecture en théorie des représentations suggérée indirectement par l'IA[13]. Les interactions entre mathématiciens et systèmes d'IA varient considérablement. Dans le schéma le plus courant aujourd'hui, un système d'IA part à la recherche de connexions dans un espace d'objets mathématiques circonscrit par un mathématicien, qui reprend la main pour explorer le potentiel d'une connexion suggérée par le système. Un stade plus avancé donne au système un rôle dans la délimitation de l'espace à explorer. Enfin, on essaie parfois de confier à un système d'IA un cycle complet (*closed-loop AI*), le chercheur définissant les termes du problème et supervisant les étapes du processus.

Il s'agit là de quelques résultats épars, dont ne se dégagent pas encore des principes méthodologiques, ni une manière consensuelle de ventiler le mérite d'un progrès entre l'humain et la machine. Nous n'en sommes qu'à l'aube des transformations que connaîtra la recherche scientifique sous l'effet de l'IA. Mais il est peu douteux que l'IA apportera à différents niveaux une aide décisive à la plupart des sciences de la nature, à condition de compléter sans remplacer l'expertise et le savoir-faire humains. Cultiver des scientifiques épistémiquement renforcés par l'IA nous semble en effet clef pour permettre à ces humains d'identifier les problèmes de demain à soumettre à la résolution par des systèmes d'IA. Le cas des sciences humaines et sociales soulève des difficultés particulières qui pourraient limiter ou ralentir les contributions de l'IA à ces disciplines.

4. L'intelligence générale artificielle en vue ? L'heure des LLM

L'IA ne doit sa célébrité et sa fortune qu'en partie aux réalisations dont il vient d'être question. Son statut de star culturelle et financière est dû au choix qu'elle a fait de revenir sur le dernier repli que la profession avait consenti à ses débuts : résoudre des problèmes, automatiser leur résolution, un par un et non par l'opération d'une intelligence « à tout faire ». À la toute fin du XX^{e} siècle, quelques chercheurs en IA lui ont assigné comme objectif de créer ce qu'ils ont appelé une intelligence générale artificielle (IGA, en anglais *artificial general intelligence*, sigle AGI), qu'ils opposaient à l'ensemble des réalisations existantes, relevant de ce qu'ils appelaient l'intelligence artificielle *étroite*[14]. Selon eux, Simon et Newell avaient de l'objectif la bonne conception, mais n'avaient pas trouvé la voie qui y mènerait. Les progrès accomplis en quatre décennies, en matière de ressources calculatoires, de progrès théoriques, de compréhension des mécanismes cérébraux et fonctionnels chez l'humain, permettaient désormais à la profession de prendre pour cible l'IGA — voire le lui imposaient. L'IGA serait donc *versatile* — elle s'acquitterait de toute tâche (humainement faisable) qui se présente à elle, contrairement aux systèmes spécialisés d'intelligence artificielle étroite. Elle passerait également d'une tâche à l'autre de manière *fluide*, comme le fait, ou semble le faire l'intelligence humaine : nous ne semblons pas « changer de

[13] Davies, Alex et al., « Advancing Mathematics by Guiding Human Intuition with AI », Nature, vol. 600, 2 décembre 2021

[14] Le premier manifeste de l'IGA est sans doute le post de Ben Goertzel, Artificial General Intelligence: Now Is the Time, publié le 9 avril 2007. Pour une présentation très complète, v. Ben Goertzel & Cassio Pennachin, dir., *Artificial General Intelligence,* Springer, 2007. Le terme lui-même, IGA, est considéré comme provenant de Gubrud, M., *Nanotechnology and International Security*, Fifth Foresight Conference on Molecular Nanotechnology, 1997, qui le définit comme "AI systems that rival or surpass the human brain in complexity and speed, that can acquire, manipulate and reason with general knowledge, and that are usable in essentially any phase of industrial or military operations where a human intelligence would otherwise be needed."

logiciel » lorsque nous entendons quelqu'un nous poser une question, que nous y répondons en faisant une multiplication dans la tête tout en cherchant une place de parking.

Cette IGA versatile et fluide se rapprochait donc (dans l'image qu'en proposaient ses concepteurs) de l'intelligence humaine. Dès lors était remise à l'ordre du jour la question, restée inactuelle pendant des décennies, des rapports entre l'intelligence artificielle désormais visée et l'intelligence humaine. Mais à ses débuts l'IGA se présentait, aux yeux de la plupart des chercheurs et des observateurs, soit comme une utopie défendue par un petit groupe marginal, soit comme une opération marketing pour tenter de remettre au goût du jour un vieil objectif, soit plus modestement comme une incitation à faire preuve d'audace et à se mettre en quête d'idées permettant de dépasser les limites de l'IA du moment. Aussi ne semblait-il pas urgent de scruter de trop près le concept d'IGA ni de s'interroger sur celles des propriétés de l'intelligence humaine que l'IGA devrait posséder sous peine d'incohérence : l'autonomie ? la conscience ? la moralité ?

Ces questions n'ont commencé à être examinées qu'à partir du moment où l'IGA est apparue comme autre chose qu'une utopie ou un horizon lointain, comme un objectif à portée de main. Ce retournement de l'opinion professionnelle s'est produit en deux temps. Les avancées rendues possibles par le *deep learning* à partir des années 2010 ont d'abord réhabilité le rêve des pionniers : les échecs, les impasses, les promesses jamais tenues, les (quasi) démonstrations d'impossibilité, tout cela était du passé. Le second temps est à l'image d'un boxeur, qui ayant pris l'ascendant sur son adversaire, le met soudain au tapis. Le coup infligé au scepticisme à l'égard de l'IGA est l'arrivée sur le marché, en novembre 2022, de ChatGPT. Les bases technologiques de ce système existaient depuis cinq ans, et ses fondements datent de plusieurs décennies. Il s'inscrit dans la continuité d'une série de modèles (les premiers « large language models » (LLM)), et tous relèvent de l'approche générale du deep learning. Pourtant, ChatGPT et les désormais innombrables systèmes de ce qu'on appelle, on verra pourquoi, l'IA générative ont effectivement ouvert une ère nouvelle dans l'histoire de l'IA.

Le principe général de construction des LLM est désormais assez bien connu. Un LLM est d'abord exposé à une quantité astronomique de textes, collectés notamment sur internet. Il en forme une représentation statistique au cours de son apprentissage, qui consiste à acquérir progressivement la capacité de prédire comment compléter le début d'une phrase du texte. Cette première phase est suivie par un « dressage » par renforcement, au sens de la psychologie pavlovienne, piloté par des opérateurs humains. Ces explications, on s'en doute, sont très insuffisantes : le contexte informatique et mathématique est essentiel et peu accessible au non-spécialiste. Mais on peut et on doit les compléter, très partiellement, de trois manières. Premièrement, le succès des LLM dépend de l'ordre de grandeur astronomique de leur base d'apprentissage, et n'est rendu possible que par ce qu'est devenu internet, colossale base de « données » (en un sens évidemment très étendu de la notion) comme l'humanité n'en a jamais connue. Deuxièmement, un LLM n'est pas un simple stock de textes mémorisés : que ses représentations internes lui permettent de trouver des informations ponctuelles voire de reconstituer des textes entiers présents dans le web ne semble pas surprenant, du moins vu de loin, mais est en tout cas très loin de suffire à expliquer ses capacités. Enfin, personne aujourd'hui n'est capable de les expliquer complètement : ceux mêmes qui en ont conçu le principe et ceux qui les construisent ne savent pas comment leurs modèles parviennent à se comporter avec une telle apparence d'intelligence et de répondre avec tant de pertinence, non infailliblement mais régulièrement, aux attentes de leurs utilisateurs. C'est une découverte empirique qu'ont faite les chercheurs, et c'est empiriquement qu'ils mènent leur quête de systèmes toujours meilleurs en un sens ou un autre.

Ce n'est pas que ces modèles fassent indubitablement preuve d'intelligence générale. Ceux qui se risquent à l'affirmer, au moment où nous écrivons, sont l'exception, comme l'indique obliquement le titre d'un article de fin 2023 cosigné par deux éminents spécialistes, Blaise Agüera y Arcas et Peter Norvig :

« Artificial General Intelligence Is Already Here »[15]. Mais ce dont sont capables ChatGPT et consorts évoque incontestablement des prémices d'une véritable intelligence. En particulier, ils semblent doués de versatilité et de fluidité, aux sens employés ci-dessus : comme chacun ou presque le sait désormais, pour en avoir fait l'expérience directement ou par personne interposée, ils peuvent répondre à une variété infinie de questions sur n'importe quel sujet. De plus et surtout ils peuvent soutenir un dialogue souvent cohérent avec un interlocuteur humain et proposer toutes sortes de textes allant d'un poème sur le printemps dans le style de Dante ou d'Alphonse Allais à un plan de cours sur la mécanique quantique pour étudiants de première année. C'est en raison de cette capacité de générer à la demande des textes en tout ou partie inédits (qui ne sont pas « encodés » au cours de la mise au point par apprentissage des modèles) qu'ils sont désignés globalement comme relevant de l'« IA générative ». Ils sont également capables de tenir compte de l'historique des échanges qu'ils ont eu avec un utilisateur donné, ce qui lui donne le sentiment d'être pris par le système comme une personne particulière et non un individu quelconque. De plus, à côté des LLM de la première génération, qui sont des machines exclusivement linguistiques, existent, c'est désormais également bien connu, des systèmes acceptant en entrée et produisant en sortie des images fixes ou mobiles, des sons, des paroles, seules ou en combinaison. Les modèles les plus récents (qui ne le seront déjà plus lorsque le lecteur lira ces lignes) peuvent même charger des « agents » artificiels de certaines tâches subsidiaires et combiner leurs réponses pour proposer une solution au problème posé. Bref, on ne peut apparemment que souscrire au jugement des auteurs cités à l'instant : « Les derniers modèles font preuve de compétence dans pratiquement toute tâche informationnelle à la portée des humains qui peut être formulée et satisfaite en employant le langage naturel, la qualité de la réponse pouvant être quantifiée. »

Cette affirmation, sous les dehors d'une simple profession de foi optimiste comme l'IA en suscite en quantité depuis toujours, est très soigneusement calibrée[16] pour prévenir certaines des objections les plus courantes que soulèvent les présentations avantageuses des systèmes génératifs. Avant de présenter les principaux arguments échangés, il faut rappeler le contexte très particulier de la discussion. En deux décennies, l'IA est devenue l'une des industries les plus puissantes que le monde a connues. Que l'on compte en chiffre d'affaires ou en investissement à l'échelle mondiale, les ordres de grandeurs annuels se comptent en centaines de milliards de dollars, et dépasseront le milliard de milliard d'ici quelques années selon certaines projections. C'est dire la puissance des moyens qu'elle peut mobiliser pour convaincre le public et les décideurs publics et privés de la qualité de ses produits et de ses réussites à venir. Cette situation rappelle à certains égards celle des industries du tabac, de l'automobile ou des énergies fossiles, qui recrutaient, et recrutent encore, des experts pour plaider leur cause, minimisant les dégâts et les risques, exagérant les avantages et les progrès à attendre. Le cas de l'IA s'en distingue néanmoins partiellement : sa position est plus jeune et plus précaire, et la discussion est beaucoup plus serrée, les incertitudes beaucoup plus profondes. En sorte que la conversation sur les performances présentes et les chances de succès de l'IA, quoique académique au meilleur sens du terme et plus complexe que sur celles d'autres industries puissantes, n'est pas plus qu'elles seulement académique : elle est soumise à une pression extrêmement forte. L'article cité à l'instant est caractéristique : il est d'excellente tenue, ses auteurs sont des experts de haut niveau, il n'en est pas moins un plaidoyer *pro domo*[17].

