Regards philosophiques sur l'intelligence artificielle : est-ce que les machines pensent ?

Auteur : Belloi, Théo
Promoteur(s) : Leclercq, Bruno
Faculté : Faculté de Philosophie et Lettres
Diplôme : Master en philosophie, à finalité didactique
Année académique : 2024-2025
URI/URL : http://hdl.handle.net/2268.2/24536

Travail de fin d'études

(PTFE0014-1)

« Regards philosophiques sur l'intelligence artificielle : est-ce que les machines pensent ? »

Présenté par Théo Belloï

en vue de l'obtention du master en philosophie, à finalité didactique.

À l'attention de Bruno LECLERCQ, promoteur,

Et aux membres du jury :

Laurence BOUQUIAUX

Denis SERON

Année académique 2024-2025

REMERCIEMENTS

À mes grands-parents, qui ont encouragé leurs enfants (et leurs petits-enfants) à poursuivre des études, et plus particulièrement à Josée, qui nous a quittés l'année passée.

À mes parents pour leur soutien indéfectible.

À mes amis, en Belgique comme au Brésil, pour leur présence et leur amitié.

Je tiens aussi à exprimer ma gratitude à mon promoteur Bruno LECLERCQ, pour sa disponibilité, sa motivation et une pédagogie qui est une source d'inspiration.

Introduction

En 1956, Herbert Simon et Allen Newell, deux chercheurs en sciences cognitives et en informatique, ont présenté lors d'une célèbre conférence à Dartmouth ce qu'ils ont appelé le « *Logic Theorist* », un programme capable de prouver des théorèmes de logique mathématique. Ce programme parvenait ainsi à simuler certaines capacités du raisonnement humain. Ils ont ensuite avancé une thèse surprenante et renversante pour l'époque : les machines peuvent simuler des processus de la pensée humaine.

L'ambition de ces chercheurs à cette époque était celle-ci : puisque le cerveau humain est composé de principes généraux d'organisation résultant du traitement d'informations diverses, il était dès lors possible selon eux de reproduire ces derniers à travers des systèmes artificiels que l'on retrouve aujourd'hui dans les ordinateurs. Ils pensaient ainsi que ces machines allaient livrer à l'humanité les secrets des rouages du fonctionnement de la pensée humaine pour ensuite aboutir à la création d'une machine pensante qui bouleverserait le monde des sciences cognitives. Naturellement, au moment de la présentation du résultat de leur recherche, ces deux pionniers de l'IA estimaient que cette révolution allait avoir lieu dans les années à venir. Mais qu'en est-il vraiment, à la lumière de toutes les informations dont nous disposons au sujet de l'IA et de son évolution ?

Plus proche de nous, la France a organisé le 10 et le 11 février 2025 un sommet pour l'action sur l'IA rassemblant une panoplie d'experts et décideurs du monde entier. Parmi les objectifs principaux fixés par ce sommet, il a été également question d'insister sur l'importance de l'accessibilité, de la transparence, de l'éthique et de la sécurité liées à l'IA. Ce sommet a été aussi l'occasion de consacrer des milliards d'euros au financement du développement de l'IA ainsi que de promouvoir la formation dans ce secteur en pleine effervescence. L'IA est désormais devenue une préoccupation mondiale.

En parallèle de ces perspectives optimistes, certains experts se sont aussi réunis en organisant des « contre-sommets » pour mettre en garde contre les impacts sociétaux de l'IA. Parmi les points qui ont été relevés et qui suscitent un certain nombre d'inquiétudes, il y a ceux liés aux biais algorithmiques, aux conséquences sur le marché du travail ainsi qu'à l'impact environnemental causé par le développement et l'utilisation de cette technologie.

En somme et depuis ses débuts, l'IA agite beaucoup les esprits et ces dernières années ont été particulièrement marquées par un développement fulgurant ainsi qu'une grande médiatisation de cette technologie informatique. Dès lors, quelle position pouvons-nous raisonnablement adopter, entre, d'un côté, les partisans d'une IA forte qui défendent l'idée qu'elle surplombera l'intelligence humaine et qu'elle révolutionnera radicalement des domaines tels que la médecine, l'éducation, le travail ainsi que la science, etc.

Ou de l'autre, les partisans de l'IA faible, qui, à l'inverse des premiers, estiment qu'il sera impossible de concevoir une machine consciente capable de penser par elle-même. Ces derniers affirment également que cette IA n'est qu'une simulation d'intelligence et que donc, par conséquent, il n'y a pas de raison de penser qu'elle remplacera l'apport humain.

Afin de mener à bien notre question de recherche et cette enquête sans *a priori*, nous avons décidé de convoquer des philosophes emblématiques dans le champ de la philosophie de l'esprit qui se sont déjà exprimés sur la question déjà à partir des années quatre-vingt et nonante et qui ont également partagé leur positionnement dans des articles ultérieurs et plus proches de nous dans le temps.

La première partie de ce travail consistera d'abord à définir ce que l'on entend par intelligence artificielle. Pour ce faire, nous ferons appel à Daniel Andler qui nous enjoint à distinguer trois grandes périodes qui caractérisent cette technologie. Nous verrons que chacune possède son fonctionnement propre ainsi que des implications différentes. De plus, cette observation nous permettra d'examiner dans un deuxième temps si les arguments philosophiques et critiques de l'IA sont toujours d'actualité ou non.

C'est donc en deuxième partie de ce travail que nous allons faire état des différents arguments philosophiques qui ont marqué le débat au sein de la philosophie de l'esprit, entre d'un côté ceux qui vont dans le sens de l'IA forte ou les optimistes qui estiment que celle-ci sera capable de surpasser les capacités humaines et de l'autre, ceux qui abondent dans le sens de l'IA faible en postulant qu'elle ne dépassera jamais certaines limites et qu'elle ne parviendra pas à résoudre des problèmes singulièrement humains.

En lien avec ces différents arguments, nous allons discuter et essayer de déterminer si ces derniers sont toujours d'actualité. Par exemple, est-ce que l'argument de la chambre chinoise et sa démonstration de la cécité sémantique sont toujours pertinentes ? Est-ce que les limites de l'IA soulevées précédemment ont été dépassées ? Est-ce que l'IA connexionniste change la donne par rapport aux critiques soulevées par les partisans de l'IA faible ?

D'une façon plus large et dans la troisième partie de ce travail, nous allons nous demander s'il existe toujours une spécificité proprement humaine que l'IA ne pourra pas reproduire. Comment la phénoménologie pourrait-elle nous éclairer pour aborder cette question ? Est-ce que les arguments soulevés par Hubert Dreyfus en son temps sont-ils toujours pertinents ? En filigrane, qu'est-ce qui manque à l'IA actuelle pour se rapprocher de l'intelligence humaine ? Ces interrogations constitueront le fil rouge de notre analyse après avoir examiné différentes positions philosophiques, de Turing à Andler, en passant par Dennett, Searle, Chalmers et Hubert Dreyfus.

De cette façon, nous allons ici nous employer à défendre une pensée qui se fonde sur une compréhension réalisée à partir d'un corps percevant, lui-même immergé dans un monde au sein duquel il interagit. Nous tenterons ainsi de démontrer que notre propre rapport au monde n'est pas reproductible par une machine. Cette dernière partie consiste donc à défendre une position qui se situe entre l'IA faible et l'IA forte tout en répondant à la question initiale de notre travail.

PARTIE 1 : Qu'est-ce que l'intelligence artificielle ? Caractéristiques, exemples, limites et horizon

Afin de bien comprendre de quel sujet nous parlons, il convient tout d'abord d'expliciter ce que l'on entend par l'intelligence artificielle. C'est pourquoi nous allons nous appuyer sur le travail du philosophe et mathématicien Daniel Andler qui précise qu'il existe non pas une, mais trois types d'intelligence artificielle avec chacune leur mode de fonctionnement propre.

Ensuite, lorsque nous aborderons la deuxième partie de notre enquête, nous examinerons et apporterons un regard critique sur un certain nombre d'auteurs tant optimistes que pessimistes quant à l'IA. En effet, ceux-ci se sont notamment exprimés sur des IA qui occupaient les devants de la scène au moment de la rédaction de leurs articles. C'est pourquoi nous prendrons le soin d'exemplifier ces différentes IA à travers celles qui ont été citées par ces auteurs. Cette démarche a pour but d'apporter de la clarification quand elles seront abordées plus tard dans le développement de notre propos.

I.1. Qu'est-ce que l'IA symbolique et comment fonctionne-t-elle ?

Daniel Andler décrit l'IA symbolique comme un système fondé sur des états et sur les processus de transition entre ces états[1]. Par « état », il entend des représentations symboliques d'informations, c'est-à-dire des propositions ou expressions que l'IA tient pour vraies. Ainsi, l'IA symbolique repose sur un stock de connaissances codées, auquel elle accède pour accomplir une tâche spécifique.

Prenons l'exemple d'un jeu d'échecs. L'IA dispose d'un ensemble de règles internes, comme les règles du jeu et elle reçoit une consigne : jouer avec les pions blancs ou noirs pour vaincre l'adversaire. Pour accomplir cela, elle applique des règles logiques pour transformer son état initial en un état final désirable (c'est-à-dire gagner la partie), ce qui implique, comme le souligne Andler, qu'elle « se conforme à la rationalité »[2].

---

[1] Daniel Andler, *Intelligence artificielle, intelligence humaine : la double énigme*, Paris, Gallimard, « NRF Essais », 2023, p. 77.

[2] *Ibid.,* p. 79.

Contrairement à l'intelligence humaine qui renvoie à l'imagination, à l'intuition ou à l'expérience, cette IA ne peut pas recourir à ces caractéristiques. Ces notions sont de l'ordre du mystère et lui sont donc inaccessibles, car elle ne fait qu'appliquer des règles qui lui ont été transmises

« Autrement dit, c'est le modélisateur humain, le programmeur, qui découvre une solution par ses moyens propres de chercheur ou d'ingénieur, et l'implémente dans le programme : il met ainsi l'intelligence « en conserve ». La difficulté est pour le programmeur, pas pour le système artificiel intelligent »[3].

Toutefois, il existe une difficulté majeure à laquelle l'IA ne peut échapper : celle qui met ses ressources dont elle dispose et qui sont mises à rude épreuve. Ainsi, l'IA ne pourra jamais aller au-delà de ses propres capacités, c'est-à-dire de sa propre mémoire de travail. Il s'agit d'une limite technique majeure et constitutive.

De l'autre côté, l'être humain doit sans cesse dans son monde composer avec des objets, des événements ou des relations dont l'ordinateur ne peut en prendre la mesure[4]. L'IA symbolique, ne fait que manipuler des symboles. Par conséquent, elle ne « sait » pas concrètement ce qu'est un nombre. C'est donc à l'interprète humain dont il revient la tâche de donner du sens aux symboles qui sont manipulés par la machine. Ce mode de fonctionnement n'est évidemment pas sans rappeler la critique émise par le philosophe John Searle pour défendre la thèse de l'IA faible. Ainsi, l'ordinateur ne fait que manipuler des données qui n'ont de sens que pour le programmeur. Nous aurons l'occasion d'y revenir plus en détails dans la deuxième partie de travail.

Ce qui fait néanmoins la force de cette IA, selon Andler, c'est sa capacité à utiliser des principes encodés par le programmeur[5]. Une fois que ces principes sont intégrés, l'IA peut fonctionner de manière autonome dans certaines limites, ce qui justifie qu'on puisse parler d'une forme d'« intelligence ». Toutefois, cette intelligence reste fondamentalement différente de celle des

---

[3] *Ibid.,* p. 80.

[4] *Ibid.,* p. 83.

[5] *Ibid.,* p. 41.

êtres humains : elle repose sur un traitement booléen, discontinu, qui reste toutefois extérieur aux dynamiques biochimiques et neurobiologiques de l'intelligence humaine[6].

I.1.1. Quelques exemples d'IA symbolique et leurs caractéristiques

Parmi les IA symboliques les plus emblématiques et qui ont bouleversé le champ de la philosophie de l'esprit, il nous est impossible de ne pas citer *Deep Blue*. Cette IA est à la fois citée chez Dennett dans son article *Deep Blue and Beyond de* 1997, ainsi que Searle dans le *mystère de la conscience*, car cet événement a aussi bénéficié d'une très grande couverture médiatique. Le regard que portait le grand public sur l'IA avait radicalement changé.

Il s'agit d'un programme informatique qui a réussi à battre le champion d'échecs russe Garry Kasparov en 1996. Cette « victoire artificielle » sur l'humain avait donc suscité de l'inquiétude auprès du grand public, bien que les experts informatiques ne furent pas particulièrement surpris. La réaction de ceux-ci n'est pas si étonnante que ça lorsque l'on sait que comme nous l'avons vu, cette IA symbolique avait déjà fait ses premiers pas en 1956.

C'est en tout cas de l'IA symbolique la plus connue, mais Fjelland[7] en évoque une autre qui l'est un peu moins, mais tout aussi intéressant à signaler. Il s'agit de l'ordinateur d'IBM nommé *Watson*. Cette IA avait été expressément développée dans l'optique de participer à l'émission de télévision appelée *Jeopardy*!. Le but de cette compétition était simple : les participants recevaient les réponses et à partir de celles-ci, elles devaient trouver la bonne question accolée. Pour réussir le jeu, et contrairement à *Deep Blue* qui requérait une compétence bien particulière, les participants devaient disposer de vastes connaissances sur des sujets liées à la politique, la culture, l'histoire, etc. Pour répondre aux questions, l'IA baptisée Watson communiquait à l'aide du langage naturel, mais elle n'avait pas accès à internet. En revanche, elle avait accès à quelques centaines de millions de pages d'informations. Grâce à toutes ces informations dont elle disposait, Watson a eu un avantage indéniable et cela lui a permis de gagner la compétition à plusieurs reprises.

---

[6] *Ibid.,* p. 104.

[7] Fjelland, Ragnar, « Why general artificial intelligence will not be realized », *Humanities & Social Sciences Communications*, 7(1):10, 17 juin 2020, p. 3.

Après ces plusieurs victoires, son développeur a eu l'idée de mettre à profit les compétences de cette IA au service de la médecine et le succès ne s'est pas fait attendre. Sa base de données comprenait alors toutes les informations médicales utiles en vue de poser des diagnostics et proposer des traitements adéquats telles que des articles scientifiques, les fiches des patients, une liste de médicaments, etc. Cependant, il s'est avéré que cette transition vers le domaine médical ne s'est pas passée comme prévu et que, par conséquent, le projet n'a pas pu se pérenniser. Au lieu de produire de super docteurs, ces IA ont finalement été cantonnées à exécuter des tâches répétitives.

Enfin, nos auteurs en abordent encore d'autres qui méritent aussi notre attention et dont ils se baseront pour émettre des critiques. Certaines se sont notamment inspirées du test Turing, c'est-à-dire sous une configuration où seule les capacités d'interaction étaient évaluées. Sous cette même configuration, mentionnons notamment le programme ELIZA conçu entre 1964 et 1967 au MIT par l'informaticien Joseph Weizenbaum.

Ce programme avait pour fonction de simuler un psychothérapeute et il avait été entraîné pour reconnaître et à cibler des mots-clés formulés qui lui été soumis puis il reformulait les phrases de ses interlocuteurs. Le but visé par Weizenbaum était donc de montrer comment il était possible de simuler une conversation humaine sur base de règles très simples et cette expérience lui permettait également de critiquer l'idée selon laquelle l'ordinateur pouvait réellement comprendre. Néanmoins, malgré cette facilité opérationnelle, l'expérience a montré que les participants de cette étude avaient été facilement trompés en pensant avoir affaire à un interlocuteur humain au lieu d'un programme conversationnel.