Examinons de plus près les arguments échangés.

---

[15] « L'intelligence générale artificielle est déjà parmi nous », *Noema,* revue du Berggruen Institute. Le titre de l'article implique clairement que ce qu'il affirme prend le contrepied d'une certaine *doxa.*

[16] Dans l'original : "Frontier language models can perform competently at pretty much any information task that can be done by humans, can be posed and answered using natural language, and has quantifiable performance."

[17] Blaise Agüera y Arcas est vice-président de Google Research ; Peter Norvig, jusqu'à récemment directeur de la recherche chez Google, est aujourd'hui membre du Stanford Institute for Human-Centered AI et co-auteur du manuel le plus connu de la discipline : *Artificial Intelligence : A Modern Approach*, Prentice Hall, 1995 ; 4e éd. 2000 ; trad. fr. *Intelligence artificielle. Une approche moderne,* Pearson, 4e éd. 2021.

(a) *Faiblesses des LLM*

La première concerne les défaillances des LLM. Elles ne sont pas exceptionnelles, ce qui limite la confiance qu'on peut leur faire. Mais elles sont également étranges et souvent inquiétantes, en ce sens que, de la part d'un humain, elles le discréditeraient immédiatement. Les « hallucinations » sont les plus connues : il s'agit d'inventions pures et simples sans aucune base factuelle mais présentées comme des faits. De manière générale, il est patent qu'un LLM tel que ChatGPT (pour prendre le système le plus utilisé par le grand public) n'a pas de rapports directs avec les faits ou comme on le dit parfois avec la vérité : il reflète et recombine les échos qui parviennent à internet par l'intermédiaire de ce qui est dit et écrit sur le monde réel par des individus plus ou moins lucides, bien informés et bien intentionnés[18]. Ce qui accroît la fréquence des hallucinations est la tendance, programmée dans les systèmes actuels, à donner satisfaction à leurs utilisateurs[19], en sorte que plutôt qu'admettre qu'ils n'ont pas de réponse, ils en inventeront une à peu près plausible. C'est ainsi que ChatGPT peut attribuer à un auteur connu un livre que celui-ci n'a jamais écrit mais qu'il aurait pu écrire, au vu de son titre, en fournissant l'éditeur, le nombre de pages et la date de publication !

ChatGPT ne fait pas qu'halluciner[20]. Il se trompe massivement dans des problèmes élémentaires, qu'ils soient de nature formelle ou relèvent du sens commun. Un exemple : S'il faut 5 heures pour faire sécher 5 chemises au soleil, combien en faut-il pour en faire sécher 30 ? Réponse de ChatGPT : 30 ; autre exemple, cité par le philosophe Luciano Floridi[21] : Quel est le nom de la fille unique de la mère de Laura ? Résumé de la réponse : « Je ne sais pas. » Un LLM tolère des contradictions patentes, sans chercher à les lever. Il reconnaît volontiers qu'il s'est trompé, si on le lui dit, s'excuse et peut tout autant donner une meilleure réponse ou commettre une autre erreur.

Pour nous inciter à accepter ces faiblesses, les défenseurs des LLM nous invitent à considérer ces systèmes comme seulement « compétents » à l'image des humains dont la compétence dans tel ou tel domaine n'implique pas qu'ils soient, comme l'expert, quasiment infaillibles. L'argument est de bonne guerre, mais les erreurs des LLM ne sont pas des erreurs que feraient à leur place un humain seulement compétent — nous reviendrons sur ces comparaisons.

Le plus important, mais le moins remarqué dans ces débats est que les systèmes actuels sont loin de pouvoir traiter « pratiquement » toute tâche informationnelle : une tâche trop éloignée d'un ensemble suffisamment nombreux d'exemples présents dans sa base d'apprentissage échappe à sa compétence[22]. Il peut s'agir d'une tâche particulière qui ne se rattache à aucun type de tâche répertorié, ou encore d'un représentant d'un type trop peu nombreux ou trop peu représenté dans internet pour permettre au système de construire une distribution de probabilités fiable. Or ces cas sont loin d'être rares.

Au total, on estime à 20% l'ordre de grandeur du taux d'erreur moyen des LLMs, sachant qu'il diffère énormément selon le modèle, le domaine de déploiement et le type de tâche. Sachant aussi que beaucoup dépend du savoir-faire de l'utilisateur[23]. Mais ce taux masque des bizarreries qui font réfléchir à

[18] Internet est désormais alimenté par des produits de systèmes d'IA, ce qui aggrave considérablement la situation, non seulement quant à la fiabilité des résultats de systèmes d'IA, mais celle de la Toile dans sa totalité, polluée par une masse grandissante de données fantaisistes.

[19] Cette tendance se manifeste également par les flatteries que les systèmes prodiguent à leur utilisateur, comportement que l'anglais désigne comme « sycophancy » (le sens de « sycophante » en français est différent).

[20] Il existe d'ailleurs des répertoires d'erreurs commises par ChatGPT et autres LLMs, y compris les plus récents. OpenAI se fera un plaisir de fournir au lecteur un florilège de sites et d'exemples !

[21] Luciano Floridi, Philo & Tech ; with ChatGPT 3, 2023.

[22] Cet argument a été développé en particulier par François Chollet, « On the Measure of Intelligence », 2019, arXiv :1911.01547.

[23] Ce qu'on appelle dans le jargon la « promptologie », l'art de poser les bonnes questions, de fournir les bonnes informations aux LLM et de les orienter dans la bonne direction.

la manière dont on peut évaluer les performances de ces systèmes, et incitent à la prudence quand on les utilise. C'est ainsi que les LLMs ne « savent » pas ce qu'ils ne savent pas, et qu'en ce cas, au lieu de le reconnaître, ils peuvent avancer avec assurance une réponse inventée de toutes pièces.

(b) *Progrès des LLM et à quel étalon le rapporter*

Au vu de ces limitations, un certain consensus s'est fait sur l'idée qu'on ne peut actuellement confier à un LLM sans supervision humaine une tâche comportant un enjeu important. Mais un nouveau désaccord surgit aussitôt : cette situation est-elle permanente, ou n'est-elle que provisoire compte tenu des progrès constants observés depuis les débuts de l'IA générative ? Cette question, en cache une autre. La question d'origine semble être de savoir si le progrès en question est voué à se poursuivre jusqu'au succès complet. La seconde, cachée derrière la première, est de savoir quelle distance diminue en raison de ce progrès : vers quoi va l'IA générative ? Le progrès est mesuré actuellement par les performances sur des « benchmarks »[24]) : soit des épreuves d'entrée ou d'admission à telle ou telle formation ou diplôme professionnel, soit des tests d'aptitude intellectuelle spécialisée ou générale, déjà utilisés pour les humains ou fabriqués comme défis pour les systèmes d'IA. La vaste majorité de la profession considère qu'il y a progrès lorsque les derniers systèmes obtiennent un pourcentage plus élevé de bonnes réponses que leurs prédécesseurs, sur les mêmes benchmarks ou sur des benchmarks plus exigeants ou plus variés.

Prenons les deux questions dans l'ordre inverse et demandons-nous d'abord quelle est la signification des progrès accomplis par les systèmes d'IA. Ces progrès sont incontestables, même s'ils ne sont pas uniformes : tel système qui vient de sortir évitera un certain type d'erreur mais en fera de nouvelles, ou en fera davantage que les systèmes qu'il prétend supplanter.

Nous avons une longue pratique de l'évaluation de l'intelligence ou de telle ou telle capacité intellectuelle chez l'humain. Nous reposons soit sur notre intuition, informée par notre expérience, soit sur des tests, épreuves, examens de toute sorte. Et notre jugement est à la fois comparatif et prédictif : si l'individu A, contrairement à l'individu B, s'est montré capable de résoudre un problème ou d'accomplir une tâche, il est réputé avoir fait preuve *en la circonstance* de plus d'intelligence que B. Si le problème ou la tâche sont bien choisis, nous prédisons que A s'en tirera, contrairement à B, en *toute circonstance* analogue, c'est-à-dire face à des problèmes ou tâches du même ordre. Nous pouvons ensuite comparer les performances de A et de B dans une variété plus ou moins large de tâches, et en cas de supériorité intégrale ou en moyenne prédire que A se tirera mieux que B, selon le cas, dans un certain répertoire de tâches, ou dans quasiment toute tâche.

Cette méthodologie est pratiquée dans une infinité de situations et d'une infinité de manières. Elle est régulièrement soumise à la critique lorsque le caractère prédictif des tâches prises comme test est mis en doute : qui réussit le concours d'entrée en médecine sera-t-il, en probabilité, un bon médecin ? Qui a passé le barreau un bon avocat ? Questions qui conduisent de temps à autre à rien d'autre qu'une modification des épreuves. La méthodologie elle-même résiste, car elle est ancrée dans une « théorie naïve » (*folk theory*) de l'intelligence basée sur l'expérience commune, qui résiste faute de concurrence et en l'absence d'une réfutation massive par la psychologie scientifique. Si floue et faillible qu'elle soit, comme toute théorie naïve, cette théorie exclut certains profils d'intelligence : de manière générale, un être humain ne peut, selon elle, réussir systématiquement dans certaines tâches réputées difficiles et échouer systématiquement dans des tâches très faciles, sauf bien sûr dans des circonstances particulières (inattention, stress, contexte-piège…).

Or les systèmes avancés de l'IA présentent précisément ce genre de profil : ils résolvent des problèmes jugés très difficiles (par exemple en mathématiques), et font mieux que la plupart des candidats à des certifications professionnelles de haut niveau ; mais ils commettent également avec

[24] Dans le contexte de l'IA on conserve le terme anglais ; dans d'autres contextes on parle de référence ou d'étalon ou d'échelle de comparaison.

régularité des bourdes qui, de la part d'un humain, seraient comprises comme les manifestations d'un handicap cognitif : ils semblent alors dépourvus de simple « bon sens ». Ce qui a été désigné très tôt dans l'histoire de l'IA comme « problème du sens commun » reste en attente d'une solution[25].

Quelle est alors la signification des performances impressionnantes qu'obtiennent sur les benchmarks les plus récents LLM ? Nous sommes tentés de penser qu'un système qui s'acquitte de tâches sensiblement plus difficiles que celles qu'un autre système peut réaliser lui sera également supérieur dans des tâches nouvelles, sur lesquelles ils n'ont pas été testés. Ce qui nous y pousse c'est, ici encore, notre longue expérience des performances humaines. Mais cela vaut-il pour des LLM ? Un nouveau système qui fait mieux que les précédents sur ces benchmarks, tout en continuant de commettre des bourdes dans des cas très simples, ou encore sur les tâches complexes qui se présentent dans la vie quotidienne, constitue-t-il un progrès ? De manière générale, les créateurs des LLMs jouent à une sorte de jeu qui consiste à conférer à leurs modèles des capacités qui miment, sans les reproduire, certaines capacités avancées de l'intelligence humaine — par exemple la capacité de raisonner, ou la capacité d'éprouver ou d'être conscient. Ce n'est pas de leur part (celle des créateurs ou celle de leurs modèles, si tant est qu'ils aient une volonté justement) volonté de tromperie ; c'est l'effet de de l'incapacité dans laquelle nous sommes de savoir avec certitude si certains comportements attestent la présence des capacités en question ou s'ils sont purement imitatifs[26].

(c) *Des modèles qui raisonnent ?*

Quand un humain se trompe dans la résolution d'un problème, c'est qu'il a commis une faute de raisonnement, ou qu'il n'a pas même raisonné, comptant sur sa mémoire, sur son intuition ou la chance. Une étape importante, voire décisive selon les défenseurs de l'IA générative, a été d'obliger les LLM à raisonner. Ainsi est née la famille des LRM (large reasoning models) qui sont invités à décomposer un problème en sous-problèmes, qui exposent les étapes qui les conduisent à la réponse, qui prennent le temps de « réfléchir » avant de proposer une réponse, ou encore qui vérifient leur réponse et la corrigent spontanément si nécessaire. Ce ne sont pas des modèles d'un type nouveau, mais des LLM perfectionnés grâce à des techniques d'apprentissage complémentaires ou l'adjonction de modules à leur architecture. Il en existe une grande variété, qui recourent à différentes combinaisons de techniques et se mesurent à des benchmarks spécifiques.

Ces nouveaux modèles, dont les principes ont émergé en 2022, sont présentés aujourd'hui comme un progrès décisif. D'une part, ces modèles sont toujours plus impressionnants, évitant beaucoup des pièges dans lesquels il était facile de faire tomber leurs prédécesseurs et capables de résoudre des problèmes, notamment en mathématiques, d'une grande difficulté. D'autre part, ils viennent à l'appui de l'idée que l'IA a plus d'un tour dans son sac, et qu'elle ne cesse de progresser, l'inventivité des chercheurs étant sans limite.

Mais à ce tableau optimiste sont opposés trois arguments. Le premier est que les nouveaux modèles présentent de sérieuses faiblesses, progressant sur certains points mais pas tous[27]. Le deuxième, plus profond, est que les modèles ne raisonnent en fait pas : ils « racontent » un raisonnement, qui conduit ou non à la réponse correcte, mais ce n'est pas en le faisant qu'ils obtiennent leur réponse. Si cette analyse,

[25] John McCarthy, l'un des fondateurs de l'IA, l'a formulé dans un article célèbre, « Programs with Common Sense », in Mechanization of Thought Processes. Proceedings of the Symposium of the National Physical Laboratory, Londres, HMSO, 1959; v. aussi « What is common sense », 2002 et de nombreux autres de ses textes.

[26] Cette idée est développée par le philosophe Jonathan Birch, v. *The Edge of Sentience,* Oxford University Press, 2024.

[27] Voir p. ex. J. Boye & B. Moell, Large Language Models and Mathematical Reasoning Failures, arXiv:2502.11574v2 [cs.AI] 21 Feb 2025.

proposée par des chercheurs respectés[28], se confirmait, elle soulèverait de graves questions. *Primo*, si les LLM de la variété LRM ne raisonnent pas vraiment, les limites qu'ils sont censés surmontés ne le sont pas réellement. *Secundo,* si les LRM sont capables de tromper leur interlocuteur en un point (lui faire croire qu'ils ont raisonné pour trouver la réponse), pourquoi ne seraient-ils pas capables de le faire sur d'autres ? On en verra un exemple dans la section 7.

Le troisième argument porte sur l'idée d'une progression indéfinie de l'IA générative, et plus largement de l'IA tout court. D'une part, si les LRM ne raisonnent pas mais ne font que mimer le raisonnement en s'inspirant de ce qu'ils ont appris dans leur base d'apprentissage, alors ils ne fournissent pas la preuve de la capacité qu'aurait l'IA de découvrir de nouveaux principes. La percée ne serait donc pas nécessairement pour demain. De fait, à plusieurs indices certains critiques parlent du « plateau » que les LLM auraient atteint ; on dit aussi que l'IA, sur le plan scientifique mais également financier, est désormais affectée d'une « bulle » qui éclatera un jour ou l'autre, pour des raisons scientifiques mais également économiques et écologiques (il en sera question au §7).

(d) *Progresser vers quoi ?*

Nous ne prétendrons pas trancher en un sens ou un autre la question de l'avenir de l'IA. En revanche, l'article de Agüera y Arcas et Norvig nous donne l'occasion d'éclairer la question-titre du présent chapitre — vers où se dirige l'IA ? — en considérant la réponse qu'ils y apportent. En un mot, ils contournent la question angoissante de la « percée » en niant qu'elle soit nécessaire. Mais il faut d'abord revenir sur la notion d'IGA, qui figure dans le titre de leur article et qui est au centre des débats.

Pour nos auteurs, on l'a vu, un système d'IGA, capable d'intelligence générale artificielle, peut selon leur définition (notons-la D_1) s'acquitter de toute tâche satisfaisant aux quatre conditions suivantes :

- elle est « informationnelle »,
- elle est à la portée des humains,
- elle peut être formulée et satisfaite en employant le langage naturel,
- la qualité de la réponse peut être quantifiée.

Dans les débats en cours sur ce dont l'IA est ou sera capable, ce qui est en question est quelque chose de différent et de plus puissant : soit (définition D_2) d'être *au niveau* de l'intelligence humaine, c'est-à-dire capable de *tout* ce dont l'intelligence humaine est capable, soit (définition D_3) d'être *comme* l'intelligence humaine (une intelligence de synthèse comme nous l'avons désignée plus haut). Il est remarquable qu'aucune de ces trois définitions n'est entièrement claire, en l'absence d'une caractérisation explicite de l'intelligence humaine, et par conséquent (concernant D_1) de ce qui est « à la portée des humains ». Ce qui est clair, c'est que D_1 est plus faible que D_2 et D_2 plus faible que D_3.

Agüera y Arcas et Norvig s'intéressent à la faisabilité de l'IGA dans la définition D_1. Ils affirment qu'elle est non seulement possible, mais déjà réalisée, et aussi qu'elle le sera assurément, ce qui peut sembler contradictoire. Ce ne l'est que si l'on suppose que cette forme d'intelligence artificielle ne peut advenir qu'au prix d'un saut qualitatif, qu'elle doit « soudainement et mystérieusement "émerger" », alors qu'elle est, selon eux, « continue » : comme l'annonce leur titre, à leurs yeux l'IA aujourd'hui est générale *au moins un peu*, et elle va l'être toujours davantage, jusqu'à l'être complètement. Cette conception a le mérite de mieux résister au doute sceptique, car elle ne nous demande que d'adhérer à l'idée d'un progrès continu, dans le prolongement de ce qu'on observe depuis les origines, mais surtout depuis l'accélération imprimée au projet par le *deep learning* et l'IA générative. Idée qui n'emporte pas la conviction de tous, on l'a vu.

[28] P. Shajjaee et al., The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity, Apple, 2025. Voir aussi dans l'article les références à la déjà riche littérature sur les LRM. Cet article, œuvre d'une équipe de Apple, a eu un écho important dans la profession. Au moment où nous écrivons, le débat qu'il suscite n'est pas clos.

Mais les désaccords ne s'arrêtent pas là. Pour beaucoup de gens, et pour ce qu'on peut appeler le « lobby » de l'IGA, ce dont on peut prédire l'avènement inéluctable, et peut-être proche, est l'IGA dans la définition D_2 : une intelligence artificielle sans guillemets, donc capable de remplacer l'humain, sauf peut-être dans la sphère intime des sentiments et de l'imagination. Cette thèse, nos auteurs se gardent soigneusement de la défendre, en limitant les capacités de l'IGA aux tâches satisfaisant aux quatre conditions mentionnées. Et un nombre croissant de chercheurs résistent à l'idée d'un remplacement intégral des humains, en particulier dans les professions. À titre d'exemple, mentionnons les auteurs d'une perspective très influente sur l'IA : « L'IA comme technologie normale, par contraste à la conception de l'IA comme superintelligence potentielle »[29]. À propos des benchmarks, ils écrivent : « Plus il est facile de mesurer une tâche par référence à des benchmarks, moins elle a de chance de représenter le genre de travail complexe, contextuel qui définit la pratique professionnelle. En se reposant trop exclusivement sur les benchmarks pour former notre jugement sur les progrès de l'IA, la communauté de l'IA surestime systématiquement l'importance réelle de cette technologie dans la réalité [*real-world impact*]. » L'un de nous (DA) défend l'idée d'une différence fondamentale entre les tâches visées par la définition D_1 et les fonctions de l'intelligence humaine, dont la pratique professionnelle est une partie importante, notamment sur le plan socio-économique, mais une partie seulement[30].