De plus, Fjelland[8] fait remarquer que les recherches de Weizenbaum, notamment reprises dans son ouvrage *Computer Power And Human Reason* publié en 1976, lui avait permis d'établir une distinction entre la puissance computationnelle et la raison humain. De fait, la puissance computationnelle peut certes s'entrainer avec des algorithmes à une vitesse impressionnante, mais elle ne pourrait pas selon lui reproduire la raison humaine, qui, au sens aristotélicien du terme, fait entrer en jeu les notions de prudence et de sagesse. Cependant, il faut garder à l'esprit que cette critique s'appliquait du temps de l'IA symbolique et que comme nous le verrons plus tard, le *deep learning* permet d'interroger quelque peu cette position.

[8] *Ibid.,* p. 2.

Comme autre exemple d'une IA conversationnelle inspirée du test de Turing, nous pouvons aussi citer le programme PARRY conçu par le psychiatre Kenneth Colby et qui simule une interaction avec un patient paranoïaque obsédé par l'idée d'être observé par la mafia. Des psychiatres devaient donner leur avis, sur la base d'un téléscripteur, pour dire s'il s'agissait d'un programme informatique ou bien d'un véritable patient humain. Comme pour le programme ELIZA, le résultat de l'étude fut édifiant et la crédulité des participants avait été démontrée.

Pour finir, citons ce que l'on appelle les « systèmes-experts ». Ce sont des systèmes tels que LUNAR, MYCIN et CYRIS. Pour aller à l'essentiel, ceux-ci ont été avant tout développés pour répondre à des besoins particuliers. Ainsi, ces systèmes se concentrent uniquement sur une expertise dans un domaine spécifique, comme le diagnostic médical ou l'analyse de données, sans avoir la nécessité de simuler des comportements humains, car ce n'est tout simplement pas ce qui leur est demandé. Ils n'ont ainsi pas vocation à tromper leur interlocuteur, contrairement à ELIZA ou à PARRY dont nous avons abordé plus haut.

I.2. L'IA connexionniste (ou *deep learning*) et en quoi est-elle si différente de l'IA symbolique ?

L'IA connexionniste, ou *deep learning* (DL), se distingue profondément de l'IA symbolique, tant dans son architecture que dans son mode de fonctionnement. Le DL est d'abord connu sous l'appellation de réseaux de neurones artificiels (*artificial neural networks*, ou ANN). Ces modèles sont constitués de réseaux de petits automates appelés « neurones formels », dont le comportement, comme le note Andler, rappelle celui des neurones biologiques[9]. Chaque automate est connecté à d'autres, et une simple modification d'état, en passant de passif (0) à actif (1), peut affecter l'ensemble du réseau auquel il est relié.

Les réseaux de neurones sont organisés en couches superposées. La couche d'entrée (*input layer*) reçoit les données brutes (images, textes, etc.), qui sont ensuite traitées par des couches cachées successives, avant d'être transmises à une couche de sortie (*output layer*) produisant le résultat final. Contrairement à l'IA symbolique, qui repose sur la manipulation de symboles et de

[9] Daniel Andler, *Intelligence artificielle, intelligence humaine : la double énigme*, Paris, Gallimard, « NRF Essais », 2023, p. 105.

règles logiques explicites, le DL fonctionne à partir de vecteurs, c'est-à-dire des listes ordonnées de nombres représentant mathématiquement les données d'entrée.

L'apprentissage du DL se fait par un mécanisme d'essai-erreur. Pour ce faire, on lui fournit un grand nombre d'exemples annotés (par exemple des images d'animaux avec leur nom), qu'il convertit alors en vecteurs numériques. Il apprend ensuite à dégager des régularités, d'abord à partir de formes simples, puis en identifiant des caractéristiques plus fines (comme la forme des oreilles ou du museau). En cas d'erreur, il ajuste ses paramètres internes pour corriger sa réponse. Plus le corpus est riche, plus l'IA devient performante et capable ensuite de reconnaître des cas inédits à partir des régularités précédemment apprises[10], que ce soit par rapport à des images ou à des corpus de textes.

Ce qui constitue la force du DL et sa différence fondamentale par rapport à l'IA symbolique, c'est sa capacité inductive : il tire des règles de l'expérience plutôt que de s'appuyer sur des principes préétablis. Comme le souligne Andler, cette méthode d'apprentissage automatique permet au DL de surpasser les capacités humaines dans le traitement massif de données, en termes de rapidité et de volume de données;

Andler remarque que, de ce point de vue, le connexionnisme se rapproche davantage de l'intelligence humaine, dans la mesure où il repose également sur l'induction par l'expérience. À l'inverse, l'IA symbolique s'apparente davantage à une approche rationaliste : elle repose sur la logique, sur des règles explicites et sur des représentations symboliques définies à l'avance[11]. Elle ne nécessite donc pas d'apprentissage progressif, mais dépend d'un programmeur humain qui encode en amont ce qu'elle doit faire.

Andler associe le connexionnisme (ou le *deep learning*) à la modernité, en rupture avec ce qu'il appelle l'ère classique de l'IA, dite symbolique. Selon lui, le règne du *deep learning* s'étend de 2012 à 2017, voire au-delà. Cette technologie apporte avec elle un lot de nouveautés impressionnantes : elle est désormais capable d'effectuer de la reconnaissance vocale, de poser un

[10] *Ibid.*, p. 107-108.

[11] *Ibid.*, p. 121.

diagnostic médical précis, d'identifier des personnes ou encore de piloter automatiquement des véhicules pour prévenir des collisions.

Autant de compétences qui présentent un intérêt croissant pour notre vie quotidienne et pour un large éventail de métiers. Andler estime ainsi que le DL a ravivé l'intérêt général pour l'IA, mais ce n'est qu'à partir des années 2010 qu'il connaît, selon lui, un véritable triomphe, inaugurant ce qu'il appelle une troisième époque, dont nous aborderons plus loin les caractéristiques.

I.2.1. Quelques exemples d'IA connexionniste et ses caractéristiques

En 2016, l'apparition de l'algorithme AlphaGo de la société DeepMind sur les devants de la scène médiatique a constitué un tournant majeur en déjouant les prévisions selon lesquelles une IA ne pourrait jamais intégrer les règles si particulières de ce type de jeu[12]. AlphaGo est en quelque sorte le *Deep Blue* de l'ère connexionniste, à la différence fondamentale que ce nouveau programme possède une architecture de réseaux neuronaux profonds combinée à un système d'apprentissage par renforcement. Cette architecture si particulière pour l'époque porte le nom de *deep reinforcement learning*. Ainsi, la machine était programmée pour évaluer plusieurs positions sur le plateau de jeu et elle était capable de choisir à un moment donné quelle était la meilleure stratégie à adopter.

Là où AlphaGo se démarquait particulièrement de, c'est qu'on ne l'avait pas programmée avec des règles explicites à suivre. Quand une machine n'est programmée qu'avec des règles explicites, il est possible d'avoir accès à sa boîte noire, contrairement à l'IA connexionniste où l'on ne peut pas voir ce qu'il se passe à l'intérieur, ni les paramètres qui ont été actionnés[13].

Qui plus est, toujours selon Fjelland et contrairement à l'IA connexionniste, AlphaGo a démontré sa capacité à gérer la connaissance tacite — l'un des arguments phares d'Hubert Dreyfus et Michael Polanyi contre l'IA forte. Ces auteurs expliquaient que dans notre vie quotidienne, nous appliquons souvent des règles implicites : par exemple, savoir faire du vélo ne suppose pas de pouvoir expliquer les lois physiques en jeu. Même si on les énonçait, elles n'aideraient pas

[12] *Ibid.,* p. 141.

[13] Fjelland, Ragnar, « Why general artificial intelligence will not be realized », *Humanities & Social Sciences Communications*, 7(1):10, 17 juin 2020, p. 4-5.

davantage à l'apprentissage, car la compétence repose sur la pratique et non sur une explication explicite.

Néanmoins, malgré cette prouesse impressionnante, nous verrons en dernière partie de notre travail qu'il existe encore des dimensions auxquelles la machine n'a pas encore accès et qui sont pourtant constitutives de l'expérience humaine en nous appuyant entre autres sur les travaux de Hubert Dreyfus.

De leur côté, les modèles GPT ont révolutionné l'interaction entre l'humain et la machine en rendant possible la génération de texte fluide sur base de *prompts* qui lui sont soumis par un utilisateur humain. Les auteurs que nous allons abordé pour tenter de répondre à la question de notre enquête se sont surtout exprimés au sujet de la troisième version de ChatGPT, même si Chalmers et Andler abordent brièvement la quatrième et plus récente version.

Tout deux reconnaissent que ce modèle d'IA est devenu capable d'acquérir une variété de compétences telles que la rédaction d'essais, de poèmes, de traduction ou plus largement des capacités de dialogues bluffantes. La particularité de ces modèles repose essentiellement sur le fait qu'ils sont entraînés à partir d'immenses corpus textuels (majoritairement issus d'internet). Comme nous le voyons et à la différence d'AlphaGo, le champ d'application est beaucoup plus vaste.

I.2.2. Le DL n'est pas sans défauts majeurs

Malgré ces capacités impressionnantes, ces systèmes souffrent encore d'un manque de flexibilité qui les empêche de s'adapter à un changement dans leur environnement. Des changements mineurs peuvent particulièrement l'affecter et modifier son comportement. Or, les problématiques du monde dans lequel nous vivons se produisent dans un monde qui ne cesse de changer.

Comme nous venons de le voir, bien que ce type d'IA a relancé un nouvel intérêt après une mise en suspens de son précédent modèle symbolique, Andler signale que le modèle connexionniste présente plusieurs faiblesses, dont l'une des plus notables est sa vulnérabilité face aux attaques dites

« adversariales »[14]. Il s'agit de techniques consistant à tromper un modèle d'apprentissage automatique en lui soumettant des entrées délibérément modifiées et souvent de manière imperceptible pour l'œil humain afin de le pousser à commettre des erreurs.

Par exemple, une altération minime et invisible à l'œil nu sur une image peut suffire à perturber l'ensemble du système et à provoquer une mauvaise classification. De cette façon, l'IA peut en venir à confondre des objets ou des entités qu'elle identifiait auparavant correctement, alors qu'un observateur humain, lui, ne tomberait pas dans le panneau. Ce simple exemple permet à Andler de relativiser l'image d'Épinal souvent associée à ces technologies perçues comme surpuissantes.

Andler signale également deux autres limites majeures du *deep learning* : le *data shift* et l'*underspecification*[15]. Le premier désigne le décalage entre les données utilisées pour entraîner l'IA et celles qu'elle rencontre dans le monde réel, une fois sortie du cadre contrôlé du laboratoire. Le second, quant à lui, souligne l'imprécision du modèle : il est souvent difficile de prévoir comment l'IA se comportera face à des données nouvelles, car les règles qu'elle applique ne sont pas explicitement connues ni entièrement maîtrisées.

Malgré ces obstacles évidents, Andler salue les prouesses technologiques accomplies, notamment par *AlphaGo* de la société *DeepMind* : cet algorithme n'a même plus besoin que l'on lui enseigne explicitement les règles d'un jeu. Il apprend seul en jouant contre lui-même, sans intervention humaine. Toutefois, une autre limite persiste : l'hyperspécialisation[16]. Un joueur humain peut facilement passer du jeu d'échecs au jeu de go sans repartir de zéro, alors que l'IA, elle, doit réapprendre intégralement chaque domaine. Une fois focalisée sur une nouvelle tâche, elle perd souvent la compétence acquise précédemment. Par ailleurs, si elle est confrontée à une situation en dehors de son périmètre d'expertise, elle est tout simplement incapable de répondre.

---

[14] Daniel Andler, *Intelligence artificielle, intelligence humaine : la double énigme*, Paris, Gallimard, « NRF Essais », 2023, p. 132.

[15] *Ibid.,* p. 136-137.

[16] *Ibid.*, p. 145. Cette limite se manifestait déjà du temps de l'IA symbolique avec ce que l'on appelle les « systèmes-experts », c'est-à-dire des IA qui étaient uniquement douées dans un domaine spécifique. Dans son premier article de 1984, Dennett remarquait déjà que celles-ci donnaient une fausse impression de performance.

Face à de tels dispositifs, même des juges humains sont parfois incapables de distinguer la production d'une IA de celle d'un humain, preuve de leur puissance. Mais cette apparence de performance masque des défauts structurels : il arrive fréquemment que ces IA produisent des réponses erronées, voire absurdes, car elles ne savent pas de quoi elles parlent[17]. Ce phénomène s'apparente à des hallucinations où ces IA génératives produisent des réponses fausses, inventées de toutes pièces ou incohérentes. C'est à cette occasion qu'Andler souligne ainsi leur cécité sémantique, en insistant sur le fait qu'elles manipulent des formes sans en comprendre le sens. Cette problématique de la cécité sémantique occupera une section de la dernière partie de notre travail.

Il évoque également la critique d'Emily Bender, célèbre linguiste américaine, qui a qualifié ces modèles de « perroquets statistiques », soulignant leur tendance à imiter le langage humain sans réelle compréhension. Cette critique va de pair avec l'anthropomorphisme naïf dont font preuve certains utilisateurs, qui attribuent à ces IA des intentions qu'elles ne possèdent pas. Alors que dans les faits, ces modèles sont avant tout conçus pour prédire des suites probables de mots en fonctions des données sur lesquelles ils se sont entraînés.

Cette tendance que nous avons à projeter de l'intelligence humain et à prêter une grande performance à l'endroit des machines ne date pas d'hier, car ce phénomène existait déjà du temps de l'IA symbolique. En effet, cette question avait déjà été abordée par Dennett dans son article de 1984 où il soulignait déjà ce qu'il appelle « qualité de façade »[18] aux variantes des tests de Turing comme ELIZA, PARRY ou même ceux qui furent mis à l'épreuve durant les différentes éditions du prix Loebner.

Pour autant, Andler invite à ne pas rejeter ces IA au seul motif de leurs défauts. L'intelligence artificielle, comme l'intelligence humaine, commet des erreurs. Malgré leur illusion de compréhension, ces systèmes reposent sur un savoir-faire technique d'une extrême complexité, fruit du travail de nombreux ingénieurs. Il ne serait pas pertinent de les disqualifier d'emblée : nous risquerions de passer à côté d'une révolution techno-scientifique majeure. Leurs capacités réelles demeurent difficiles à cerner, tant leur évolution est rapide. C'est d'ailleurs l'un des enjeux de

---

[17] *Ibid.,* p. 165.

[18] Dennett, Daniel C., « Can Machines Think? », in *Brainchildren: Essays on Designing Minds*, Cambridge (Mass.), MIT Press, 1998, p. 16.

l'ouvrage d'Andler : apprendre à mieux comprendre et utiliser ces systèmes, en connaissant à la fois leurs forces et leurs limites, pour éviter de leur attribuer des compétences qu'ils n'ont pas.

I.3. Troisième ère de l'IA et conclusion

Cette troisième et dernière ère se manifeste sur deux approches, en premier lieu, il y a celle qui se caractérise par l'émergence du concept d'intelligence augmentée, qui bouleverse déjà les champs éthiques, juridiques et sociaux[19]. On associe cette phase à des avancées spectaculaires dans des domaines tels que la traduction automatique, le diagnostic médical, la prévention de la criminalité ou encore l'évaluation des risques de récidive.

De plus, ce type d'IA se distingue des deux autres par rapport à la philosophie de leur entreprise dans lesquelles elles sont imaginées, pensées et construites[20]. Cette intelligence augmentée était à la base cantonnée dans les laboratoires militaires ou universitaires et son usage était exclusivement réservé à des professionnels qui disposaient de connaissances scientifiques exceptionnelles.

L'un des buts affichés par son créateur Douglas Engelbart était d' « augmenter » l'esprit humain à l'aide des ressources informatiques dont on dispose[21]. Pour donner un exemple de ces ambitions, il est notamment question d'élaborer un partenariat avec les sciences cognitives et la robotique afin de proposer au public des augmentations qui prennent la forme d'une prothèse qui irait se greffer sur le corps humain.

Parallèlement, cette troisième ère de l'IA est aussi représentée par l'AGI (*Artificial General Intelligence*, ou intelligence artificielle générale). Celle-ci caresse le rêve de dépasser les limites inhérentes à l'IA actuelle, qui n'excelle que dans des tâches très précises. Elle ambitionne également d'atteindre des performances proches de l'intelligence humaine en termes de capacités de raisonnement, de polyvalence et d'adaptation. C'est, en réalité, la définition actuelle associée à la notion d'IA forte. Toutefois, il faut garder à l'esprit que cette IA reste avant tout une notion hypothétique : elle demeure un idéal théorique et n'existe pas encore aujourd'hui.

---

[19] *Ibid.*, p. 146.

[20] *Ibid.*, p. 189.

[21] *Ibid.*, p. 191.

La preuve en est qu'une longue étude publiée en 2023[22] a conclu que ChatGPT-4 « manifeste des étincelles d'intelligence artificielle générale » de par sa polyvalence et sa facilité à passer des tests professionnels variés. En somme, l'étude reconnait que la frontière entre IA spécialisée et IA générale commence devenir plus floue, mais qu'il demeure toujours certaines limites comme l'autonomie, une conscience de soi ou le sens commun.

La conclusion de cette étude rejoint celle défendue par David Chalmers en 2023, dans son article consacré à la question de savoir si les *Large Language Models* (LLM) sont conscients. Il reconnaît que ces modèles manifestent déjà une certaine forme d'intelligence générale par leur capacité à accomplir des tâches variées, comme la poésie, le codage, le jeu, l'argumentation, etc. Il constate ainsi une avancée évidente par rapport aux systèmes-experts, qui étaient spécialisés dans un domaine restreint. Néanmoins, cette diversité dans les tâches ne constitue pas une preuve définitive d'une conscience artificielle, même si elle peut sembler s'en approcher.

En résumé, Daniel Andler distingue trois grandes phases dans l'évolution de l'intelligence artificielle : l'ère symbolique classique, marquée par une IA fondée sur des règles explicites et une logique formelle, l'essor du *deep learning*, qui inaugure une nouvelle approche fondée sur l'apprentissage par réseaux de neurones, et enfin, l'émergence d'une troisième époque, celle de l'intelligence augmentée et de l'AGI, où l'IA commence à s'immiscer progressivement dans des sphères sensibles comme la médecine, la justice ou la traduction. Notre enquête ne se focalisera pas sur cette dernière ère, car les contours de celle-ci sont encore difficiles à délimiter étant donné son stade de développement.

Ainsi, nous observons que cette évolution montre un glissement progressif : on passe d'une intelligence artificielle qui exécute simplement des instructions explicites, à une IA qui apprend par induction à partir de masses de données, jusqu'à une IA qui co-évolue avec nos structures sociales et cognitives. Pourtant, nous le verrons plus tard, ce progrès fulgurant soulève aussi un certain nombre de limites techniques.

---

[22] Bubeck, Sébastien, et al., « Sparks of Artificial General Intelligence: Early Experiments with GPT-4 », *arXiv.org*, mars 2023. DOI : 10.48550/arXiv.2303.12712.

PARTIE 2 : Le débat philosophique entre ceux qui défendent l'IA forte et l'IA faible et leurs arguments

Le débat qui oppose les défenseurs de l'IA forte et de l'IA faible prend sa source dans la discipline de la philosophie de l'esprit, et plus particulièrement par rapport à l'épineuse question de la conscience. Lorsque nous nous interrogeons pour savoir ce qu'est penser, ce qu'est la conscience ou l'intentionnalité, nous pouvons également poser ces questions dans le domaine de l'intelligence artificielle.

En ce qui concerne la question centrale de notre enquête, de nombreux philosophes ont, dès 1950, pris position sur le fait de savoir si les machines peuvent penser. Comme nous l'avons indiqué, beaucoup d'entre eux relèvent de la philosophie de l'esprit et apportent chacun un éclairage spécifique. Cette section vise à mettre en évidence les arguments principaux déployés par ces penseurs ainsi que leurs démonstrations.

Pour plus de clarté, il convient d'abord de préciser ce que l'on entend par « IA forte » et « IA faible » :

Par IA forte, nous entendons une IA qui sera à même un jour de surpasser l'intelligence humaine. C'est-à-dire que cette IA vise à devenir une AGI, soit une IA générale qui serait à même un jour de reproduire la polyvalence et la compréhension humaine. Elle ne se limiterait donc plus à des tâches limitées et restreintes, mais elle démontrerait une flexibilité adaptative proche ou supérieure à celle de l'intelligence humaine. Nous associons également cette thèse défend également le fait que les machines sont capables de manifester de la pensée. En outre, cette AGI serait également apte à transférer des compétences d'un domaine à un autre, de répondre et s'adapter à des situations nouvelles, etc. Elle ne se cantonnerait donc plus à des domaines limités et restreints, mais elle se montrerait également capable de résoudre des problèmes en mobilisant plusieurs systèmes d'intelligence artificielle de manière autonome.

L'IA faible, quant à elle, soutient que l'IA ne pourra jamais résoudre des problèmes typiquement humains, car elle présente des limites fondamentales constatées dès sa création et qui perdurent aujourd'hui. Cette position est notamment illustrée par les travaux d'Hubert Dreyfus dans *What Computers Still Can't Do* (ouvrage sur lequel nous reviendrons en troisième partie). Dreyfus

affirmait déjà que l'IA ne comprend pas, qu'elle ne produit pas de pensée véritable et qu'elle ne manifeste aucune conscience. Sa réflexion, d'inspiration phénoménologique comme le rappelle Andler, avait particulièrement marqué l'époque de l'IA symbolique. Il est donc pertinent de s'interroger sur l'actualité de cette position à l'ère de l'IA connexionniste et des technologies actuelles.

La thèse de l'IA faible, quant à elle, renvoie particulièrement à Searle, qui est un des philosophes les plus emblématiques de ce courant, car, comme nous le verrons, il maintient que la structure elle-même de la machine ne lui permettra jamais d'expérimenter un vécu subjectif ou de mobiliser une véritable compréhension des symboles et mots qu'elle utilise. Cette compréhension de ces données n'ont de sens que pour l'observateur ou pour celui qui encode les données. C'est une position forte qu'il maintiendra, comme nous le verrons, jusqu'à son dernier article à ce sujet publié en 2015.

Tout au long de cette partie du travail, nous verrons aussi que Dennett présente à la fois des arguments en faveur de l'IA forte, mais aussi en faveur de l'IA faible et qu'il est moins évident de le cantonner à une seule posture philosophique en particulier. Sa position nuancée au sujet de l'IA découle à la fois du fait qu'il s'exprime par rapport à des technologies à l'époque où il écrit ses articles, mais aussi car il estime qu'il n'est pas déraisonnable de penser que les machines deviennent un jour conscientes. Le développement des différents arguments soulevés montrera comment sa conception de la conscience l'amène à envisager cette possibilité.

De surcroît, nous examinerons également la position de Chalmers au sujet de l'IA. Bien qu'il soit traditionnellement associé à la défense de l'IA forte, nous verrons en quoi l'arrivée des LLM actuels ne lui a pas encore permis de statuer sur l'émergence d'un esprit artificiel tel qu'il l'a théorisé. Selon lui, ils leur manquent encore certaines composantes essentielles pour atteindre cet objectif qu'il estime envisageable dans les années à venir.

Enfin, il va de soi que chacun des arguments explicités fera l'objet d'une relecture critique à partir du moment où des contre-arguments philosophiques pourront être mobilisés ou être en lien avec la technologie actuelle dont nous disposons et en particulier celle de l'IA connexionniste.

II.1. Arguments en faveur de l'IA forte

II.1.1. Turing et la puissance de calcul

Aux prémices de l'émergence des ordinateurs, la pensée dominante considérait que ces machines n'étaient pas dotées d'intelligence. C'est dans ce contexte que Alan Turing publia un article sorti en 1950[23], intitulé *Computing machinery and intelligence* où il propose d'imaginer une expérience qui tentera de montrer si une machine peut être considérée comme intelligente ou non.

En préambule de son article, Turing précise tout de même que pour répondre correctement à la question, il faudrait d'abord s'attarder sur les termes « machines » et « penser »[24]. Toutefois, compte tenu du fait qu'ils sont utilisés couramment dans notre langage, la signification que l'on leur en donne pourrait nous orienter vers de fausses pistes.

C'est sur cette base qu'il imagine une nouvelle approche expérimentale : un test fictif, connu aujourd'hui sous le nom de test de Turing. Son principe est le suivant : plusieurs personnes échangent par écrit via des interfaces, tout en étant dans des lieux séparés. L'un des interlocuteurs est en réalité un logiciel conversationnel. L'objectif, pour l'humain, est de distinguer son interlocuteur humain de l'interlocuteur artificiel.

L'ordinateur doit donc tromper l'interlocuteur en se faisant passer pour un être humain, grâce à sa capacité à formuler des réponses adaptées à divers contextes conversationnels. La machine réussit le test si l'interlocuteur humain n'est pas capable de la différencier d'un autre humain. Plus la machine imite l'humain de manière convaincante, plus elle est considérée comme « pensante ».

C'est à partir de sa proposition qu'il émet la thèse innovante que la puissance de calcul amènerait forcément l'humanité à se questionner sur la conception traditionnelle de l'intelligence. À partir du moment où une machine parvient à tromper l'humain, il estime que l'on peut supposer qu'elle est intelligente.

---

[23] Turing, Alan, « Computing Machinery and Intelligence », *Mind*, vol. 49, 1950, p. 433-460.

[24] *Ibid*., p. 433.

Ainsi, il ne fournit ni une définition précise de la pensée, ni une preuve irréfutable que les machines « pensent » au sens humain du terme. Selon lui, les notions de « machine » et de « penser » sont caractérisés par une ambiguïté inhérente à leur usage courant, ce qui rend nécessaire de dépasser les significations traditionnelles qu'on leur accorde. Il privilégie alors à travers son test fictif une approche à la fois mesurable et interactionnelle au lieu d'une discussion qui prendrait la forme d'un débat qui tournerait seulement autour de définitions.

De plus, son article est aussi une occasion pour lui d'anticiper un ensemble d'objections[25] que l'on oppose habituellement à l'affirmation que les machines pensent. Parmi celles que nous retiendrons et qui nous intéressent, il y a la 4ème, la 5ème et la 6ème qui, comme nous le verrons, sont celles à partir desquels Searle ou Andler apporteront des arguments qui vont dans les sens de la thèse de l'IA faible.

- L'objection de la conscience. Selon les opposants à la thèse de la machine pensante, la pensée suppose une conscience vécue à travers des sentiments, des intentions ainsi qu'une compréhension authentique de ses actes. Par exemple, dans le cas d'une poésie, elle est composée par une personne qui a ressenti des émotions et non par quelqu'un qui écrit des mots vides de sens. Pour qu'une machine soit considérée comme consciente, il est attendu qu'elle comprenne ce qu'elle fait et qu'elle soit également capable de juger et d'expliquer ce qu'elle a produit. Cependant, Turing rejette cette exigence, car selon lui, les humains eux-mêmes ne sont pas capables de ressentir de l'intérieur ce que ça fait d'être dans la peau d'une autre personne. Pour lui, ce qui est primordial pour évaluer la pensée, c'est d'observer ce que la machine est capable de faire. C'est-à-dire qu'il place le curseur de son attention sur ce qui émane de la machine d'un point interactionnel plutôt que sur ce qu'il se passe à l'intérieur de celle-ci. En d'autres termes, nous pouvons dire qu'il élude la question de la subjectivité vécue en se focalisant exclusivement sur les sorties et non sur les états internes.

- L'incapacité d'acquérir certaines qualités humaines. Il serait impossible de la programmer pour faire en sorte qu'elle acquiert des capacités proprement humaines telles que la gentillesse, l'humour, la distinction du bien et du mal, etc. Turing pense plutôt que ces limitations pourraient être dépassées dans le futur si l'on se penche sur l'augmentation de la taille de la mémoire des

[25] *Ibid*., p. 441-454.

machines. Affirmer cette position serait selon lui faire preuve d'anthropocentrisme en fondant uniquement son opinion sur sa propre expérience, alors que nous pouvons pas encore nous exprimer sur ce qu'il adviendra de l'évolution de ces machines. De même, il remarque qu'une machine est capable de produire des erreurs comme un humain dans le cas de figure où on la programme pour en faire. Il ajoute également qu'avec un programme adéquat, la machine pourrait être capable d'étudier son propre programme, notamment via un programme débogueur où les lignes de codes sont passées en revue afin d'y déceler un éventuel bug. En somme, il estime que les manquements que l'on reproche à la machine sont avant tout liés à sa capacité de mémoire. Autrement dit, si l'on trouvait un moyen d'augmenter celle-ci, nous serions amenés à reconsidérer l'argument de leur incapacité d'acquérir certaines qualités humaines.

- L'objection Lady Lovelace. Ada Lovelace est considérée comme une pionnière dans le champ de l'informatique lorsqu'elle a conçu le premier programme informatique. Selon elle, les machines ne seraient capables de produire que ce que nous leur demandons de faire. Celles-ci se contentent simplement d'exécuter des instructions qu'on leur donne et elles ne peuvent rien inventer par elles-mêmes. La machine est donc fondamentalement déterminée : elle ne peut surprendre son concepteur. La véritable intelligence résiderait plutôt dans celle du programmateur grâce à laquelle il a conçu le programme. Or, Turing remarque que l'absence d'initiative et d'originalité peut aussi être reproché chez les humains. En outre, il affirme avoir été souvent surpris par le résultat parfois inattendu de ses recherches en la matière.

De notre point de vue, force est de constater que son argument qui consiste à dire que la puissance du calcul allait bouleverser nos conceptions sur l'esprit humain fait écho à notre propre actualité. On le sait, la technologie informatique a fortement évolué depuis et Turing était conscient que les limitations technologiques de son temps allaient certainement être dépassées.

En ce qui concerne la capacité à acquérir certaines qualités humains, il nous semble pertinent de discuter sur ce point. De nos jours, les LLM démontrent effectivement une capacité à dialoguer ou à produire des textes fluides de manière convaincante. Nous pouvons également témoigner qu'elles peuvent produire une certaine créativité simulée à travers la génération d'images, de musique, de poèmes, etc. De même, nous savons que ces IA peuvent faire preuve d'empathie et de bienveillance dans leurs réponses, car elles ont été conçues pour répondre de cette façon à travers la manière dont elles ont été programmées en lien avec leur alignement des valeurs.

En un sens, ce que disait Turing en 1950 est proche de notre réalité. Cependant, nous pourrions nous interroger sur la manière dont elles héritent justement de ces qualités humaines. Il est très probable que les machines actuelles acquièrent ces qualités en s'entraînant à partir de fiches théoriques ainsi qu'à partir de réponses issues de d'une multitude de situations particulières.