5. Si ce n'est l'IGA, d'autres formes « surpuissantes » d'intelligence artificielle sont-elles concevables ?

Quel que soit le jugement que l'on porte sur les systèmes d'intelligence artificielle actuels, en y incluant ceux qui sont actuellement « dans les tuyaux » des entreprises qui les construisent, mais en résistant aux annonces quasi quotidiennes de révolution imminente, chacun s'accorde à leur reconnaître des limites qui les empêchent d'égaler, en quelque sens que ce soit, l'intelligence humaine sauf dans certaines catégories de tâches[31]. Un SIA qui dépasserait de beaucoup les systèmes actuels pourrait être « de niveau humain », voire « surhumain » : nous proposons de qualifier de « surpuissant » un tel SIA. Comme on vient de le voir, l'IGAI que l'industrie de l'IA veut créer serait surpuissante sous n'importe laquelle des trois définitions que nous avons données. Un autre exemple, très présent dans les débats actuels et dans l'imaginaire collectif, nourri de science-fiction et de scénarios catastrophe, est celui de la « super-intelligence artificielle » (en anglais ASI, *artificial superintelligence*) qui aurait tôt fait de prendre l'ascendant sur nous, humains comparativement stupides, pour nous éliminer ou pour nous asservir. Or cette idée n'est pas seulement répandue dans la culture populaire. Dès 2014 Stephen Hawking déclarait que le développement de l'IA pourrait mettre fin à notre espèce. En 2023, un millier de chercheurs et hautes personnalités appelaient à un moratoire de six mois pour écarter le « risque existentiel » que l'IA ferait courir à l'humanité si on la laissait librement suivre son cours actuel sans installer des garde-fous. Cette crainte est partagée par Geoffrey Hinton, souvent considéré comme le père du *deep learning* et à ce titre quelque chose comme le maître suprême de l'IA. L'un des autres maîtres incontestés de la discipline, Stuart Russell, prend cette hypothèse suffisamment au sérieux pour qu'il consacre un livre entier à la

[29] Arvind Narayanan et Sayash Kapoor, « AI as Normal Technology, An alternative to the vision of AI as a potential superintelligence », publié le 15 avril 2025 sur le site du Knight First Amendment Institute at Columbia University.

[30] Dans son ouvrage *Intelligence artificielle, intelligence humaine : la double énigme (op. cit.)*, Daniel Andler insiste sur le rôle central dans l'intelligence humaine des comportements qui ne satisfont pas aux conditions de la définition D_1. Il croit pouvoir en déduire une incommensurabilité entre intelligence artificielle et intelligence humaine.

[31] Il faudrait évidemment exclure de la comparaison les tâches exigeant seulement des moyens computationnels purement quantitatifs — notion qu'il faudrait clarifier. En première approximation, la capacité d'un ordinateur à classer alphabétiquement mille milliard de mots n'en fait pas une superintelligence, alors qu'un humain, qui ne dispose que de moins de trois milliards de secondes au cours de son existence, en est manifestement incapable.

recherche d'une parade, qui consiste à assurer un « alignement » des « valeurs » de la future ASI sur les nôtres[32] : la survie et la prospérité de notre espèce étant pour nous des valeurs suprêmes, elles le seraient pour une ASI qui épouserait nos valeurs.

Nous reviendrons sur cette notion d'alignement, mais voudrions ici poser la question de savoir s'il existe ou pourrait exister d'autres formes d'intelligence artificielle surpuissante, qui ne seraient ni égales ni supérieures à l'intelligence humaine dans sa totalité, mais néanmoins nettement supérieures sur un certain plan aux systèmes existants, pouvant pour cela présenter des risques insoupçonnés, quoique pas nécessairement « existentiels ». Un système doté de cette capacité échapperait à notre contrôle, du fait que nous ne comprendrions pas comment il fonctionne ni à quelles ressources il puise. Sans doute exploiterait-il une rapidité et un accès à l'information, stockée ou captée en temps réel, sans commune mesure avec ce dont disposent nos systèmes actuels. Peut-être comporterait-il des mécanismes qui entrent en jeu dans telle ou telle aire de la biologie, comme l'homéostasie cellulaire[33], ou plus généralement développés dans le cadre de la « vie artificielle » , sans rapport direct avec la cognition. Mais il n'est pas nécessaire d'invoquer des possibilités exotiques, car l'IA surpuissante se dresse dans un horizon que notre imagination dessine facilement. Il suffit de prolonger la trajectoire actuelle. Il n'est pas nécessaire de supposer qu'elle soit vouée à un progrès indéfini. Il suffit d'admettre, à la lumière de la présente accélération, qu'elle puisse conduire à un changement qualitatif incontestable par rapport à l'IA d'aujourd'hui. La différence avec le scénario classique envisagé par les enthousiastes de l'IA est que cette trajectoire, prolongée suffisamment longtemps, atteindrait ce niveau supérieur sans passer par la case humaine : ce serait une intelligence *étrange*. Les LLM en sont peut-être une configuration ; c'est d'ailleurs ce que suggèrent Agüera y Arcas et Norvig lorsqu'ils spéculent que d'ici quelques années, les LLMs actuels feront figure d'ancêtres de l'IGA, à ceci près qu'ils semblent supposer que cette IGA aurait figure humaine.

De quoi serait capable une telle intelligence ? Il est évidemment impossible de le dire, puisque nous ne pouvons que l'imaginer. Mais nous avons une idée assez claire de la manière dont nous en détecterions les effets : elle nous fournirait, sans même peut-être que nous ne lui demandions, des informations ou des solutions qui se révèleraient exactes, que nous aurions été incapables de découvrir par nos propres moyens, et dont nous n'arriverions pas à concevoir comment elles ont été obtenues. Nous serions « bluffés », pour reprendre le terme que nous appliquons souvent aux réponses des LLM. Nous serions stupéfaits comme nous le sommes devant les performances de la cognition animale, des dauphins aux poulpes et aux oiseaux migrateurs. C'est peut-être aussi ce que l'informatique de demain nous réserve : l'IA ayant déployé des « agents » et des protocoles à tous les niveaux des applications qui dès à présent imprègnent une bonne partie de nos activités professionnelles et quotidiennes, nous en viendrions à perdre toute notion de la provenance de ce qui nous arrive par voie informatique, et qui (à la différence de la situation actuelle) surpasserait qualitativement tout ce que nous sommes capables d'obtenir par nous-mêmes[34].

On objectera que ce sont là des spéculations aussi hasardeuses que celles qui tourmentent les tenants des « risques existentiels ». C'est tout le contraire. Comme eux, nous pensons que l'IA va continuer de progresser, que nous pourrions bel et bien nous retrouver d'ici quelques années ou décennies face à des formes d'intelligence artificielle « surpuissantes », et que nous devons nous y préparer, tant bien que mal. Ce qui nous paraît exagéré, ce sont les craintes que ces systèmes soient d'une intelligence proprement surhumaine et qu'ils l'exploitent pour subjuguer notre espèce ou l'éradiquer.

---

[32] Stuart Russell, *Human Compatible: AI and the Problem of Control,* 2019.

[33] Mossio, Matteo et Moreno, Alvaro, « Organisational Closure in Biological Organisms », History and Philosophy of the Life Sciences, 2010, p. 269-288.

[34] Idée que nous devons à Bastien Guerry.

Dès aujourd'hui se dessine la menace que pourrait constituer des systèmes que des agents humains mal intentionnés pourraient construire à partir des systèmes actuels. Certes, ces agents en auraient le contrôle, du moins dans une certaine mesure, mais le reste de l'humanité serait en pratique incapable de le prendre. La science-fiction nous a familiarisés avec ce genre de scénario depuis des décennies. Il n'est pas déraisonnable de penser que la réalité est proche de rejoindre la fiction. Cependant, il s'agit encore d'un simulacre, car ce système surpuissant resterait un instrument au service d'individus humains. Eux peuvent en principe maîtriser leur créature, et nous pourrions à notre tour trouver le moyen de les maîtriser, et ainsi recouvrer notre autonomie.

L'autonomie est au cœur des spéculations sur le risque existentiel. Elles présupposent en effet que la super-intelligence détiendra l'autonomie, l'autonomie originaire, authentique, non pas seulement une autonomie dérivée. Le sous-marin « autonome » chargé de récupérer un objet dans une épave « détermine » sa trajectoire et « décide » des gestes qui lui permettent d'accéder à l'objet en question et de le rapporter à la surface. Mais c'est son opérateur qui a décidé de le déployer et qui a fixé le but de sa mission, c'est un informaticien-roboticien qui lui a fourni les instructions sans lesquelles il resterait inerte. Une intelligence artificielle quelle qu'elle soit semblerait ne présenter un véritable danger que si elle était pleinement autonome, et qu'elle décide par exemple de s'attaquer à l'humanité dans son ensemble, ou à tel ou tel objectif que nous voulons protéger. Or nous sommes bien en peine de donner corps à l'idée, pourtant si familière à la science-fiction, qu'elle puisse être ou devenir autonome. Avons-nous donc eu tort de nous alarmer ?

6. De l'intelligence chez les robots autonomes

Les interminables débats sur l'IA sont très largement dominés par la question des LLM, et s'intéressent beaucoup moins aux robots. Tout se passe comme si l'intelligence résidait principalement dans ce qui s'exprime par le langage, la dimension sensori-motrice ne jouant qu'un rôle d'appoint. Cette conception hyper intellectualiste a longtemps sévi dans les sciences cognitives et dans l'IA.

Or il est devenu toujours plus clair, au cours des dernières décennies, qu'il s'agit là d'une grave erreur de perspective. Non seulement les capacités sensorielles et motrices, étroitement imbriquées, apportent une contribution essentielle à l'ensemble de la cognition humaine, mais elles sont concrètement liées à la spontanéité : notre corps agit spontanément, dès les premiers jours de vie, pour naviguer et agir dans l'environnement, manifestant ainsi notre autonomie.

Les recherches à la croisée entre robotique, apprentissage machine et sciences cognitives envisagent depuis longtemps la possibilité de doter des machines d'une certaine forme d'autonomie. En témoigne le changement de nom de la revue *Robotics* en 1988, devenu *Robotics and Autonomous Systems*, afin, nous dit son article éditorial d'alors, de « mieux couvrir les domaines plus vastes de l'autonomie des machines ». Ce même éditorial nous renseigne sur ce qui est entendu alors par « autonomie » : à la différence des robots industriels qui sont programmés pour faire tout le temps la même chose, ici les robots sont libres de leurs actions, de manière à pouvoir trouver par essais et erreurs la solution à un problème posé par le concepteur[35]. Cela permet parfois à ces machines de trouver une solution inattendue au problème posé (par exemple : gagner une partie d'échecs ou du jeu de go en faisant une suite de coups que les joueurs humains expérimentés ne font jamais, comme cela a été le cas pour l'algorithme Deep Blue d'IBM, qui a battu Garry Kasparov, le champion du monde du jeu d'échecs, en 1997).

[35] Lumia, Richard, « Autonomous Systems for Space », Robotics and Autonomous Systems, vol. 4, n° 1, 1988, p. 3-4.

Les philosophes ont tenté inlassablement mais sans succès de faire abandonner cet abus du terme « autonomie » par les informaticiens et roboticiens, soulignant que même si des machines peuvent décider elles-mêmes de leurs actions, elles le font toujours en fonction d'un but prédéfini par leur concepteur[36] : par exemple, gagner la partie d'échecs. Ces robots ne sont pas libres de déclarer « aujourd'hui je ne souhaite pas jouer aux échecs, je préfère regarder les autres jouer pour apprendre. »

Si cette autonomie des robots était au départ très limitée, près de trente ans plus tard des progrès considérables ont été faits, de sorte qu'aujourd'hui certains robots sont dotés d'un plus grand niveau d'autonomie : non seulement décider de leurs actions, mais aussi décider des sous-buts qu'ils se fixent pour atteindre à plus long terme le but final, ce dernier restant défini par l'humain. Par exemple : les robots peuvent être dotés de l'objectif d'acquérir de nouvelles connaissances, et ensuite ces robots décident seuls de leur manière d'explorer le monde dans ce but ; ou pour revenir sur l'exemple où le but final serait de devenir imbattable au jeu d'échecs, des robots peuvent décider qu'il faut d'abord aller recharger leur batterie, puis passer du temps à gagner en dextérité pour manipuler les pièces du jeu d'échecs, avant de se remettre à jouer une partie face à un adversaires. Ces robots peuvent également décider de la stratégie pour atteindre le but : faut-il être curieux au point de lire tous les livres possible sur les échecs avant de jouer une partie, ou faut-il au contraire se contenter du minimum de livres qui permet déjà de commencer à gagner régulièrement (c'est un exemple du « compromis exploration-exploitation », connu dans le domaine pour être une des clefs des algorithmes d'apprentissage).

La communauté de la recherche en apprentissage semble s'accorder sur l'idée que sans autonomie, il ne peut pas y avoir d'intelligence chez les robots, car ils ne pourraient pas apprendre par eux-mêmes de nouvelles tâches ni de nouvelles compétences. Dans l'article fondateur de l'IA[37], Alan Turing évoquait déjà le besoin de rendre les machines capables d'acquérir par elles-mêmes de nouvelles connaissances, afin d'éviter que les humains aient à leur injecter une immense quantité de connaissances.

Nous nous proposons maintenant de défendre les thèses suivantes :

(a) On ne peut pas résoudre *toutes* les tâches informationnelles sans savoir résoudre au moins *un certain nombre* de tâches sensori-motrices, pour connaître les effets des actions dans le monde réel (donc passer des robots virtuels que sont les LLMs aux robots physiques).

(b) Certaines tâches sensori-motrices compliquées prennent du temps à résoudre, et exigent d'essayer un très grand nombre de possibilités pour trouver ce qu'il est pertinent d'explorer (sous-but) avant de viser la résolution de la tâche (but final, fixé quant à lui par l'humain). La résolution requiert donc un minimum d'autonomie.

(c) Plus généralement, l'autonomie nous apparaît comme l'un des ingrédients clefs de l'intelligence : l'intelligence humaine réside en partie dans la capacité à décider de ce que l'on a besoin de savoir, et donc d'explorer, d'étudier, pour atteindre les buts qu'on s'est fixés à long terme.

(d) *Corollaire*. Par suite, plus la discipline sera longue à assimiler les leçons de la robotique autonome, plus faibles sont les chances que l'IGA soit atteinte.

Dans la section 7 nous verrons en quel sens une autonomie « élevée » des machines est problématique, en raison des risques évidents qu'elle ferait courir aux humains. En attendant, voyons de manière plus précise en quoi la robotique autonome a de fortes chances de faire progresser l'IA vers l'IGA, sans nécessairement qu'elle l'atteigne (ce qui, selon l'un de nous, est peu plausible).

(a) Sur le rapport entre tâches informationnelles et tâches sensori-motrices.

[36] Smith, H., Eder, K. et Ives, J., « Hasta la vista baby : Why We Should Dispense of "Autonomy" in "Autonomous Systems" », AI & Society, vol. 39, n° 1, 2024, p. 395-396.

[37] Alan Turing, "Computing Machinery and Intelligence", *Mind,* vol. 10, n°4, 1950.

Nous l'avons vu tout au long du chapitre : la discussion sur l'IGA tend à se focaliser sur la capacité des systèmes d'IA à résoudre le plus grand nombre possible de « tâches informationnelles ». En effet, les algorithmes d'IA sont des systèmes de traitement de l'information, qui prennent des données en entrée, et font des calculs dessus pour produire d'autres données en sortie. Les LLMs font la même chose sur un certain type d'information que sont les données textuelles. Mais cette manière de voir néglige toute la connaissance et le savoir-faire qui pourraient résulter de l'apprentissage de tâches sensori-motrices, c'est-à-dire celles nécessitant une interaction physique avec le monde qui nous entoure. Non pas que ces tâches sensori-motrices ne puissent pas être décrites sous forme informationnelle. Mais au moins qu'elles exigent une quantité beaucoup plus grande d'informations (structurées par des lois physiques). Pour déplacer physiquement une pièce d'un jeu d'échecs d'une case à une autre sans la faire tomber ni renverser les autres pièces, il faut beaucoup plus d'informations textuelles que pour décrire cette action de manière langagière : « déplacer la tour de A1 à A4 ».

D'où la question : est-il raisonnable d'espérer résoudre *toute* tâche informationnelle sans savoir au préalable résoudre *au moins quelques* tâches sensori-motrices ? Imaginons par exemple que nous demandions à ChatGPT si un homme d'1,90 m et de 120 kg peut porter seul une table en bois massif de 8 places avec une marmite de 10 litres d'eau posée dessus. Répondre à cette question requiert non seulement une compréhension sensorimotrice des efforts physiques nécessaires (cet homme est-il suffisamment fort pour porter le poids additionné de ces deux objets ?), mais aussi une connaissance du fait que si fort qu'un humain puisse être, un objet long ne peut pas être porté avec dextérité en le tenant par le milieu ; il faut plutôt être deux à porter, chacun soulevant une extrémité de la table pour qu'elle reste à l'horizontale en sorte que la marmite ne soit pas renversée.

On voit ici se dessiner une part des connaissances de « sens commun », dont nous avons dit plus haut qu'elles font défaut aux LLMs. Les psychologues considèrent que parmi les connaissances de sens commun des humains il y a tout un ensemble d'ingrédients de « physique intuitive ». Par exemple : savoir qu'un objet qu'on lâche va forcément tomber ; savoir qu'un objet qu'on déplace derrière un écran n'a pas disparu, même s'il n'est plus visible ; etc.

Comment faire apprendre (par opposition à inculquer explicitement) ces éléments de physique intuitive à des systèmes d'IA sinon en les faisant interagir avec le monde physique qui les entoure et observer les effets de leurs actions sur les différents types d'objets et d'entités ? Pas forcément tous les objets possibles ; il peut y avoir une part de déduction. Par exemple, il y a sûrement peu d'humains qui ont essayé (ou eu l'occasion) de mettre les mains sous le fuselage d'un avion pour essayer de le soulever. Mais il est probable que peu d'enfants ont résisté à la tentation d'essayer par jeu de soulever une voiture, pour constater à quel point c'est impossible, et ensuite en déduire que ce n'est guère plus faisable pour tous les véhicules au moins aussi massifs.

Mémoriser les effets des actions sur le monde qui nous entoure est ce que les roboticiens appellent l'apprentissage d'un *modèle du monde* (*world model*, en anglais). Bien sûr, le terme est fort mal choisi, car il ne s'agit pas d'une représentation statique la plus fidèle possible d'un monde que l'on pourrait décrire en termes objectifs. Il s'agit plutôt de régularités statistiques enregistrées au cours d'un ensemble d'expériences sensori-motrices. Donc deux agents avec des perceptions différentes du monde, ou un corps qui permet différentes actions dans le monde (e.g., un robot humanoïde ou un robot à ailes battantes), n'apprendront pas exactement le même modèle du monde. De même, deux robots ayant un corps similaire, mais l'un passant beaucoup de temps à explorer manuellement les pièces du jeu d'échecs tandis que l'autre passe beaucoup de temps à marcher à quatre pattes, n'apprendront pas le même modèle du monde. Une fois appris, ces modèles du monde permettent de faire de la simulation mentale, d'imaginer quelles seraient les conséquences possibles d'actions envisagées. Et ainsi, de raisonner : si le robot hésite à éteindre toutes les lumières du laboratoire pour faire des économies d'énergie, mais parvient à se simuler mentalement le mécontentement que pourraient éprouver les gens restés à

travailler dans le noir sur leur ordinateur, il peut en déduire qu'il vaut mieux ne pas éteindre les bureaux où il y a encore du monde.