Ensuite, elles sont recadrées par des intervenants humains ou des évaluateurs qui sélectionnent le meilleur comportement à adopter face à telle requête. Ainsi, le comportement positif se voit récompensé à mesure que ces IA emploient un ton empathique et positif. Peut-on affirmer que l'IA elle-même démontre une disposition morale ? À notre sens, il s'agit plutôt d'un choix d'ingénierie effectué sur base de valeurs considérées comme bienveillantes du point de vue des évaluateurs et qui font naturellement écho aux nôtres.

De même, nous pouvons interroger sur les capacités des IA à faire preuve d'humour, qui est pourtant une aptitude humaine, car elle doit en toute logique se baser sur des références humaines afin de générer un effet. De la même façon que les qualités humaines, nous pouvons penser que le modèle a en premier lieu assimilé de grandes quantités de blagues, scripts ou sketches. Après cela, sur base de ce grand regroupement de données, l'IA repère des régularités formelles que l'on retrouve habituellement dans cette discipline telles que le quiproquo, l'incongruité, etc. Mais de nouveau, est-ce que cela veut dire que ces modèles manifestent une intention comique ? Nous pensons au contraire qu'il s'agit essentiellement de corrélations statistiques et d'estimations qui mettent en avant des tournures perçues comme drôle aux yeux de l'interlocuteur.

N'oublions pas que les qualités humaines sont aussi en lien avec nos propres affects. Que peut-on alors dire de la phénoménologie des affects en lien avec les IA ? Sont-elles par exemple capables de reconnaître des émotions humaines ? Andler pointe le fait que les résultats actuels en la matière sont plutôt mitigés[26]. Les machines pourraient néanmoins acquérir une sorte de « théorie des affects » sous forme de fiches, mais est-ce que cela les aideraient à résoudre des problèmes pertinentes issus de contextes réels ? De nouveau, Andler écrit que l'état de la recherche reste encore très préliminaire et il affirme que malgré des décennies de spéculation, l'intégration efficace d'un système affectif à l'IA n'a toujours pas débouché sur une percée significative[27].

---

[26] Daniel Andler, *Intelligence artificielle, intelligence humaine : la double énigme*, Paris, Gallimard, « NRF Essais », 2023, p. 238.

[27] *Ibid.,* p. 241.

De son côté, Searle, dans son article intitulé « *What computers can't know* » paru en 2014[28], affirme avec force que, contrairement à Turing, les machines ne sont pas dotées de conscience. Pour comprendre, penser ou apprendre une langue implique une expérience subjective, toujours absente chez les machines. Le test de Turing, en ce sens, ne permet pas de détecter une véritable intelligence : il ne teste que le comportement externe, non une réalité mentale interne. Ceci nous fait dire que ces IA ne présentent pas de la subjectivité, ni de vécu émotionnel. Elles simulent en réalité des actes de langage qui sont conformes à nos propres attentes.

En somme, du point de vue de Turing, celui-ci réduit la notion de pensée à une notion fonctionnaliste. La pensée consiste donc pour lui de démontrer la capacité à produire un flux cohérent de propositions à partir d'une interaction conversationnelle donnée. Autrement dit, la machine produit des conclusions logiques ou probables à partir d'un ensemble d'énoncés. Et à l'aune de notre technologie actuelle, les machines deviennent effectivement de plus en plus crédibles et adaptatives. Pour reprendre le titre de film basé sur la vie Turing, les machines imitent de mieux en mieux, mais cela ne veut pas dire pour autant qu'elles ont un vécu subjectif ou qu'elles ont leurs intentions propres.

Comme nous venons de le voir, l'article et le test proposés par Turing à l'époque ont par la suite fait l'objet de nombreuses discussions. Ce fut notamment le cas avec Daniel Dennett pour qui la question de savoir si les machines pensent occupe une grande partie de son propos dans ses articles de 1984 et de 1997. Reconnaît-il des qualités au test de Turing ?

Que ce soit dans son article de 1984 ou encore celui de 1997, Dennett reconnaît que le test de Turing, tel qu'il a été conçu, est suffisamment solide pour évaluer la pensée, et il met au défi quiconque de faire mieux[29]. D'une part, parce que la configuration elle-même du fameux test ressemble à celle des auditions réalisées à l'aveugle dans le cadre du recrutement au sein d'un orchestre, ce qui permet d'éviter l'apparition de biais liés aux caractéristiques physiques afin de privilégier uniquement les compétences propres des candidats.

---

[28] Searle, John R., « What Your Computer Can't Know », *The New York Review of Books*, vol. 61, no 15, 9 octobre 2014

[29] Dennett, Daniel C., « Can Machines Think? », in *Brainchildren: Essays on Designing Minds*, Cambridge (Mass.), MIT Press, 1998, p. 5.

D'autre part, Dennett estime Turing a basé son test sur l'idée que la capacité de tenir une conversation est un signe d'intelligence, ce qui renvoie à une intuition déjà présente chez Descartes[30], lequel affirmait en que ce qui distingue la machine d'un homme est précisément la capacité à interagir par le langage. En un sens, Dennett reconnaît que le test de l'informaticien britannique repose sur une intuition philosophique cartésienne pertinente et que la configuration par écrans interposés permet de limiter les biais subjectifs.

Le philosophe reprendra d'ailleurs cet argument dans son article de 1997 où il précise que Descartes posait déjà la question de savoir comment distinguer un homme d'une machine[31]. Il avançait deux critères : d'une part, une machine ne pourrait pas produire des discours variés et adaptés à toutes les situations ; d'autre part, même si elle imitait certains comportements, elle n'aurait pas la raison pour agir de manière appropriée dans tous les contextes. Les machines seraient donc limitées par leur propre conception dans le sens où il leur manquerait une adaptation généralisée. Dennett réitérera son propos dans son article de 2019[32] en attribuant à Turing le mérite d'avoir théoriser qu'une intelligence consciente se manifeste comme une stratégie pour imiter l'humain.

II.1.2. L'expérience de pensée ou les pompes à intuition

Vers la fin de son article de 1950, Turing jette les bases d'une potentielle « machine apprenante »[33]. Pour ce faire, il repart de la sixième objection qu'il a formulée, celle qui consiste à dire que la machine ne peut pas faire autre chose que ce que nous lui disons de faire. Ensuite, pour dépasser cette objection, il va proposer l'idée de concevoir une machine apprenante qui fonctionnerait d'une façon presque similaire à l'éducation d'un enfant en y incluant un système d'apprentissage basé sur des récompenses et des punitions.

---

[30] Dennett, Daniel C., « Can Machines Think? Deep Blue and Beyond », *Studium Generale Maastricht*, 1997, p. 6.

[31] *Ibid.,* p. 12.

[32] Dennett, Daniel C., « What Can We Do? », in J. Brockman (dir.), *Possible Minds: Twenty-Five Ways of Looking at AI*, New York, Penguin Press, 2019.

[33] Turing, Alan, « Computing Machinery and Intelligence », *Mind*, vol. 49, 1950, p. 449.

Malgré l'ingéniosité de cette expérience de pensée, Turing reconnaît quand même que les avancées en la matière seront lentes et produiront certainement de nombreux échecs qui seront liés tant aux ressources computationnelles que pédagogiques. Mais cette lenteur ne contredit pas pour autant la possibilité d'une machine apprenante.

Ce passage illustre le fait que Turing a anticipé l'émergence d'une machine capable d'apprendre par elle-même en passant d'un modèle symbolique et rigide typique à son époque à un modèle à la fois évolutif et adaptatif. Il ne faudrait donc plus programmer la machine de façon explicite en « mettant en conserve » toute une série de règles à appliquer comme nous l'avons vu avec Andler, bien que celle-ci acquière des comportements par induction. Cette configuration, nous la retrouvons à l'oeuvre du côté de l'IA connexionniste qui ne mémorise pas des règles à l'avance, mais qui cerne et apprend des régularités ou tendances à partir d'un large stock de données.

Cette anticipation n'est pas non plus sans rappeler celle qu'a formulée Dennett dans son article de 1997 lorsqu'il a imaginé la construction d'un esprit artificiel nommé WUNDERKIND[34]. Cette machine irait à l'école, suivrait les cours, et acquerrait progressivement des compétences intellectuelles. Il s'agit ici d'une expérience de pensée ou d'une "pompe à intuition", comme il les appelle. Cette expérience est destinée à explorer les limites supposées de l'intelligence artificielle et de manière plus générale, à sortir des sentiers battus de la pensée, à l'instar de ce que Turing proposait dans son article fondateur de 1950.

Dennett se demande donc : si une machine suit exactement le même parcours éducatif et social qu'un enfant humain, pourrait-elle atteindre le même niveau d'intuition et d'intelligence ? Cette hypothèse, si elle était testée, permettrait alors de vérifier empiriquement les affirmations de Lucas et Penrose sur lesquelles nous reviendrons plus en détails dans la partie consacrée aux arguments en faveur de l'IA faible.

L'idée de Dennett est donc de rejeter l'affirmation selon laquelle il existerait une capacité humaine transcendantale impossible à imiter par une machine sophistiquée. D'une certaine façon, à travers son expérience de pensée, il déconstruit et retire à l'humain toute compétence mystérieuse et inaccessible aux machines.

---

[34] Dennett, Daniel C., « Can Machines Think? Deep Blue and Beyond », *Studium Generale Maastricht*, 1997, p. 29.

Cette intuition de Dennett rejoint aussi au *post-scriptum* de son article de 1984[35] où il estime qu'il sera possible un jour de doter la machine d'une *self-conciousness* si l'on part du principe que celle-ci se caractérise par le fait de se distinguer soi par rapport aux autres. Pour donner du poids à son intuition, il prend l'exemple du homard : même quand il est affamé, il ne se mange pas lui-même. Cela prouverait qu'il possède une forme de conscience de soi distincte de celle d'autrui.

Non dénuée de pertinence, cette vision de la conscience ne met toutefois pas en avant que, comme le dit Andler, nous ne disposons pas à l'heure actuelle d'une idée suffisamment claire de ce qu'est justement la conscience humaine[36]. Il mentionne qu'il existe bel et bien des études sur la conscience du côté des sciences cognitives, mais qu'il n'existe pas à proprement parler d'une théorie sur la conscience qui fasse l'objet d'un consensus. Il note qu'il existe bien sûr des tentatives pour doter l'IA d'une conscience. Les idées et les propositions ne manquent pas, mais cela relève toutefois d'un domaine fort spéculatif.

D'une certaine façon, l'ordinateur possède déjà une capacité quelque peu similaire à travers son système de *self-monitoring* et de *self-watching*. Le premier système consiste en effet en la capacité de surveiller son propre état interne ainsi que ses performances en temps réel. Il évite ainsi l'état de surchauffe et contrôle l'état et ainsi que l'usage de sa propre mémoire qui est évidemment limitée. Le second système, quant à lui, confère à l'ordinateur la capacité de s'auto-observer. Cela dit, c'est une vision de conscience pour le moins minimaliste, car elle ne prend pas en compte les interactions avec le monde extérieur.

Néanmoins, Dennett nous fait tout de même remarquer que pour qu'une entité soit pensante, il est nécessaire que celle-ci dispose d'organes sensoriels tels que des yeux, des oreilles et des mains, sans oublier des interactions au sein du monde réel[37]. Pareillement, sans interactions avec le monde réel, aucune entité ne pourrait manifester de croyances, de désirs ou d'intentions.

---

[35] Dennett, Daniel C., « Can Machines Think? », in *Brainchildren: Essays on Designing Minds*, Cambridge (Mass.), MIT Press, 1998, p. 23-34.

[36] Daniel Andler, *Intelligence artificielle, intelligence humaine : la double énigme*, Paris, Gallimard, « NRF Essais », 2023, p. 221.

[37] Dennett, Daniel C., « Can Machines Think? », in *Brainchildren: Essays on Designing Minds*, Cambridge (Mass.), MIT Press, 1998, p. 21.

Aujourd'hui, nous disposons de véhicules autonomes capables d'interagir avec leur environnement : reconnaissance du type ou de la forme de la route, détection d'autres véhicules, ajustement de la vitesse, anticipation des obstacles. Ils utilisent pour cela des capteurs, des programmes de traitement d'image et des réseaux neuronaux. Toutefois, cela reste une simulation des comportements adaptatifs en réaction à des stimuli qui se base sur des algorithmes statistiques et où la subjectivité n'a pas sa place.

De plus, nous ne pouvons pas dire que ces machines comprennent ce qu'est une route, une voiture ou démontrent un vécu subjectif. Face à ce constat, Searle dirait qu'elles sont dépourvues ce qui est « ontologiquement subjectif » ce qui les empêchent d'agir selon des désirs et de manière intentionnelle. Elles sont certes programmées et apprennent à respecter le code de la route ainsi qu'à éviter les accidents, mais elles ne peuvent ni « savoir », ni « imaginer », ni « comprendre » les conséquences dramatiques qui pourraient découler en cas d'un manquement à ces règles qu'on leur a attribuées.

II.1.3. Projets robots humanoïdes existants

Autre argument en faveur de la thèse de l'IA forte et en parallèle de son expérience de pensée sur WUNDERKIND, Dennett évoque Cog[38], un projet de robot humanoïde développé dans les années nonante au MIT par Rodney Brooks et pour lequel il a été un observateur de première main. L'ambition de ce projet était de développer un robot apprenant qui interagit avec son propre environnement, comme le fait l'être humain. Les chercheurs partaient de l'idée que la cognition allait émerger de l'interaction sensori-motrice avec le monde réel.

Cog était un robot qui disposait d'un torse, de bras, d'yeux motorisés, et qui pouvait effectuer des gestes à l'aide de son corps. Son intelligence devait donc découler d'un apprentissage ainsi que d'une interaction avec son environnement. Ce projet avait été pour Dennett l'occasion de critiquer les approches symboliques et traditionnelles de l'IA.

Cependant, aux dernières nouvelles, le projet a été abandonné en 2003 car les capacités matérielles de l'époque étaient en deçà par rapport aux ambitions affichées. Les résultats obtenus

[38] Dennett, Daniel C., « Can Machines Think? Deep Blue and Beyond », *Studium Generale Maastricht*, 1997, p. 30.

ont été particulièrement influencés par l'équipement rudimentaire dont les chercheurs disposaient. Ce projet n'a pas permis d'apporter des éléments significatifs par rapport à son ambition de départ. Naturellement, ce constat n'a pas pu justifier davantage de financements. Néanmoins, cette expérience a constitué les premiers pas de la machine apprenante sous forme d'un robot humanoïde.

Avant même le projet Cog, lorsque Dennett a répondu une première fois à la critique que lui a faite Searle sur l'hétérophénoménologie[39], il cite l'exemple que Edelman lui-même a déclaré que le robot qu'il avait conçu présentait une forme d'intentionnalité réelle, et non une simple imitation artificielle[40].

Pour rappel, Gerald Edelman avait conçu un robot nommé Darwin IV[41], dont l'architecture informatique s'inspirait de celle du cerveau humain. L'objectif du chercheur était de proposer une nouvelle théorie de la conscience, en montrant que ce système permettait au robot d'adopter des comportements à la fois complexes et adaptatifs, sans qu'il soit nécessaire de le programmer spécifiquement à l'avance. Le robot était ainsi capable d'évoluer de manière autonome dans un environnement semé d'obstacles et d'apprendre de ses erreurs.

Cette expérience a conduit Edelman à mieux comprendre comment un cerveau humain parvient à s'adapter à son environnement, grâce à un système flexible fondé sur l'apprentissage par l'expérience et l'interaction avec le monde extérieur. En conséquence, un processus de sélection neuronale s'opère, permettant au cerveau de développer de nouvelles structures et des connexions de plus en plus optimisées. Cette découverte a donc donné lieu à ce qu'on appelle le « darwinisme neural ».

En somme, ces différents projets ont permis, selon Dennett, non seulement d'en apprendre davantage sur les mécanismes de notre cerveau — notamment à travers le processus de sélection

---

[39] Par hétérophénoménologie, nous entendons la méthode mise au point par Dennett dans *La conscience expliquée* selon laquelle la voie la plus objective en vue d'atteindre une description phénoménologique fidèle aux expériences privées tout en veillant à être compatible avec la méthodologie scientifique actuelle (Dennett, 1993, p. 98). Autrement dit, il cherche à faire la jonction entre d'un côté l'objectivité et de l'autre, l'étude des expériences subjectives. C'est-à-dire que sa méthode prend en compte les descriptions introspectives des individus par rapport à leur propre expérience, mais il les interprète d'un point de vue extérieur.

[40] John Searle, *Le mystère de la conscience*, Paris, Éditions Odile Jacob, 1999, p. 127.

[41] Edelman, Gerald M., et al., « Synthetic Neural Modeling Applied to a Real-World Artifact », *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no 15, 1992, p. 7267-7271.

neuronale mis en lumière par Edelman — mais aussi de développer des initiatives telles que Cog, qui ambitionnaient de concevoir un robot capable de découvrir son environnement par auto-apprentissage.

II.1.4. La théorie computationnelle de l'esprit

Dans *La conscience expliquée*, Dennett compare l'esprit humain à une machine virtuelle, celui-ci peut être réductible à des fonctions computationnelles. En ce sens, il s'inscrit dans le sillage de Turing et de Von Newmann. La conscience est donc selon lui un programme d'ordinateur évolué qui façonne le fonctionnement de notre cerveau.

« La conscience humaine est elle-même un énorme complexe de mèmes (ou plus exactement d'effets de mèmes dans le cerveau) ; elle fonctionne en quelque sorte comme une machine virtuelle à la von Newmann implémentée dans l'architecture parallèle d'un cerveau, lequel n'est pas conçu pour de telles activités »[42].

Derrière cette approche, Dennett défend une vision à la fois fonctionnelle et computationnelle de l'esprit humain. Il retire à la conscience son caractère immatériel et mystérieux et la perçoit plutôt comme le résultat d'un programme complexe fonctionnant dans le cerveau. Son explication lui permet ainsi de réfuter le « Théâtre Cartésien », une idée tenace selon laquelle la conscience prendrait ses quartiers uniquement dans le cerveau et de manière centralisée[43] et dans lequel il y aurait un « soi » ou une sorte de spectateur intérieur qui à la fois observerait et traiterait toutes les informations liées à l'expérience. Selon le philosophe, cette vision de la conscience est, comme son nom l'indique, issue de l'influence de Descartes qui avait popularisé l'idée qu'il existerait une glande pinéale[44] dans le corps de l'homme qui serait en quelque sorte le siège de l'esprit conscient.

C'est après toutes ces considérations citées précédemment que Dennett développe son modèle des Versions Multiples, autrement appelé le *multiple drafts model*. En effet, selon lui, « nous ne faisons pas directement l'expérience de ce qui se passe dans nos rétines, dans nos oreilles, sur la

[42] Daniel C. Dennett, *La conscience expliquée*, Paris, Éditions Odile Jacob, 1993, p. 262.

[43] *Ibid.*, p. 141-142.

[44] *Ibid.,* p. 214.

surface de notre peau. Ce dont nous faisons effectivement l'expérience est le produit de nombreux processus d'interprétation »[45].

De cette façon, ces processus d'interprétation reçoivent des représentations brutes, puis, ces dernières s'insèrent dans un flux d'activité distribué qui se produit dans plusieurs parties du cerveau. Autrement dit, le travail exercé par la conscience est réparti dans le temps et l'espace au sien même du cerveau. Ensuite, le cerveau fonctionne en gérant de multiples flux de traitements simultanés et de manière non synchronisée. La conscience devient donc un phénomène distribué et non centralisé.

Comme le souligne Bruno Leclercq[46], l'information n'est pas centralisée au niveau du cerveau, mais que cette dernière est plutôt largement distribuée au sein du système nerveux ainsi que plus généralement dans l'organisme. C'est en fait un puissant argument neurophysiologique contre la notion de théâtre cartésien. En effet, nous sommes dotés de différents arcs réflexes distribués à travers tout le corps afin entre autres de nous prémunir face à la douleur. Il prend ainsi l'exemple du doigt que l'on approche trop près du feu. Le corps, voulant ainsi se protéger du danger de la brûlure, dispose alors de tout un réseau nerveux qui fait en sorte que l'individu retire immédiatement sa main du feu sans pour autant que toute cette information ne soit gérée et uniquement logée dans le cerveau.

Une fois que son hypothèse est posée, la pensée de Dennett l'amène à affirmer que si la conscience n'est qu'une machine virtuelle fonctionnelle, alors il n'est pas inenvisageable d'imaginer qu'un robot puisse un jour en principe être conscient[47].

Cette théorie computationnelle de la conscience qu'il défend renvoie aussi à son article de 1997, mais aussi au dernier publié en 2019 :

---

[45] *Ibid.,* p. 147-148.

[46] Leclercq, Bruno, « Ni fantôme, ni zombie : L'émergence de la conscience subjective dans le flux des expériences », *Bulletin d'Analyse Phénoménologique*, vol. 10, no 3, 2014.

[47] Daniel C. Dennett, *La conscience expliquée*, Éditions Odile Jacob, 1993, p. 535.

In one sense of 'machine' it is already quite clear that we are machines: we are composed of cells - there is no reason to suppose that we have any other secret ingredient - and cells are machines. A cell can't think, in any interesting sense of 'think', but a mega-machine made of a few trillion cells can think, since you and I can think, and that is what we are[48].

« After all, everything we now know suggests that, as I have put it, we are robots made of robots made of robots . . . down to the motor proteins and their ilk, with no magical ingredients thrown in along the way »[49].

Même si Dennett défend une théorie computationnelle, il n'affirme cependant pas que nous ayons actuellement atteint l'IA forte et qu'il existe donc encore un écart important entre les systèmes actuels et ceux issus de la science-fiction qui continuent de bercer notre imaginaire collectif. Sa position consiste à dire que puisque c'est possible, ce n'est donc pas souhaitable.

C'est cette position qu'il défendra lorsque dans son article de 2019, lorsqu'il en appelle à effectuer une distinction franche entre un agent (ou collègue) et un outil afin de critiquer le phénomène d'humanisation dans lequel certains projettent leur propre humanité dans ces robots. Il ajoute même que nous n'avons pas besoin d'agent intelligent, car la puissance de calcul dont nous disposons actuellement se suffit à elle-même pour nous aider dans la vie quotidienne et que nous n'avons pas besoin de recourir à ceux-ci.

Le modèle computationnel tel que défendu par Dennett ne manquera pas de faire réagir John Searle dans le *mystère de la conscience.* Ce dernier lui reprochera une approche anti-biologique qui réduit l'esprit à un programme d'ordinateur. Or, comme nous le verrons plus en profondeur dans la

---

[48] Dennett, Daniel C., « Can Machines Think? Deep Blue and Beyond », *Studium Generale Maastricht*, 1997, p. 31.

[49] Dennett, Daniel C., « What Can We Do? », in J. Brockman (dir.), *Possible Minds: Twenty-Five Ways of Looking at AI*, New York, Penguin Press, 2019.

section consacrée à l'argument neuro-biologique associé à la thèse de l'IA faible, Searle soutient que ce sont les processus cérébraux qui génèrent la conscience[50].

II.1.5. L'argument de l'invariance organisationnelle

II.1.5.1. Qu'est-ce que le principe d'invariance organisationnelle ?

Dans cette section, nous allons désormais nous pencher sur une approche philosophique radicalement opposée à celle de Searle, c'est-à-dire une approche qui conserve l'importance de l'expérience consciente, sans pour autant la limiter à une base biologique.

C'est précisément ce que propose David Chalmers, philosophe contemporain de l'esprit, dont l'œuvre développe une théorie pour le moins originale. Dans *The Conscious Mind: In Search of a Fundamental Theory*[51], Chalmers avance notamment l'idée d'invariance organisationnelle : une hypothèse selon laquelle la conscience pourrait, en principe, émerger de toute structure fonctionnelle équivalente, indépendamment de son substrat physique.

Dans le chapitre 7 de *The Conscious Mind*, Chalmers s'interroge sur les conditions de possibilité de l'émergence de la conscience. Il avance que la conscience ne dépend pas du matériau biologique du cerveau, mais bien de son organisation fonctionnelle abstraite. Autrement dit, ce qui importe, ce n'est pas la substance physique, mais le schéma causal des interactions entre les composantes du système, autrement dit ses entrées, ses sorties et la manière dont elles sont connectées.

"A natural suggestion is that consciousness arises in virtue of the functional organization of the brain [...] What counts is the brain's abstract causal organization, an organization that might be realized in many different physical substrates […] Conscious experience arises from fine-grained functional organization. More specifically, I will argue for a principle of organizational invariance,

[50]John Searle, *Le mystère de la conscience*, Paris, « Éditions Odile Jacob », 1999, p. 197.

[51] David J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory,* New York, Oxford University Press, 1996.

holding that given any system that has conscious experiences, then any system that has the same fine-grained functional organization will have qualitatively identical experiences »[52].

À partir de cette base, il défend ce qu'il appelle le principe d'invariance organisationnelle. Il avance que si deux systèmes partagent la même organisation fonctionnelle, leurs expériences conscientes seront qualitativement identiques, et ce, indépendamment de leur substrat physique. En d'autres termes, la conscience peut selon lui émerger indépendamment de sa base matérielle.

Ce point est particulièrement intéressant par rapport au parallèle que l'on peut faire avec la machine. En effet, si l'on en croit Chalmers, si une machine possédait une organisation fonctionnelle semblable à celle d'un système conscient, alors nous pourrions dire que la machine possèderait elle aussi une conscience. Mais nous verrons plus tard dans son article de 2023 que malgré les prouesses des LLM, nous ne pouvons pas encore dire que ces modèles ont déjà acquis un esprit artificiel et qui leur manque un certain nombre de composantes.

II.1.5.2. Démonstration via l'expérience de pensée sur les *qualia* absents, s'estompant et dansants

Pour défendre sa position, il mobilise une série d'expériences de pensée fondées sur les *qualia* absents, s'estompant, et dansants[53]. Il commence par l'hypothèse des *qualia* absents. Il nous invite alors à comparer deux entités :

D'un côté, DAVID[54], un être humain constitué de neurones biologiques et dont l'organisation cérébrale est bien fonctionnelle et consciente. De l'autre, nous avons ROBOT, un système purement artificiel ayant la même organisation fonctionnelle que David, à la différence que contrairement à celui-ci, ROBOT est uniquement composé de puces de silicium. Ainsi, ROBOT n'est pas constitué de matières biologiques.

---

[52] *Ibid.,* p. 247-248.

[53] Chalmers, David J., « Absent Qualia, Fading Qualia, Dancing Qualia », in Thomas Metzinger (dir.), *Conscious Experience*, Paderborn, Schöningh, 1995, p. 309-328.

[54] Chalmers parle de lui-même dans l'article, alors, par question de facilité, nous choisissons de nommer l'entité humaine par son prénom.

L'hypothèse qui consiste à dire que ROBOT, malgré sa similarité avec DAVID, ne possède pas de conscience est appelée celle des *qualia* absents. ROBOT se comporterait donc comme DAVID, sauf qu'il n'éprouverait rien intérieurement. Il s'agirait en fait que d'une simple simulation d'un esprit conscient. Ce système serait, pour ainsi dire, un zombie philosophique.

Une fois que Chalmers a posé le décor, il décide de mettre à l'épreuve cette hypothèse. Chalmers introduit alors un scénario de remplacement progressif. C'est celui des f*ading qualia* ou bien des *qualia* qui s'estompent. Concrètement, les neurones biologiques de DAVID sont petit à petit remplacés par des puces en silicium assurant exactement les mêmes fonctions. Lorsque le remplacement est partiel, DAVID devient JOE et à 100 %, il devient ROBOT. Or, à chaque étape, le comportement demeure inchangé, de même que les jugements introspectifs. Cela le conduit à se demander : à partir de quel moment la conscience disparaît-elle ?

Ensuite, Chalmers s'interroge pour savoir à partir de quel moment l'expérience consciente de DAVID disparaît ou change au cours de ce processus. Il propose de répondre à cette question en imaginant deux scénarios : le premier, celui où la conscience disparaitrait graduellement, c'est-à-dire le scénario qu'il associe au *qualia* s'estompant, et le deuxième, celui la conscience qui disparaitrait de manière subite.

Dans le premier cas de figure, Chalmers imagine qu'à mesure que le processus de remplacement s'opère, DAVID perdrait progressivement toute expérience phénoménale jusqu'à sa disparition. Ainsi, la couleur rouge vif apparaitrait de plus en plus terne et les expériences sonores deviendraient de moins en moins perceptibles. Sauf que Chalmers réfute ce cas de figure et ce, pour une raison majeure : celle de la continuité fonctionnelle.

En effet, le système JOE continuerait de faire l'expérience de la couleur rouge vif, mais il percevrait plutôt du rose terne. Il n'est donc pas raisonnable de penser que JOE serait systématiquement dans l'erreur. Chalmers rejette l'idée selon laquelle un être conscient puisse être aussi déconnecté de ses propres sensations.

Dans le deuxième cas de figure, celui où la conscience disparait soudainement, le philosophe australien imagine que DAVID perde subitement sa conscience à un stade précis, par exemple à partir du moment où 25 %, 50 % ou 75 % des neurones seraient remplacés. Si l'on s'en

tient à ce scénario, cela voudrait dire qu'un tout petit changement au sein du système aboutirait à une transition brutale, où l'on passerait subitement d'un état pleinement conscient à complètement inconscient.

Chalmers épingle ce scénario pour le réfuter comme le premier, car une telle transition semble à la fois arbitraire et absurde. Cela voudrait donc dire qu'il existerait un seul neurone qui jouerait un rôle décisif qui aboutisse à un changement aussi abrupte et radical ? Le philosophe rejette cette idée selon laquelle il existerait un seuil arbitraire et caché au sein d'un processus continu. Pire encore : cette idée s'inscrit en opposition avec celle d'envisager la conscience comme un phénomène global et distribué dans le cerveau.

Enfin, la seconde expérience de pensée vise cette fois-ci l'hypothèse basée sur ce qu'il appelle les *qualia* dansants. Il nous propose donc cette fois-ci d'imaginer deux systèmes distincts : DAVID et BILL, deux entités isomorphes fonctionnelles dans le sens où elles possèdent la même organisation fonctionnelle et qu'elle agissent de la même façon. Leur expérience consiste à voir une pomme. Lorsque DAVID en voit une rouge, BILL en voit une verte. Cette différence de perception s'explique par le fait qu'un circuit composé de puces en silicone de BILL a été implémenté au cerveau de DAVID.

Ensuite, à l'aide d'un interrupteur, on peut passer d'un circuit à un autre. De cette manière, DAVID passerait d'une expérience du rouge à une expérience du vert en appuyant simplement sur un bouton. Son expérience subjective serait alors radicalement changeante. Chalmers s'interroge ensuite sur ce scénario en notant toutefois que malgré les changements phénoménaux brusques, DAVID ne remarquerait rien d'anormal.