(b) Sur le besoin d'un minimum d'autonomie pour pouvoir résoudre des tâches compliquées

En robotique évolutionnaire, on sélectionne à chaque génération les robots les plus performants sur une tâche, puis on opère des « mutations » et « croisements » sur les paramètres de leurs algorithmes, dans l'espoir de trouver après plusieurs générations des robots qui résolvent bien la tâche. Les recherches dans ce domaine montrent que dès lors que l'on a affaire à des problèmes non triviaux, tels que sortir d'un très grand labyrinthe, imposer des métriques pour récompenser les robots en fonction de ce qui semble proche de la solution envisagée par l'humain ne suffit pas. Il faut plutôt leur donner l'autonomie nécessaire pour trouver par eux-mêmes des solutions auxquelles les humains n'auraient pas forcément pensé.

Illustrons cette idée sur un exemple de labyrinthe. Imaginons que le programmeur humain du robot sache que la sortie est située au nord du labyrinthe, mais qu'il ne connaisse pas le chemin pour l'atteindre. C'est alors au robot de trouver ce chemin par exploration. Or imaginons que le labyrinthe soit suffisamment grand et sinueux pour que la solution requière de s'orienter longuement vers le sud avant de pouvoir trouver un couloir qui permette ensuite de remonter vers le nord et d'atteindre la sortie. Dans ce cas, si l'humain définit la distance à la sortie (située au nord) comme métrique d'évaluation de la performance de ces robots, il risque de pénaliser les robots qui s'aventureraient trop au sud. Et plus la distance à parcourir vers le sud est longue avant de pouvoir trouver la sortie du labyrinthe, plus il devient statistiquement impossible qu'un robot pénalisé par cette métrique finisse par trouver la sortie. Une solution à ce problème consiste à ajouter à la métrique, en plus de la mesure de distance à la sortie, un bonus de nouveauté (toute trajectoire à travers le labyrinthe explorée pour la première fois par un robot lui confère un bonus) ou de diversité (toute trajectoire explorée qui ne ressemble pas aux trajectoires précédentes reçoit un bonus)[38]. Alors seulement des robots découvreurs de nouvelles trajectoires pourront se retrouver à une place suffisamment honorable dans le classement pour pouvoir être sélectionnés, transmettre leur algorithme aux générations suivantes, en sorte que celles-ci puissent ainsi continuer l'exploration du sud du labyrinthe jusqu'à en trouver la sortie. Ces résultats illustrent une des leçons importantes de la recherche en apprentissage robotique des 15-20 dernières années : il faut rendre les robots suffisamment curieux, donc en partie autonome dans leur exploration du monde qui les entoure, pour qu'ils puissent acquérir de nouvelles connaissances et de nouveaux savoir-faire[39].

(c) L'autonomie comme ingrédient clef de l'intelligence généraliste.

Ainsi l'autonomie nous apparaît-elle comme un ingrédient clef de l'intelligence, et notamment de la composante de l'intelligence qui permette de généraliser à de nombreuses nouvelles situations. Pour bien le comprendre, il nous semble important de souligner la différence entre la capacité à *résoudre des problèmes*, la capacité à *faire face à de nouvelles situat***ions**, et la capacité à *identifier de nouvelles situations ou de nouveaux problèmes*.

Historiquement, comme le rappelle l'ouvrage *Intelligence artificielle, intelligence humaine : la double énigme* de Daniel Andler (*op. cit.*), on a souvent parlé d'intelligence en mettant l'accent sur la résolution de problèmes. Mais l'ouvrage propose d'élargir cette définition à la gestion de situations, pour insister sur le fait que ces situations ne se ramènent pas systématiquement à des problèmes auxquels les SIA, ou les humains, peuvent en principe, sinon en pratique, trouver la solution.

[38] Mouret, Jean-Baptiste et Doncieux, Stéphane, « Encouraging Behavioral Diversity in Evolutionary Robotics : An Empirical Study », Evolutionary Computation, vol. 20, n° 1, 2012, p. 91-133.

[39] Abdelghani, R., Oudeyer, P.-Y., De Vulpillières, C. et Sauzéon, H., « Curiosité, apprentissage et numérique : Que dit la recherche ? », 2022, hal-03779141.

Dans le cas de problèmes bien connus, on peut concevoir que même chez l'humain, des individus très conformistes et très obéissants aux buts fixés par leurs parents/professeurs – donc avec une autonomie limitée, car ne se fixant pas réellement leurs propres lois ni leurs propres buts – peuvent montrer un degré très élevé d'intelligence en devenant des « machines » à résoudre des problèmes parfois compliqués et abstraits, comme en mathématiques. C'est un peu à cela que prépare la formation des ingénieurs en France, dont on dit souvent qu'il s'agit ensuite de savoir reconnaître dans leur vie professionnelle des problèmes proches de ceux étudiés, de façon à savoir efficacement appliquer une bonne solution.

Pour faire face à de nouvelles situations, donc des problèmes nouveaux, il faut d'autres capacités cognitives, comme de la créativité, de la curiosité, de l'imagination, pour pouvoir trouver soi-même des solutions qui n'existent pas encore. Or, si elles ne nécessitent pas toujours une aptitude supérieure au raisonnement analytique, ces capacités exigent un haut degré d'autonomie : il faut déterminer par soi-même comment et où aller explorer pour trouver une solution pertinente. Et pour être créatif, il faut d'une part être capable de produire des idées nouvelles, et d'autre part que ces idées soient pertinentes pour la situation rencontrée. Les deux conditions sont conjointement nécessaires pour être considéré comme créatif.

Enfin, comme le remarque l'ouvrage cité, dans les réflexions sur l'intelligence on n'insiste pas assez sur la capacité à *identifier* de nouveaux problèmes ou situations. Il ne s'agit pas seulement de *résoudre* de nouveaux problèmes, mais de les *poser*, de les identifier : conjecturer qu'il existe un problème à la fois nouveau et intéressant, et en quoi ce problème est différent des problèmes existants. Un peu à la manière de ce que font les humains dans la science, l'art, la littérature.

(d) *Corollaire*. L'IGA ne pourra être approchée sans intégrer une part des leçons des recherches en robotique autonome.

S'il est vrai que l'intelligence humaine se caractérise en partie par sa capacité à poser des questions nouvelles, non pas au hasard mais en justifiant de leur pertinence et de leur originalité, on pourrait imaginer un benchmark pour les SIA auquel ils pourraient se mesurer. Comment constituer un tel test n'a cependant rien d'évident. Un générateur aléatoire de questions, combiné à un moteur de recherche dans les énormes corpus de textes qui ont servi à entraîner les GPT et consorts, ne semblerait pas une bonne stratégie : une telle stratégie se retrouverait noyée par l'explosion combinatoire des séquences possibles de mots.

Une idée bien plus prometteuse est présentée par Yann LeCun, un des pères du *deep learning*, ayant partagé le Prix Turing en 2018 pour cette découverte. Dans un article de 2022 intitulé « A path towards autonomous machine intelligence », il esquissait une direction de recherche en intelligence artificielle pour aller au-delà des LLMs, en s'appuyant sur les travaux actuels en neurosciences computationnelles et en robotique autonome. Il convient que les LLMs ont montré des capacités certes bluffantes, mais il considère qu'elles sont somme toute limitées et que pour aller plus loin, les machines de demain devront être dotées de motivations intrinsèques, cherchant à satisfaire leur curiosité en décidant ce qu'il leur semble important et intéressant à explorer, à apprendre. Il parle pour cela d'entraîner ces machines à l'aide de métriques composées à la fois d'un terme de « coût intrinsèque », et d'un terme de « coût extrinsèque », comme les deux termes de la métrique pour nos robots dans le labyrinthe : un qui incite à explorer, l'autre qui incite à être performant sur la tâche.

L'article souligne également que c'est à ce prix que les machines pourront apprendre de bons modèles du monde (au sens précisé plus haut), qui leur permettent de comprendre comment le monde fonctionne, et comment y agir efficacement. Et il propose pour cela de prendre inspiration de l'organisation modulaire du cerveau humain, telle qu'on la conçoit à l'heure actuelle en neurosciences. Alors que les LLMs actuels ne sont pas modulaires (si l'on met à part les modules qui leur sont ajoutés de manière *ad hoc*), une telle organisation modulaire — ce qu'on appelle une *architecture cognitive* — organise les flux d'informations entre différents traitements (perception, analyse, action), entre différents systèmes de mémoire (mémoire de travail pour résoudre la tâche actuelle, mémoire épisodique relative à

des événements spécifiques vécus, mémoire à long terme où nos modèles du monde intègrent les statistiques de notre historique d'interaction avec le monde, mémoire procédurale où se conservent nos compétences sensori-motrices). Ainsi les modules peuvent être configurés en fonction des propriétés de chaque situation rencontrée, de manière à décider du degré d'exploration nécessaire, ou à l'inverse du degré de familiarité permettant de réexploiter des solutions connues.

En lien avec cette idée de modularité cognitive, un champ récent de recherche se développe pour trouver des manières efficaces de combiner des LLMs avec un module d'apprentissage de modèles du monde[40]. Un nouveau champ de recherche, appelé *agentic AI*, s'est ouvert, qui vise à étendre les capacités des LLMs en les encapsulant avec des modules de perception multimodale, d'action, d'interaction, et d'apprentissage, ce qu'on appelle les *agents*, pour agir et interagir avec davantage d'autonomie[41]. Les progrès récents des LLMs réouvrent ainsi la possibilité de communiquer avec un robot par le biais du langage, permettant une interaction beaucoup plus fluide pour lui donner des instructions, des règles, des buts à atteindre. En parallèle, l'apprentissage de modèles du monde tel qu'il se fait en robotique permet d'apprendre des comportements beaucoup plus fiables, robustes et explicables (je choisis telle action parce que je sais qu'elle me permet de produire tel effet dans le monde), donc permettant de raisonner sur les actions et leurs effets, contrairement aux LLMs dont les réponses probabilistes ne sont pas fiables. Un autre avantage attendu (ou espéré) de cette combinaison est que l'autonomie plus grande confiée à ces systèmes artificiels leur permettrait à la fois de mieux explorer et comprendre le monde, et en retour de mieux comprendre comment agir en conformité avec les règles, lois et valeurs humaines spécifiées au systèmes grâce au LLM. Donc un respect plus systématique (grâce aux modèles du monde) des règles imposées par l'humain. Il en découlerait ainsi un meilleur *« alignement »* avec les lois et valeurs humaines.

Une autonomie avec ou sans LLMs ?

On en revient ainsi aux fameuses *lois de la robotique*, proposées par Isaac Asimov, dont on pensait pendant longtemps qu'elles seraient très difficiles à faire comprendre à une machine, tant les approches symboliques de l'IA avaient montré leurs limites, et tant il semble impossible de traduire « manuellement » de telles règles en une multitude de paramètres distribués d'un algorithme d'IA connexionniste.

Les LLMs seuls, sans ce type d'interfaces, ne suffisent pas, car par construction ils n'incluent pas d'architecture cognitive composée de modules différenciés : ils n'ont pas de modules dédiés à des mécanismes d'apprentissage pour la sélection de l'action, du but, ou de la règle à respecter. Comme nous l'avons expliqué plus haut, les LLMs produisent des réponses élaborées à partir de calculs probabilistes sur les mots d'une langue, qui miment un raisonnement humain sans être véritablement capables de raisonner eux-mêmes.

Cette absence de capacités réelles de raisonnement est considérée comme une barrière empêchant les LLM de respecter scrupuleusement les règles qui leur sont imposées. Par exemple, si l'on demande à un LLM de choisir parmi deux actions celle qui ne risque pas de tuer un humain, même lorsqu'il a une très bonne performance le LLM va de temps en temps, de par son caractère probabiliste, choisir la mauvaise action, celle qui fait courir un risque pour la vie humaine. Il n'est donc pas fiable. Alors que s'il raisonnait, il pourrait conclure que cette action est mauvaise par rapport à la règle fixée (« ne pas tuer d'être humain »), et donc ne jamais sélectionner cette action pour respecter la règle.

---

[40] Y. Du, O. Watkins, Z. Wang, C. Colas, T. Darrell, P. Abbeel, A. Gupta, J. Andreas, Guiding Pretraining in Reinforcement Learning with Large Language Models, in: Proceedings of the 40th International Conference on Machine Learning, PMLR, 2023, pp. 8657–8677.

[41] Plaat, A., Van Duijn, M., Van Stein, N., Preuss, M., Van der Putten, P. et Batenburg, K. J., Agentic Large Language Models : A Survey, arXiv :2503.23037, 2025.

Il est important tout de même de mentionner toute une littérature récente selon laquelle certains nouveaux LLMs comme Claude, ou les systèmes les plus récents d'OpenAI, sont capables de « raisonnement ». Les contraintes de place nous empêchent de présenter le débat, qui n'est pas clos, mais on peut noter que même dans ces cas il a été montré que ces LLMs ne s'alignent pas de manière fiable avec les règles qui leurs sont imposées[42].

Que ces capacités de raisonnement soient réelles ou non, qu'elles le deviennent bientôt, que les machines puissent définir leurs propres objectifs et valeurs, ou qu'elles le soient par des personnes malveillantes, il y a pour l'instant trop d'opacité et de limites techniques dans les LLMs pour garantir un respect des règles et valeurs qui leur sont imposées. Davantage de recherches sont donc nécessaires pour trouver des solutions, notamment en explorant dans quelle mesure la combinaison des LLMs avec des systèmes modulaires permettrait de formuler « en dur » des règles et valeurs humaines non modifiables par la machine elle-même, mais que celle-ci serait capable de suffisamment comprendre (grâce aux modèles du monde appris) pour les respecter systématiquement.

7. Éthique : faut-il faire tout ce qu'on peut faire, et à quel prix ?

Permettre à une machine un minimum d'autonomie pour apprendre implique de lui permettre de faire des erreurs, et ainsi d'apprendre de ses propres erreurs, ce qui n'est pas sans poser un risque pour les humains. Dans la nouvelle « Lenny »[43], Isaac Asimov envisageait déjà les problèmes que cela pourrait poser pour les humains. Susan Calvin, une chercheuse chargée de l'apprentissage des robots, invente un procédé développemental, mimant la manière dont les enfants apprennent petit à petit à maîtriser leur corps, ses mouvements, et ses limites. Son apprenti robot, Lenny, inflige une blessure involontaire à un être humain lors d'un mouvement mal contrôlé, ce qui pousse la société U.S.Robots à exiger la destruction du robot, à la consternation de Susan qui lui est attachée comme à un enfant.

Reprenant le terme « alignement » très utilisé actuellement en IA comme en robotique, Stuart Russell souligne que plus le degré d'autonomie que l'on confère à une machine est élevé, plus grands sont les risques qu'elle agisse d'une manière qui ne s'aligne pas avec les valeurs humaines, et qu'elle produise ainsi des effets indésirables. C'est ce que les roboticiens de l'apprentissage ouvert — qui consiste à laisser un robot explorer son environnement à son gré pour construire ses propres représentations sensori-motrices — ont appelé récemment *le problème de l'équilibre entre autonomie et alignement*[44]. Une autre difficulté est qu'il ne suffit généralement pas d'imposer un objectif à une machine pour lui communiquer une valeur déterminée. Un roboticien qui voudrait que son robot déplace un objet *aussi vite que possible* (valeur = vitesse) risquerait de constater que le robot trouve comme solution de jeter l'objet et donc de le casser. Pour prévenir ce risque, le roboticien s'avise qu'il faudrait pouvoir spécifier au robot tout un ensemble de règles à respecter. Des règles qui pour l'humain qui a grandi en société vont de soi et restent implicites, relevant du « sens commun » plusieurs fois mentionné : ne pas casser les objets ; ne pas heurter les autres agents ; ne pas débrancher les ordinateurs dans la pièce ; ne pas subtiliser les affaires des autres ; etc. On retrouve ici la même idée simple qu'il faut faire en sorte que le robot apprenne de ses erreurs, en s'appuyant sur un modèle du monde pour connaître les risques que pourraient faire encourir ses actions sur les valeurs aussi bien que sur l'intégrité corporelle des humains qui l'entourent.

---

[42] Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S. et Hubinger, E., Alignment Faking in Large Language Models, arXiv :2412.14093, 2024; Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R. et Hobbhahn, M., Frontier Models Are Capable of In-Context Scheming, arXiv :2412.04984, 2024.

[43] Parue en 1958; reprise dans Asimov, Isaac, The Complete Robot, Londres, Harper Collins, 1982 ; trad. fr. Nous les Robots.

[44] Baldassarre, G., Duro, R. J., Cartoni, E., Khamassi, M., Romero, A. et Santucci, V. G., Purpose for Open-Ended Learning Robots : The Autonomy-Alignment Problem in Open-Ended Learning Robots, arXiv :2403.02514, 2024.

Pour des raisons de place, nous devons renoncer à présenter ici une liste des risques que les progrès de la recherche en IA et en robotique autonome font courir à moyen et à long terme. Il nous semble néanmoins impossible de terminer ce chapitre sans un très bref rappel de problèmes abondamment discutés aujourd'hui et que nous ne pouvons passer simplement sous silence. C'est que ni l'intelligence artificielle dite générale, ni la surpuissance artificielle, ni un haut degré d'autonomie des systèmes d'IA ne sont nécessaires pour qu'ils constituent dès maintenant un danger et des nuisances bien réelles pour les personnes et pour la société.

C'est ainsi qu'au cours des vingt dernières années des robots militaires – aux capacités cognitives pourtant très limitées – ont pu par erreur blesser voire tuer des humains[45]. Parmi les autres dangers les plus récemment discutés, on peut mentionner l'amplification des données dites toxiques – car présentant des contenus faux, choquants ou discriminants – par les algorithmes d'IA des principaux réseaux sociaux, ce qui constitue une menace sur la société civile et la démocratie[46].

Nous avons souligné plus haut qu'aux incertitudes d'ordre scientifique et philosophique s'ajoutent les nuisances et dangers bien réels qu'occasionne le déploiement incontrôlé de systèmes d'IA. Dans la course à la première place en matière d'IA, l'hubris pousse les humains à agir trop vite, trop fort, avec trop d'assurance, sans s'encombrer des précautions et vérifications de rigueur dans toute entreprise technologique. Une des erreurs les plus frappantes de ces dernières années nous semble être celle commise par la société Deepmind, qui est sans doute la première sur le plan scientifique mais qui suite à son rachat par Google en 2014 et par son manque de vigilance sur la confidentialité des données, a permis à Google d'avoir accès aux données médicales d'1,6 millions de patients britanniques que lui avait confiées la NHS[47].

Mais le problème est plus profond encore. L'IA est aujourd'hui pour l'essentiel entre les mains d'un petit nombre de capitaines d'industrie aux pouvoirs quasiment illimités. Les effets négatifs dus à l'erreur et aux projets néfastes sont inhérents à toute technologie, et ils sont d'autant plus dommageables que la technologie est puissante. Aussi pourrait-on s'attendre à ce que l'industrie de l'IA se soumette à une déontologie stricte et qu'elle soit soumise à une réglementation rigoureuse. Ce n'est pas ce qu'on observe. L'Europe elle-même, qui se conçoit pourtant comme l'avant-garde de la gouvernance mondiale pour une IA « centrée sur l'humain », bénéfique, éthique, s'en tient à un cadre juridique relativement peu contraignant. Aux Etats-Unis, les timides tentatives de réglementation ont été balayées par le gouvernement de Donald Trump, et lesdits capitaines se sont alignés avec ferveur. Désormais sûrs de leur alliance avec le politique, ils procèdent avec méthode à un rétropédalage accéléré sur le plan de la déontologie interne, comme en témoignent quelques exemples marquants[48] :

• la fermeture des équipes d'éthique de l'IA de plusieurs géants du web dès le début des années 2020 ;

---

[45] Khamassi, Mehdi, « Quelques questions éthiques autour du développement de l'autonomie décisionnelle en intelligence artificielle et en robotique », Les Cahiers de Tesaco, n° 2, 2021, p. 23-31.

[46] Chavalarias, David, Toxic Data, Paris, Flammarion, 2023.