Comment peut-on expliquer cela ? Car, selon le philosophe, l'organisation fonctionnelle même du système de DAVID, ses jugements ainsi que son comportement resteraient les mêmes. Il serait en effet invraisemblable que des expériences phénoménales aussi radicalement différentes puissent co-exister au sein d'une organisation fonctionnelle inchangée. Et si tel était le cas, il faudrait admettre que notre expérience consciente change brusquement sans que nous nous en rendions compte. Cela reviendrait à dire que la conscience pourrait changer radicalement sans impact cognitif perceptible.

En résumé, Chalmers estime qu'un esprit artificielle peut apparaître à partir de substrats non-biologiques comme du silicium et que par conséquent, la conscience n'est pas uniquement réservée aux humains, mais qu'elle dépend surtout de la structure fonctionnelle et non du matériau de base. En principe, si l'on suit son raisonnement, une IA suffisamment développée et ayant la même organisation causale et fonctionnelle qu'un humain pourrait avoir des *qualia*. Il ne s'agirait donc pas d'une simple simulation de ceux-ci, mais bien le fait de vivre une véritable expérience subjective.

Une fois son principe ainsi que sa démonstration posés, la question qui serait désormais intéressante à se poser une fois que nous avons abordé le principe d'organisation fonctionnelle est de savoir si avec l'arrivée de l'IA connexionniste, Chalmers estime que nous pouvons dire que les IA actuelles démontrent une conscience ? Comme nous allons le voir, la réponse n'est pas aussi tranchée que l'on pourrait imaginer.

Nous savons effectivement que la particularité de l'IA connexionniste repose sur son système de réseau de circuit neuronal qui reproduit ce qu'il se passe dans un cerveau organique. Mais est-ce suffisant pour affirmer que les LLM sont conscients ?

Dans son article de 2023[55], Chalmers ne pense pas qu'à l'heure actuelle, les LLM soient conscients. Il rappelle que la conscience reste ce que l'on appelle une expérience subjective, c'est-à-dire la sensation de savoir ce que cela fait d'être un être. En outre, une conscience ne se mesure pas uniquement à la performance conversationnelle. Tout comme Dennett avant lui, il met en garde par rapport à notre tendance à projeter de la conscience chez les machines qui s'était déjà manifestée du temps du programme ELIZA dont nous avons parlé en première partie de ce travail. Il faut garder à l'esprit que malgré ces prouesses conversationnelles, ces IA génèrent du contenu sans intentionnalité et sur la base de probabilités.

Toutefois, il reconnaît tout de même que les LLM font preuve d'une intelligence générale dans le sens où ils désormais sont capables d'accomplir des tâches très variées, comme la poésie, le codage, le jeu, l'argumentation, etc. En effet, cette diversité dans les tâches n'était pas présente du côté des systèmes-experts qui, comme leur nom indique, n'étaient spécialisés que dans un domaine

[55] Chalmers, David J., « Could a Large Language Model Be Conscious? », *Boston Review*, 9 août 2023.

en particulier. De même, les premières IA conversationnelles telles que PARRY ou ELIZA dont nous avons parlé dans la première de notre travail n'étaient pas capables de remplir ces fonctions. Néanmoins, cette diversité dans les tâches ne constitue pas une preuve définitive en faveur d'une conscience artificielle, même si cette compétence tend à s'en rapprocher.

Mais faut-il pour autant qu'il faut faire le deuil d'une conscience artificielle telle qu'il l'a théorisée dans les années nonante ? Là est toute l'originalité de sa position. Chalmers ne ferme pas la porte à l'hypothèse d'une conscience artificielle. Cela demandera du temps, mais il estime que les récentes avancées en matière de LLM pourraient accélérer son apparition.

En revanche, il évoque une série de critères que les LLM ne remplissent pas encore et qui sont, selon lui, nécessaires à la conscience. Parmi ceux-ci, l'absence d'un corps et de sens : ils ne sont pas incarnés et ne possèdent donc pas de perception directe. Cependant, Chalmers considère que cette absence n'est pas incompatible avec l'existence d'une conscience ou d'une faculté de compréhension. Il ajoute que même sans les sens, une IA pourrait raisonner sur elle-même et, plus largement, sur le monde.

Sa thèse, selon laquelle l'incarnation et la perception ne sont pas nécessaires à la conscience ou à la compréhension, mérite discussion. Si l'on se base sur la cognition incarnée, penser, percevoir ou comprendre ne relèvent pas d'une simple manipulation de symboles, mais d'une interaction en temps réel avec le monde. Or, dans le cas des LLM, il s'agit précisément d'IA désincarnées : elles manipulent du texte, mais ne reproduisent pas pour autant la structure réelle de la cognition humaine. C'est dans la troisième partie de notre travail que nous défendrons, contrairement à Chalmers, que la compréhension implique un corps ainsi qu'un rapport au monde. Selon nous, la pensée ne se manifeste pas par une simple simulation.

Pour en revenir à Chalmers, il adopte une posture prudente quant à la question de la possibilité d'une conscience artificielle voit le jour. En effet, il n'accorde pas une foi aveugle en l'IA forte en l'état des choses, mais il tente, à la manière prospectiviste, de poser les conditions de possibilité d'une conscience artificielle. Naturellement, ces avancées technologiques vont forcément nous amener à s'interroger sur le plan éthique. Comme ce qu'Andler avait identifié avec « l'irrésistible ascension », il n'y a en effet pas de raison de penser que les chercheurs en la matière interrompent leurs recherches.

Par conséquent, tout comme Dennett, il estime qu'il n'est souhaitable de construire de tels systèmes conscients et il alerte en particulier au sujet de la création d'agents artificiels dotés d'une forme d'unité psychologique. À travers son article, le philosophe australien en appelle en fait à la lucidité collective quant à l'éventualité d'une conscience artificielle et de ses implications éthiques et morales. Cela suppose donc d'anticiper les responsabilités qui en découleraient.

En conclusion, même si les IA connexionnistes partagent une similitude avec le principe d'invariance organisationnelle à travers son architecture sous forme de réseau neuronal, cela ne veut pas dire pour autant que nous nous trouvons actuellement face à des IA conscientes. Il reste encore des caractéristiques, comme le corps et la biologie, qu'elles ne possèdent pas. Nous reviendrons précisément sur l'importance du corps dans la dernière partie de notre travail en nous appuyant sur l'apport de Hubert Dreyfus et de Daniel Andler.

II.2. Arguments en faveur de l'IA faible

II.2.1. Le test de Turing, ses variantes et la critique de Dennett

Dans son article publié en 1984, Dennett note que nous sommes devenus de plus en plus dépendants aux machines ainsi que de leurs capacités cognitives, mais qu'on aurait bien tort de surestimer leurs capacités. Il estime que le premier article paru sur la question et rédigé par l'informaticien Alan Turing ne répond pas véritablement à la question qui constitue la base de son article paru en 1950.

Dennett considère que Turing n'a pas imaginé son test comme un outil qui pourrait être exploité dans le cadre de la psychologie scientifique[56]. La proposition formulée par l'informaticien britannique doit plutôt être vue comme un outil de réflexion philosophique. Par conséquent, il regrette que ce test ait bénéficié d'une attention excessive, sans pour autant permettre de cerner précisément les enjeux liés à l'intelligence artificielle.

[56] Dennett, Daniel C., « Can Machines Think? », in *Brainchildren: Essays on Designing Minds*, Cambridge (Mass.), MIT Press, 1998, p. 4.

Un exemple souvent cité dans les débats sur l'intelligence artificielle et la validité du test de Turing est celui du programme PARRY[57], conçu par le psychiatre Kenneth Colby. Ce programme simule un patient paranoïaque convaincu d'être constamment observé et persécuté par la mafia. Pour évaluer ses capacités de simulation, Colby a organisé une série d'échanges via téléscripteur entre PARRY et des psychiatres, à qui il était demandé de déterminer s'ils avaient affaire à un véritable patient humain ou à un programme informatique. Les résultats furent surprenants : le taux de réussite pour identifier correctement le programme ne dépassait pas 48 %, soit un résultat comparable à une réponse aléatoire.

Daniel Dennett analyse cette expérience en soulignant un élément souvent négligé, mais pourtant décisif : les psychiatres engagés dans ces tests adoptaient une posture clinique standard, fondée sur l'écoute empathique et l'absence de confrontation. En suivant leur déontologie médicale, ils ne cherchaient ni à piéger ni à déstabiliser leur interlocuteur. En conséquence, les questions posées à PARRY étaient celles que l'on pose habituellement à un patient paranoïaque, ce qui conférait un avantage évident au programme, car son manque de coopération apparent durant l'interaction renforçait l'impression de paranoïa qui en découlait.

Cela démontre que la machine avait adopté une stratégie défensive crédible, reposant sur la logique propre de la paranoïa, laquelle inclut souvent une certaine résistance à la coopération. Dans ce rôle, le programme s'avérait étonnamment convaincant. Sauf que PARRY n'était pas conçu pour simuler l'intelligence humaine dans sa globalité, mais bien pour optimiser ses réponses dans un domaine restreint. Ce choix s'explique par un impératif de rentabilité computationnelle. Le programme se limitait aux répliques nécessaires pour maintenir l'illusion du trouble paranoïaque, sans mobiliser de connaissances générales sur le monde, ce qui aurait alourdi son fonctionnement sans nécessairement améliorer sa performance durant cette expérience.

Si l'on suit le raisonnement de Dennett, ces variantes du test de Turing relèvent d'une version affaiblie : elles ne reposent que sur des conversations restreintes ou sur une expertise particulière, et non sur des compétences générales. Il parle à leur propos d'une « qualité de façade ». Selon lui, ces approches conduisent à surestimer l'intelligence des systèmes, comme nous le verrons

[57] *Ibid.,* p. 13-14.

dans la section suivante consacrée au prix Loebner. Cette surestimation soulève un problème social : elle peut amener certains à attribuer un statut moral ou cognitif à de simples artefacts.

Cet aspect rejoint également les travaux de Ned Block, qui affirmait que, même avec un vocabulaire limité, les variantes du test de Turing produiraient une « explosion combinatoire » donnant toujours l'illusion de capacités conversationnelles impressionnantes. Cette observation montre à quel point nous avons tendance à surestimer les performances des machines.

En somme, Dennett nous met en garde contre le risque de nous illusionner sur les capacités des machines et de leur attribuer une intelligence qu'elles ne possèdent pas. La source de cette erreur de jugement tient à la fois à la conception de ces systèmes et à notre propre propension à projeter de l'intelligence sur les systèmes experts et sur les variantes du test de Turing.

Toutefois, ce n'est que dans le *post-scriptum* de l'article qu'il reconnaît quand même que le test de Turing à lui seul ne permet pas de justifier l'attribution de la pensée. En effet, sans une confrontation au réel à travers son propre vécu, il n'y a pas d'authenticité. Lorsque l'on parle de confrontation, nous pensons au fait de s'adapter à des contextes variés et à interagir avec des personnes ainsi que dans leur environnement physique.

Enfin, Dennett ne tranche pas définitivement sur le fait qu'une machine ne puisse jamais avoir un conscience subjective. Néanmoins, ce *post-scriptum* met en évidence son idée selon laquelle l'intelligence est incarnée et ancrée dans un vécu[58].

II.2.2. Critique du prix Loebner

À la fois dans son *post-scriptum* de 1984 et dans son article *Deep Blue and Beyond* de la même année, Dennett apporte son regard critique quant au prix Loebner, une compétition organisée au *Computer Museum* de Boston. Le but de ce concours est de mettre à l'épreuve des intelligences artificielles, selon les principes du test de Turing. Ainsi, les participants avaient pour mission de classer les AI en fonction de leur « humanité » perçue.

[58] *Ibid.,* p. 21.

Il relève avoir été surpris par le manque d'esprit critique de certains juges lors du prix Loebner, bien qu'il reconnaisse que les règles du concours limitaient leur liberté d'interrogation. En ne pouvant pas poser de questions incisives, les juges laissaient en réalité les IA mener la conversation, ce qui facilitait leur capacité à les tromper.

Ainsi, bien qu'il admette que ce prix ait constitué une expérience sociale intéressante, Dennett identifie de nombreux biais ayant conduit les participants à surestimer les capacités de ces chatbots. Parmi eux figure le biais d'attente, observé lors de la troisième édition du concours : des journalistes, persuadés d'évaluer des IA particulièrement performantes, avaient tendance à attribuer aux machines les propos les moins éloquents… qui provenaient pourtant d'êtres humains.

Ce que l'on retient de cette expérimentation, c'est que ce concours n'a pas permis de mettre en place de manière rigoureusement scientifique le test de Turing. Ce qui pousse Dennett à affirmer que le celui-ci est encore trop compliqué en mettre en oeuvre et que cette expérience n'a pas contribué à considérer ce test comme sérieux. D'un côté, les IA de l'époque telles que ELIZA étaient encore trop « primitives » et, de l'autre, les conditions mêmes du prix Loebner a permis de révéler un ensemble de biais qui ont rendu les conclusions peu fiables.

Mais peut-on dire que cela soit toujours le cas actuellement ? En tout cas, plusieurs articles de presse ont clamé que ChatGPT-4 avait battu le test de Turing à plate couture. Mais qu'en est-il vraiment ?

Très récemment, un article paru en 2025[59], bien qu'encore en *preprint*, semble répondre à ce qu'exigeait Dennett à l'époque, c'est-à-dire une procédure scientifique rigoureuse en lien avec le test de Turing classique. Pour ce faire, l'étude met en place une configuration à trois participants, à l'instar du test éponyme. Il y avait donc la participation d'un juge humain (soit l'interrogateur), un agent humain et d'un agent IA (tels que GPT-4.5 ou LLaMa de Meta). Le tout était réalisé sous forme d'une conversation textuelle simultanée. L'objectif était évidemment que le juge devine lequel de ses deux correspondants est humain.

[59] Jones, et Bergen, « Large Language Models Pass the Turing Test », *arXiv.org*, mars 2025.

Il est à noter qu'il y avait deux sortes d'IA : l'une nommée « NO-PERSONA » est un LLM qui répond de manière standard et l'autre, appelée « PERSONA » reçoit l'instruction d'imiter une personnalité humaine spécifique. Dans le cas de l'étude, les chercheurs ont choisi de la programmer comme un jeune adulte introverti sensible à l'humour que l'on retrouve typiquement sur internet et qui présente un intérêt pour la musique *underground*.

Une fois que c'est fait, le juge entre en conversation pendant cinq minutes avec ses deux interlocuteurs (l'un humain, et l'autre une IA), sans évidemment savoir qui est qui. Le juge pouvait poser des questions libres soit à l'un soit aux deux. Une fois les cinq minutes écoulées, celui-ci devait répondre par oui ou par non lequel des deux était humain.

Contrairement au prix Loebner, cette étude a bénéficié d'un large échantillon et les IA « NO-PERSONA » et « PERSONA » ont été clairement définies et nous savons aussi que ces IA conversationnelles sont plus performants que celles testées du temps des années nonante.

Il s'est donc avéré que GPT-4.5 PERSONA a été confondue pour un humain dans 73 % des cas. Force est de constater que lorsque l'on « prompte » l'IA d'une personnalité propre, elle parvient à mieux tromper le juge. À titre de comparaison, le programme ELIZA, qui servait de témoin et malgré naturellement l'explosion combinatoire dont elle bénéficie comme l'avait déjà fait remarquer Ned Block, a été le moins confondu comme un humain.

Cette étude, bien qu'elle ne soit pas encore passée par le filtre du *peer-review*, nous permet quand même d'interroger la vision de Dennett selon laquelle un test de Turing ne pourrait pas être configuré selon la méthodologie scientifique. En tout cas, des chercheurs se penchent actuellement pour évaluer scientifiquement ces IA génératives à travers cette proposition de Turing.

Nous pouvons dès lors raisonnablement affirmer que les compétences conversationnelles des machines s'améliorent et que cela convaincrait sans doute Turing lui-même que les machines « pensent ». Ces études mettent en évidence qu'il est possible de nos jours de nous rapprocher du test du célèbre informaticien britannique à travers une méthodologie scientifique adaptée faisant notamment appel à un échantillonnage, un groupe témoin, une comparaison des différents programmes, des critères d'évaluation, etc.

Reste maintenant à savoir si la pensée peut se réduire à des capacités conversationnelles, sans rapport direct au monde. Ce point sera examiné dans la dernière partie de ce travail, en lien avec la notion de super-intelligence prométhéenne développée par Andler

II.2.3. L'ambiguïté linguistique et la connaissance du monde

Reprenant les précédents travaux du chercheur américain et informaticien Terry Winogard, Dennett rappelle que ce dernier avait mis en évidence que deux phrases, *a priori* linguistiquement proches, contenaient une forme d'ambiguïté par rapport au pronom « ils ». Dans l'article original, il s'agit en anglais des phrases[60] : « *The committee denied the group a parade permit because they advocated violence* » et « *The committee denied the group a parade permit because they feared violence* ».

Comme nous pouvons le constater, ces deux phrases n'ont effectivement pas tout à fait le même sens, et c'est donc à l'ordinateur de devoir lui-même décider du sens à leur attribuer, en interprétant correctement à qui renvoie le pronom. En bref, l'ordinateur doit répondre lui-même à la question : à qui fait-on référence ?

Pour Dennett, la seule chose qui permet de trancher définitivement cette ambiguïté est la question de la connaissance sur le monde, c'est-à-dire les éléments contextuels relatifs à la politique, les circonstances sociales, etc. Sans cette compréhension du monde dans lequel nous vivons, il est donc impossible de donner du sens à ces deux phrases. Evidemment, cette connaissance du monde était particulièrement limitée du temps où Dennett avait rédigé son article.

Tout comme Turing avant lui, il avait constaté que les bases de données de l'époque ainsi que leur mémoire étaient particulièrement limité. Qui plus est, internet n'était pas encore accessible, or nous savons que les LLM exploitent actuellement des bases de données gigantesques et majoritairement issues de la toile à travers de grands corpus de textes.

[60] Dennett, Daniel C., « Can Machines Think? », in *Brainchildren: Essays on Designing Minds*, Cambridge (Mass.), MIT Press, 1998, p. 7.

Dans son article de 2019[61], Dennett note tout de même que Turing n'avait pas anticipé la compétence que pourrait avoir les machines à traiter une immense masse d'informations à travers le *Big Data*[62], c'est-à-dire l'ensemble des données que l'on retrouve sur internet. Cette faculté permet donc d'accroitre considérablement la connaissance du monde que peut par exemple avoir un LLM. Dennett, souligne que cette technologie peut effectivement donner l'impressionner d'une compréhension authentique, car elle est très performante lorsqu'il s'agit de repérer des tendances ou ce qu'il appelle des schémas probabilistes.

L'avancée technologique de l'IA permet donc de dire que dans ce cas-ci que la limite de la connaissance du monde présentée par Dennett en 1984 n'est plus aussi évidente de nos jours. Toutefois, dans la dernière partie de notre travail, nous explorerons la question des éléments contextuels, qui, comme le souligne Hubert Dreyfus et Andler, sont une composantes essentielles de nos vies auxquelles les machines n'ont pas à ce jour encore accès pour aborder une Situation.

En ce qui concerne l'ambiguïté linguistique, une question se pose : cette limite existe-t-elle encore avec des LLM comme ChatGPT-4 ? Récemment, un article encore en *preprint* et intitulé *We're afraid language models aren't modeling ambiguity*[63] publié en 2023 démontre, par exemple, que mêmes les modèles les plus avancés, comme ChatGPT-4, rencontrent des difficultés quant il s'agit pour ces derniers de reconnaître et désambigüiser correctement des énoncés ambigus. Pour donner un ordre d'idée, ces modèles n'ont obtenu que 32 % de réussite, contre 90 % pour des participants humains. Les auteurs de l'étude en profitent pour préciser que l'ambiguïté est indissociable de notre langage et que c'est grâce à celle-ci que nous pouvons anticiper les malentendus ou revoir notre propre interprétation à la lumière des informations que l'on nous donne. Or, ChatGPT-4 a démontré des difficultés à « désambigüiser » des phrases qu'on lui a soumises.

---

61 Dennett, Daniel C., « What Can We Do? », in J. Brockman (dir.), *Possible Minds: Twenty-Five Ways of Looking at AI*, New York, Penguin Press, 2019.

62 *Idem*.

63 Liu, Alisa, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith et Yejin Choi, « We're Afraid Language Models Aren't Modeling Ambiguity », *arXiv.org*, avril 2023. DOI : 10.48550/arXiv.2304.14399.

Dans le même esprit, une autre étude, *Linguistic Ambiguity Analysis in ChatGPT*[64], a dressé une typologie des ambiguïtés, c'est-à-dire en classant les lexicales, les syntaxiques et les sémantiques et en les soumettant au système d'OpenAI.

Pour donner un exemple de la manière dont ils s'y sont pris pour chaque type d'ambiguïté, la lexicale contenait des mots qui pouvaient avoir plusieurs significations, comme le mot *bank* qui renvoie tant à la rive qu'à la banque. Pour la syntaxique, ils ont par exemple utilisé cette phrase : « *Mary saw John with a telescope* ». Comme nous le voyons, cette phrase, si elle n'est pas précisée, possède plusieurs sens : qui des deux personnages possède un télescope ? Enfin, pour l'ambiguïté sémantique, les auteurs de l'article ont recouru notamment à la phrase *My mother and my sister were sad after she shouted at her*. Sans information contextuelle supplémentaire, nous pouvons hésiter quant à savoir à qui se réfère le pronom. Les exemples sont ensuite soumis à ChatGPT qui doit lui-même reconnaître les ambiguïtés puis proposer plusieurs lectures possibles ou bien observer s'il en impose qu'une seule.

L'étude reconnaît que même si l'IA s'est montrée plutôt performante, l'ambiguïté linguistique reste un obstacle structurel majeur. Mais s'agit-il seulement d'une simple question d'entraînement, ou bien d'un obstacle infranchissable pour l'IA actuelle ?

En tout cas, nous pouvons raisonnablement affirmer que la nature du langage est fondamentalement ambiguë. Par exemple, une même phrase peut avoir plusieurs sens selon le contexte, l'intonation ou même les références culturelles. L'ambiguïté linguistique restera toujours un défi fondamental pour l'IA, car la langage humain lui-même n'est pas toujours clair. De même que les humains peuvent mal interpréter ce que disent d'autres personnes, quand bien même ils s'expriment de leur langue maternelle. Ces études montrent en tout cas que les IA progressent vite et qu'un modèle d'IA générative comme GTP-4 est capable de mieux en mieux comprendre certains problèmes liés à l'ambiguïté linguistique.

De plus, il est également à noter que l'ironie, le sarcasme et les métaphores restent difficiles à inférer, car cela implique à la fois une connaissance du monde ainsi que du bon sens afin de comprendre certaines significations cachées. Autrement dit, dans notre vie quotidienne, nous avons

[64] Ortega-Martín, et al., « Linguistic Ambiguity Analysis in ChatGPT », *arXiv.org*, février 2023.

besoin d'indices et d'éléments contextuels afin de comprendre quelles sont les intentions des personnes avec qui nous sommes en interaction. L'approche contextuelle sera primordiale dans la critique que pose à la fois Hubert Dreyfus et Andler lorsqu'ils évoquent l'incapacité de l'IA à faire face à une situation. Ce point, nous y reviendrons en dernière partie de ce travail comme annoncé plus haut.

Dans tous les cas, nous pouvons raisonnablement affirmer que les limites soulevées par Dennett en 1984 sont progressivement en train de s'estomper. En effet, l'IA actuelle bénéfice donc de plus larges bases de données grâce au *Big Data* qui lui permettent d'acquérir davantage d'informations et d'identifier des schémas ou des tendances à partir desquels elle fournit des réponses de plus en plus convaincantes et basées sur des prédictions.

Enfin, pour ce qui est de l'ambiguïté linguistique, nous remarquons que l'IA a grandement amélioré ses capacités par rapport à ses débuts. Les limites d'antan ne sont plus les mêmes qu'actuellement. Néanmoins, il nous semble que comme elle se base sur notre langage qui contient fondamentalement des ambiguïtés, elle ne pourra pas toutes les lever. Elle reste encore sur une mode d'inférence probabiliste sur base de ses apprentissages qui ne sont pas issus de sa propre expérience du monde.

II.2.4. L'argument Penrose/Lucas

Dans son article *Deep Blue & Beyond*, Dennett évoque ensuite l'argument de John Lucas, repris plus tard par Roger Penrose, selon lequel le théorème d'incomplétude de Gödel montre que l'esprit humain ne peut pas être réduit à une simple machine. En effet, selon ce raisonnement, il existerait des vérités que l'humain peut saisir intuitivement, mais que les machines ne pourraient jamais démontrer.

Pour rappel, dans un article scientifique initialement publié en 1961 dans la revue Philosophy[65], le philosophe John Lucas prétend qu'à travers le théorème de Gödel, les esprits ne peuvent pas être assimilés à des machines. Autrement dit, il y aura toujours des vérités qui ne leur seront pas accessibles[66]. Et ce, même si le philosophe imagine que dans le futur, les machines seront

[65] Lucas, John R., « Minds, Machines and Gödel », *Philosophy*, vol. 36, 1961, p. 112-127.

[66] *Ibid*., p. 115.

capables de surpasser les capacités humaines. Les êtres conscients, c'est-à-dire les humains, au contraire des machines, peuvent percevoir la vérité dans certaines propositions que les machines ne peuvent pas démontrer.

Cette thèse philosophique consiste donc à dire que les esprits humains ne peuvent pas être entièrement modélisés, car les machines n'auront pas accès à certaines vérités. Même si des modèles mécaniques de l'esprit humain pourront être conçus, ceux-ci ne pourront jamais rendre complètement compte de la richesse et de la complexité de l'esprit humain[67]. La science aura beau progresser dans notre compréhension de celui-ci, il subsistera toujours des dimensions qui seront inaccessibles à la production scientifique.

Qui plus est, celui-ci plaide également en faveur de la complexité de l'esprit humain dont les études scientifiques n'ont pas encore pu à ce jour délimiter les contours[68]. Puisque le machinisme ne peut rendre compte de cette complexité, il en conclut que les êtres humains ne sont pas réductibles à des ordinateurs.[69].

Théorisé comme un véritable argument massue contre les velléités technophiles, cet argument a tout de même fait l'objet d'un certain nombres de critiques. De nouveau, celui-ci se présente comme un horizon indépassable qui place en quelque sorte l'humain au-dessus de tout. Néanmoins, cette forme d'idéalisation de l'esprit humain implique que ce dernier possèderait une infaillibilité logique.

En outre, cet argument s'appliquait surtout à l'IA symbolique qui, comme nous l'avons vu, fonctionne à travers des systèmes formels. Mais peut-on dire que ce raisonnement s'applique aussi à l'IA connexionniste ? Cette dernière repose sur une architecture fondée sur des réseaux de neurones, dépourvue de structure formelle au sens gödelien. En d'autres termes, le *deep learning* ne s'appuie pas sur un formalisme axiomatique : il ne s'agit pas de manipuler des symboles selon des règles

---

[67] *Ibid*., p. 126.

[68] *Ibid*., p. 127.

[69] Cet argument, nous le verrons plus loin, a évidemment donné du grain à moudre à Searle lorsque celui-ci défend l'idée d'une IA faible et donc limitée face à une hypothétique IA forte qui serait capable de reproduire tout ce qu'un humain est capable de faire (Searle, 1999, p. 65). La théorie de l'IA forte qu'il dénonce est ce qu'il appelle la théorie computationnelle.

explicites de déduction, mais de modifier les pondérations d'un réseau de neurones à partir d'exemples empiriques issus de corpus textuels et d'images.

Ainsi, nous constatons donc que même si cet argument peut paraître séduisant de prime abord et qu'il a incontestablement constitué l'un des arguments phare en faveur de l'IA faible, il n'est par contre plus tout à fait d'actualité avec nos technologies actuelles et plus particulièrement avec l'IA connexionniste. Le raisonnement de cet argument ne laisse donc aucune place à la remise en question de la « supériorité humaine », mais il en fait une sorte de pétition de principe en retirant le fait que l'IA au sens large puissent développer certaines compétences proches des nôtres.

II.2.5. La chambre chinoise : la différence fondamentale entre la syntaxe et la sémantique

Peut-on vraiment dire qu'une machine "comprend" ou est "consciente" ? Ou n'est-elle simplement qu'un dispositif électronique dépourvu de sens, vide de toute expérience subjective ? Cela fait partie des questions auxquelles John Searle tente de répondre dans *Le mystère de la conscience*.

Ce qui constitue le fondement de son argument de la chambre chinoise, c'est qu'il s'oppose à de nombreux penseurs de la philosophie de l'esprit qui assimilent le cerveau à un ordinateur numérique et considèrent que l'esprit conscient fonctionnerait comme un programme informatique. Il y aurait donc, selon cette conception, d'un côté le cerveau interprété comme un *hardware*, et de l'autre l'esprit, assimilé à un *software*. Le cerveau humain manipulerait des symboles et traiterait des informations à la manière d'un ordinateur exécutant des algorithmes.

Searle s'oppose ainsi à ce qu'il appelle la théorie computationnelle de l'esprit[70]. Il s'inscrit donc dans la critique de cette conception de l'IA forte. Car selon lui, la définition d'un ordinateur, c'est avant tout un dispositif qui manipule des symboles formels, notamment avec des 0 et des 1. Il fonctionne en trois étapes : d'abord l'encodage de l'information en binaire, ensuite sa traduction sous forme d'impulsions électriques, et enfin son traitement selon les règles définies par un programme.

[70] John R. Searle, *Le mystère de la conscience*, Paris, Éditions Odile Jacob, 1999, p. 21.

Mais, selon Searle, ce qui se passe dans notre esprit ne se réduit pas à une manipulation mécanique de symboles. Les esprits, en plus de manipuler des signes, sont porteurs de contenu : les mots que nous utilisons ont une signification, une charge sémantique.

C'est à partir de cette critique que Searle propose sa célèbre expérience de pensée de la chambre chinoise. Imaginons une personne enfermée dans une pièce, avec des manuels de traduction chinois. Des questions lui sont envoyées depuis l'extérieur, dans une langue qu'elle ne comprend pas. À l'aide des manuels, elle parvient à répondre correctement aux messages reçus, donnant ainsi l'illusion qu'elle maîtrise le chinois. Pourtant, elle ne comprend rien au contenu de ses réponses. Elle ne fait qu'associer des symboles de façon syntaxique, sans jamais accéder à leur signification.

Cette expérience vise donc à montrer que le traitement syntaxique de l'information ne suffit pas à produire de la compréhension. Searle avance que l'ordinateur fonctionne exactement comme la personne dans la chambre : il ne fait que manipuler des symboles sans pour autant en saisir le sens.

Cela étant dit, à l'instar d'Andler, Searle reconnaît que le cerveau est aussi capable d'effectuer ce qu'une machine accomplit, par exemple des calculs. Mais l'argument de la chambre chinoise ne vise pas à nier toute forme de traitement dans les machines : il veut montrer qu'elles ne font que manipuler des symboles formels sans accès à la sémantique. Cette expérience souligne que l'architecture des programmes repose uniquement sur des règles syntaxiques, sans lien direct avec une quelconque compréhension.