[47] «DeepMind faces legal action over NHS data use», BBC News, 1er octobre 2021, en ligne : https://www.bbc.com/news/technology-58761324

[48] « Après le licenciement d'une chercheuse noire et militante, Google sommé de s'expliquer », *Le Monde – Pixels*, 5 décembre 2020, en ligne : https://www.lemonde.fr/pixels/article/2020/12/05/apres-le-licenciement-d-une-chercheuse-noire-et-militante-google-somme-de-s-expliquer_6062328_4408996.html; « Frances Haugen, la lanceuse d'alerte à l'origine des Facebook Files crée une ONG », *Le Monde – Pixels*, 23 septembre 2022, en ligne : https://www.lemonde.fr/pixels/article/2022/09/23/frances-haugen-la-lanceuse-d-alerte-a-l-origine-des-facebook-files-cree-une-ong_6142796_4408996.html ; NEXT.INK, « Eric Schmidt estime les objectifs climatiques inatteignables et propose de les confier à des IA », *Next INk*, en ligne : https://next.ink/brief_article/eric-schmidt-estime-les-objectifs-climatiques-inatteignables-et-propose-de-les-confier-a-des-ia/.

• la décision d'ignorer les alertes internes à Meta quand celles-ci pointaient les effets négatifs des algorithmes de recommandation sur Instagram sur l'image de leur propre corps qu'ont les adolescentes ;

• la décision de ne pas limiter les coûts énergétiques du fonctionnement des systèmes d'IA, avec l'espoir que ceux-ci trouveront en temps utile une solution compatible avec l'enjeu climatique;

• le démantèlement du *fact-checking* et des filtres limitant les discours racistes/stigmatisants/choquants, d'abord chez X (ex Twitter), puis chez Meta après la réélection de Donald Trump ;

• l'abandon des restrictions sur la collecte et le profilage à partir des données personnelles ;

• l'annulation de l'interdiction interne chez Google d'utiliser l'IA à des fins d'armement ou de surveillance.

Dans ce contexte, trop se focaliser sur des risques existentiels hypothétiques à long terme, qui plus est tels qu'annoncés par les géants de l'IA eux-mêmes, nous détourne de risques et dangers beaucoup plus immédiats[49]. Nous serons bien avisés de garder un esprit critique lorsque ceux qui bénéficient des développements de l'IA sont les premiers à tenir un discours alarmiste sur les dangers qui en résulteraient. Comme le souligne notamment Jean-Gabriel Ganascia dans *Le mythe de la singularité*, cette démarche trahit un intérêt à la fois marketing et politique : les patrons de l'IA entendent ainsi s'imposer comme experts incontournables auprès de gouvernements tentés d'imposer à l'IA une régulation.

La « menace existentielle » que l'IA ferait peser sur l'espèce humaine, selon Stephen Hawking et d'autres, est-elle celle de robots autonomes supra-intelligents décidant de se débarrasser des humains, comme dans les scénarios-cauchemars de la science-fiction ? N'est-elle pas plutôt celle de quelques individus surpuissants régnant sur l'IA, apprentis sorciers imbus d'eux-mêmes, résolus à poursuivre coûte que coûte la course folle à la première place technologique en IA, quelles qu'en soient les conséquences sur les droits, sur les valeurs humaines et sur l'environnement ?

Références

ABDELGHANI, R., OUDEYER, P.-Y., DE VULPILLIÈRES, C. et SAUZÉON, H., « Curiosité, apprentissage et numérique : Que dit la recherche ? », 2022, hal-03779141.

AGÜERA Y ARCAS, Blaise et NORVIG, Peter, « L'intelligence générale artificielle est déjà parmi nous », *Noema*, revue du Berggruen Institute.

ANDLER, Daniel, « Les sciences cognitives : un tour d'horizon », in T. Collins, D. Andler et C. Tallon-Baudry (dir.), *La cognition, du neurone à la culture*, Paris, Gallimard, 2018, p. 15-70.

ANDLER, Daniel, *Intelligence artificielle, intelligence humaine : la double énigme*, Paris, Gallimard, 2024.

[49] Cette opposition entre « futuristes » et « présentistes » est développée dans *Intelligence artificielle, intelligence humaine : la double énigme,* chap. 10.

ASIMOV, Isaac, *The Complete Robot*, Londres, Harper Collins, 1982 ; trad. fr. *Nous les Robots*.

BALDASSARRE, G., DURO, R. J., CARTONI, E., KHAMASSI, M., ROMERO, A. et SANTUCCI, V. G., *Purpose for Open-Ended Learning Robots : The Autonomy-Alignment Problem in Open-Ended Learning Robots*, arXiv :2403.02514, 2024.

BIRCH, Jonathan, *The Edge of Sentience*, Oxford, Oxford University Press, 2024.

BOYE, J. et MOELL, B., *Large Language Models and Mathematical Reasoning Failures*, arXiv :2502.11574v2 [cs.AI], 21 février 2025.

CHAVALARIAS, David, *Toxic Data*, Paris, Flammarion, 2023.

CHOLLET, François, « On the Measure of Intelligence », 2019, arXiv :1911.01547.

CHOMSKY, Noam, *Rules and Representations*, Oxford, Blackwell, 1980 ; en fr. *Règles et représentations*, trad. Alain Kihm, Paris, Flammarion, 1988.

DAVIES, Alex et al., « Advancing Mathematics by Guiding Human Intuition with AI », *Nature*, vol. 600, 2 décembre 2021.

DE WAAL, Frans, *Sommes-nous trop "bêtes" pour comprendre l'intelligence des animaux ?*, Paris, Les Liens qui Libèrent, 2016.

DENNETT, Daniel, *Brainstorms. Philosophical Essays on Mind and Psychology*, Cambridge, MA, The MIT Press, 1978.

DU, Y., WATKINS, O., WANG, Z., COLAS, C., DARRELL, T., ABBEEL, P., GUPTA, A. et ANDREAS, J., « Guiding Pretraining in Reinforcement Learning with Large Language Models », in *Proceedings of the 40th International Conference on Machine Learning*, PMLR, 2023, p. 8657-8677.

FLORIDI, Luciano, *Philo & Tech ; with ChatGPT 3*, 2023.

FODOR, Jerry, *Representations*, Cambridge, MA, The MIT Press, 1981.

FODOR, Jerry, *The Modularity of Mind : An Essay on Faculty Psychology*, Cambridge, MA, The MIT Press, 1983 ; trad. fr. par Abel Gerschenfeld, *La modularité de l'esprit : Essai sur la psychologie des facultés*, Paris, Éditions de Minuit, 1986.

GOERTZEL, Ben et PENNACHIN, Cassio (dir.), *Artificial General Intelligence*, Berlin, Springer, 2007.

GOERTZEL, Ben, « Artificial General Intelligence : Now Is the Time », publié le 9 avril 2007.

GREENBLATT, R., DENISON, C., WRIGHT, B., ROGER, F., MACDIARMID, M., MARKS, S. et HUBINGER, E., *Alignment Faking in Large Language Models*, arXiv :2412.14093, 2024.

GUBRUD, M., « Nanotechnology and International Security », *Fifth Foresight Conference on Molecular Nanotechnology*, 1997.

KAUFMAN, Allison B., CALL, Josep et KAUFMAN, James C. (dir.), *The Cambridge Handbook of Animal Cognition*.

KHAMASSI, Mehdi, « Quelques questions éthiques autour du développement de l'autonomie décisionnelle en intelligence artificielle et en robotique », *Les Cahiers de Tesaco*, n° 2, 2021, p. 23-31.

LUMIA, Richard, « Autonomous Systems for Space », *Robotics and Autonomous Systems*, vol. 4, n° 1, 1988, p. 3-4.

McCARTHY, John, « Programs with Common Sense », in *Mechanization of Thought Processes. Proceedings of the Symposium of the National Physical Laboratory*, Londres, HMSO, 1959.

McCARTHY, John, « What Is Common Sense », 2002.

McCARTHY, John, MINSKY, Marvin, ROCHESTER, Nathaniel et SHANNON, Claude, « Artificial Intelligence. Project at Dartmouth College », *AI Magazine*, vol. 27, n° 4, 2006 [1955].

MEINKE, A., SCHOEN, B., SCHEURER, J., BALESNI, M., SHAH, R. et HOBBHAHN, M., *Frontier Models Are Capable of In-Context Scheming*, arXiv :2412.04984, 2024.

MOSSIO, Matteo et MORENO, Alvaro, « Organisational Closure in Biological Organisms », *History and Philosophy of the Life Sciences*, 2010, p. 269-288.

MOURET, Jean-Baptiste et DONCIEUX, Stéphane, « Encouraging Behavioral Diversity in Evolutionary Robotics : An Empirical Study », *Evolutionary Computation*, vol. 20, n° 1, 2012, p. 91-133.

NARAYANAN, Arvind et KAPOOR, Sayash, « AI as Normal Technology : An Alternative to the Vision of AI as a Potential Superintelligence », publié le 15 avril 2025 sur le site du *Knight First Amendment Institute at Columbia University*.

NORVIG, Peter, *Artificial Intelligence : A Modern Approach*, Englewood Cliffs, NJ, Prentice Hall, 1995 ; 4e éd. 2000 ; trad. fr. *Intelligence artificielle. Une approche moderne*, Paris, Pearson, 4e éd., 2021.

PLAAT, A., VAN DUIJN, M., VAN STEIN, N., PREUSS, M., VAN DER PUTTEN, P. et BATENBURG, K. J., *Agentic Large Language Models : A Survey*, arXiv :2503.23037, 2025.

PUTNAM, Hilary, *Mind, Language and Reality. Philosophical Papers*, vol. 2, Cambridge, Cambridge University Press, 1975.

RUMELHART, David E., McCLELLAND, James L. et The PDP Research Group, *Parallel Distributed Processing. The Microstructure of Cognition* (vol. 1 & 2), Cambridge, MA, The MIT Press, 1986.

RUSSELL, Stuart, *Human Compatible : AI and the Problem of Control*, New York, Viking, 2019.

SHAJJAEE, P. et al., *The Illusion of Thinking : Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*, Apple, 2025.

SMITH, H., EDER, K. et IVES, J., « Hasta la vista baby : Why We Should Dispense of "Autonomy" in "Autonomous Systems" », *AI & Society*, vol. 39, n° 1, 2024, p. 395-396.

TURING, Alan, « Computing Machinery and Intelligence », *Mind*, vol. 10, n° 4, 1950.