Searle ajoute encore que la syntaxe n'est pas inhérente à la nature physique d'un système, mais qu'elle est dans "l'œil de l'observateur". Autrement dit, le calcul n'est pas une propriété objective des systèmes, mais une interprétation produite par un esprit conscient en dehors du système.

Il va plus loin encore : les impulsions électriques d'un ordinateur sont objectives, certes, mais l'idée qu'elles représentent un calcul ne prend sens que par l'interprétation qu'en donne un observateur humain. En somme, le calcul est un processus abstrait, et n'a de sens qu'à travers une conscience qui lui attribue une signification. Le cerveau, selon Searle, est une machine organique

dont la conscience résulte de processus neuronaux. L'ordinateur, quant à lui, peut simuler certains aspects du fonctionnement cérébral, mais ne pourra jamais simuler un état mental. L'ordinateur est certes capable d'effectuer des millions d'opérations par seconde, mais cela ne constitue en rien une simulation de la psychologie humaine[71].

Searle poursuit sa critique en évoquant l'enthousiasme autour des performances des ordinateurs, comme dans le cas du programme d'échecs, qui avait battu les meilleurs joueurs mondiaux. Les médias s'étaient alors inquiétés d'un possible danger que représenterait une telle machine. Mais pour Searle, l'ordinateur ne comprend strictement rien aux échecs, aux pièces, ni aux stratégies. Il ne fait que manipuler des symboles vides de sens en suivant les instructions qui lui sont données.

Il réitère son idée dans son dernier article en date publié en 2015[72]. Toujours selon lui, Aussi impressionnante que fût la victoire de *Deep Blue* sur Kasparov, il est essentiel de souligner que le programme n'avait aucune conscience de ce qu'il faisait. Or, la conscience est la condition *sine qua non* de toute forme d'activité cognitive authentique.

De plus, il réaffirme que les ordinateurs ne calculent pas réellement : ils effectuent des transitions d'états électroniques que nous interprétons comme des opérations computationnelles. La computation est donc, pour lui, une notion relative à l'observateur. Là où un humain fait une addition avec intention et compréhension, la machine se contente simplement de manipuler des symboles selon des règles syntaxiques, sans pour autant en saisir le sens.

Son argument lui permet donc de répondre aux inquiétudes soulevées par Bostrom qui alerte de son côté sur l'apparition imminente d'ordinateurs intelligents capables de surpasser l'humanité, voire de se retourner contre elle si aucune précaution n'est prise. Bostrom défend donc une vision clairement catastrophiste de l'avenir, mais Searle tempère ces propos en lui répondant qu'il n'y a aucune raison de croire que les ordinateurs puissent représenter un danger réel pour l'humanité, car ils ne peuvent devenir des agents autonomes dotés de croyances, de désirs ou de motivations. Même si un robot était programmé pour tuer, il ne le ferait pas de son propre chef, car il lui manque

---

[71] *Ibid.*, p. 67.

[72] Searle, John R., « What Your Computer Can't Know », *The New York Review of Books*, vol. 61, no 15, 9 octobre 2014.

précisément une intentionnalité propre. Il pourra bien exécuter des ordres, mais sans passer par une médiation psychologique consciente.

II.2.5.1. Critique de la chambre chinoise par Dennett

Cet argument de la chambre chinoise n'a évidemment pas manqué de faire couler beaucoup d'encre dans le débat philosophique. Parmi les auteurs que nous avons abordés, nous allons désormais présenter les critiques qui ont été formulées à l'égard de cet argument phare de Searle en commençant par Dennett puis par Chalmers.

Dans son célèbre ouvrage *La conscience expliquée*, Dennett reprend l'argument de la chambre chinoise formulé par Searle[73]. Il propose alors à son lecteur une expérience de pensée en imaginant une conversation organisée selon les modalités du test de Turing, entre un juge et un interlocuteur isolé dans une chambre chinoise, tel que décrit par Searle. Le juge entame la discussion en racontant une blague potache sur les Irlandais et leur goût prononcé pour la *Guinness*, dans un contexte où l'un d'entre eux formule un vœu à un génie sorti d'une lampe magique. L'interlocuteur répond alors par une réaction de gêne ou de malaise. Par la suite, le juge révèle que cette blague avait été insérée dans la conversation afin qu'elle soit expliquée par l'interlocuteur situé dans la chambre. Ce dernier s'exécute en livrant une explication très détaillée du ressort humoristique de l'histoire.

Dennett s'appuie sur cette mise en scène pour remettre en cause l'idée selon laquelle une telle réponse pourrait être produite par un simple traitement syntaxique de symboles[74]. Selon lui, un programme suffisamment sophistiqué pourrait bel et bien produire une réponse de ce type, non pas en comprenant réellement, mais en s'appuyant sur la manière dont la question est formulée. Il adopterait une stratégie d'interprétation conditionnée par les formulations reçues. Dans ce contexte, nous pouvons dire que Dennett anticipe déjà en quelque sorte l'importance cruciale des "*prompts*" dans l'interprétation informatique qui deviendra une notion centrale dans le développement des modèles de langage.

[73] Daniel Dennett (trad. Pascal Engel), *La Conscience expliquée*, Paris, Éditions Odile Jacob, 1993, p. 540.

[74] *Ibid.,* p. 542.

Or, selon Dennett, Searle ne nie pas qu'un programme surpuissant puisse un jour exister, mais il refuse d'imaginer concrètement une situation semblable à celle que Dennett décrit. Plus encore, il l'accuse de conduire son lecteur sur une fausse piste. Car si un programme parvient à produire une forme de compréhension, ce n'est pas uniquement grâce à l'homme enfermé dans la chambre, qui n'en serait qu'un engrenage, sans vision d'ensemble ni accès au sens[75].

Dennett pousse plus loin son objection en affirmant que ceux qui adhèrent à l'argument de la chambre chinoise sont, en fin de compte, des dualistes cartésiens. Nous rappelons que c'est une position qu'il combat farouchement dans l'ensemble de son œuvre. Pourquoi ? Parce que le dualisme postule que le cerveau seul ne suffit pas à produire de la compréhension, et qu'il faudrait y ajouter une âme ou un principe immatériel. Or, le philosophe soutient au contraire que la compréhension authentique naît d'un processus complexe, fait d'interactions entre sous-systèmes multiples, présents dans l'organisme, mais inaccessibles à notre conscience réflexive[76]. Ainsi, lorsque Searle est dans sa chambre chinoise, il ne comprend certes pas le chinois, mais il n'est pas le seul acteur du processus. Il oublie, selon Dennett, le rôle du "système" global, que ce dernier désigne par l'acronyme CR (*Chinese Room*). Cette idée du philosophe basée sur une analyse plus globale de la chambre chinoise sera reprise chez Chalmers, comme nous allons le voir plus précisément dans la section suivante.

En conclusion, Dennett reproche à Searle que son expérience de pensée mène à une impasse[77]. À l'inverse, ses propres "pompes à intuitions », c'est-à-dire des expériences mentales conçues pour stimuler la réflexion et qui permettent d'explorer d'autres pistes conceptuelles. Elles ne fournissent pas des preuves au sens strict, mais ouvrent l'imaginaire à des possibilités explicatives inédites. Cela dit, Dennett reconnaît lui-même que ce type d'outil relève davantage de l'art de la philosophie que d'une méthode scientifique rigoureuse[78].

---

[75] *Ibid.,* p. 543.

[76] *Ibid.,* p. 544.

[77] *Ibid.,* p. 545.

[78] *Ibid.,* p. 546.

II.2.5.2. Critique de Chalmers ou le malentendu sémantique de la chambre chinoise

Dans un premier temps, il est important de rappeler que Chalmers partage le constat initial de Searle : l'homme enfermé dans la chambre chinoise ne comprend effectivement pas le chinois. Cependant, il estime que l'argument de la chambre chinoise échoue à démontrer que les machines ne peuvent être conscientes ou comprendre quoi que ce soit. La raison est on ne peut plus simple : Searle confond la compréhension du système global avec celle d'un de ses composants isolés. Ainsi, comme l'écrit Chalmers :

"Proponents of strong AI have typically replied by conceding that the demon does not understand Chinese, and arguing that understanding and consciousness should instead be attributed to the system consisting of the demon and the pieces of paper"[79].

Autrement dit, le système dans son ensemble, composé de l'homme, des règles, des symboles ainsi que leur organisation causale, est seul élément pertinent pour juger de l'existence éventuelle de la conscience. Chalmers fait le parallèle avec le cerveau humain : aucun neurone seul ne comprend quoi que ce soit, mais c'est l'organisation causale d'ensemble qui produit la compréhension. Autrement dit, il ne faut pas envisager l'expérience de la chambre chinoise comme un simple individu qui jongle et manipule des symboles abstraits, mais plutôt comme un individu qui est inséré au cœur d'un système plus vaste et dont chaque manipulation produit des conséquences dans l'ensemble.

Chalmers rejoint cependant Searle sur un point fondamental : un simple programme abstrait, écrit sur papier ou stocké dans une mémoire, ne saurait produire par lui-même une compréhension intrinsèque. Cependant, il s'en distingue radicalement sur ce qui permet à un système d'accéder à la conscience. Pour Chalmers, ce qui compte, c'est l'implémentation effective de ce programme dans un système physique dont l'organisation causale reflète celle d'un système conscient.

Pour illustrer cette idée, Chalmers propose une analogie culinaire. Une recette de cuisine écrite sur papier, en elle-même, ne produit aucun gâteau. Mais lorsqu'un cuisinier suit cette recette

[79] David J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory*, New York, Oxford University Press, 1996, p. 323.

en y associant des ingrédients et du matériel adapté, le système global — composé du cuisinier, des outils et des ingrédients — produit effectivement un gâteau[80]. Il en va de même, selon lui, pour la conscience : le programme abstrait à lui seul ne suffit pas, mais son implémentation ou son processus physique inscrit au sein d'une architecture causale appropriée peut engendrer une expérience consciente.

Chalmers critique également l'intuition qui sous-tend l'argument de la chambre chinoise, en mobilisant ses expériences de pensée des *qualia* s'effaçant et *qualia* dansants. Il montre que, si l'organisation fonctionnelle du système est maintenue, alors l'expérience consciente ne peut disparaître progressivement ni changer subitement sans incohérence logique[81]. En conséquence, un système qui reproduit fidèlement les relations causales d'un cerveau humain doit être considéré comme conscient, même s'il est fait de puces en silicium.

En somme, Chalmers pense que Searle se trompe de cible : ce n'est pas l'exécution locale de règles syntaxiques par un individu (comme l'homme dans la chambre) qui est pertinente, mais bien l'organisation causale globale du système. Si cette organisation reflète celle du cerveau humain, alors la conscience peut émerger, indépendamment du substrat matériel.

Autrement dit, Chalmers reformule le cœur du problème : ce n'est pas la manipulation syntaxique des symboles qui compte en soi, mais la manière dont elle s'intègre dans un système organisé causalement, capable de générer des propriétés émergentes telles que la conscience. Il rejette donc l'argument de Searle en soulignant que celui-ci confond la nature abstraite d'un programme avec les effets réels d'une implémentation physique.

II.2.5.3. Quelle actualité pour l'argument de la chambre chinoise ?

Après avoir examiné les différentes critiques philosophiques à l'égard de ce célèbre argument, il convient désormais de se poser la question de savoir si celui-ci tient toujours à l'ère de l'IA connexionniste. Notons que Searle n'a pas publié d'article sur l'IA après celui de 2015 que nous avons utilisé.

[80] *Ibid.,* p. 327.

[81] *Ibid.,* p. 325-326.

Pour éclairer cette interrogation, il convient de s'attarder sur un aspect fondamental du fonctionnement de l'IA connexionniste : l'apprentissage supervisé ou *training set*. Cette étape est indispensable à la mise en œuvre de l'apprentissage profond. Certes, les réseaux neuronaux artificiels sont capables d'identifier des régularités dans de vastes ensembles de données, mais cette forme d'« apprentissage » ne s'accompagne ni d'intentionnalité, ni de compréhension consciente du contenu traité.

Concrètement, pour rendre ces systèmes performants, on leur présente des exemples (par ex. des images) accompagnés de leur étiquette. Ce processus peut être vu comme un étiquetage : on assigne à chaque donnée une signification que la machine ne fait que reproduire. Au début, le système est très peu fiable et dépend entièrement de l'intervention humaine : ce sont les humains qui choisissent les données, attribuent les bonnes réponses et valident les résultats. L'IA apprend donc dans un cadre strictement défini, sans jamais en sortir.

Ce point est clairement illustré par l'article de LeCun, Bengio et Hinton paru en 2015[82]. Les auteurs décrivent l'apprentissage supervisé comme un processus visant à minimiser une fonction d'erreur entre la sortie produite et la sortie attendue, en ajustant automatiquement des millions de paramètres internes par rétropropagation. Il s'agit d'un processus purement computationnel et statistique : les poids sont modifiés en réponse à des signaux d'erreur, sans qu'aucune compréhension des contenus ne soit impliquée.

« The most common form of machine learning, deep or not, is supervised learning. Imagine that we want to build a system that can classify images as containing, say, a house, a car, a person or a pet. We first collect a large data set of images of houses, cars, people and pets, each labelled with its category. During training, the machine is shown an image and produces an output in the form of a vector of scores, one for each category. We want the desired category to have the highest score of all categories, but this is unlikely to happen before training. We compute an objective function that measures the error (or distance) between the output scores and the desired pattern of scores. The machine then modifies its internal adjustable parameters to reduce this error. These adjustable parameters, often called weights, are real numbers that can be seen as 'knobs' that define the input–output function of the machine. In a typical deep-learning system, there may be hundreds of

[82] LeCun, Yann, Bengio, Yoshua, et Hinton, Geoffrey, « Deep Learning », *Nature*, vol. 521, 2015, p. 436-444. https://doi.org/10.1038/nature14539

millions of these adjustable weights, and hundreds of millions of labelled examples with which to train the machine »[83].

L'analyse du fonctionnement de l'IA connexionniste met en évidence sa dépendance à un cadre conceptuel et technique défini par l'humain. Cette dépendance renforce les critiques adressées par Searle à l'IA forte : les systèmes actuels, bien qu'efficaces dans certaines tâches, ne disposent ni de compréhension authentique, ni de conscience. Leur performance repose entièrement sur l'intervention humaine, qui demeure centrale en tant que source du sens projeté sur des opérations purement syntaxiques.

Bien que l'argument de la chambre chinoise ait été formulé dans les années nonante du temps où il n'existait qu'une IA symbolique, force est de constater que cet argument de la cécité sémantique résonne encore de nos jours. Ainsi, la question de la cécité sémantique décrite par Searle à travers cet argument trouve un écho dans les analyses plus récentes produites par Andler sur le *deep learning*. Chez chacun des deux, la machine n'est pas capable d'accéder au sens de ce qu'elle traite, car elle reste tributaire de l'intervention humaine pour fixer l'interprétation.

De son côté, Andler présente, comme nous l'avons vu dans la première partie de ce travail, ce qui constitue pour lui des limites structurelles de l'IA connexionniste : les attaques adversiales, le *data shift* et l'*underspecification.* Malgré une architecture fondée sur l'empirisme et l'apprentissage inductif, cette IA reste elle aussi victime de cette forme d'opacité[84]. Elle peut effectivement reconnaître des motifs, des corrélations ou générer du langage, mais sans nécessairement en comprendre le sens, comme l'attestent les attaques adversariales, qui démontrent à quel point ces systèmes peuvent être trompés par des modifications invisibles à l'œil humain. Une infime modification peut conduire la machine a donné un résultat faussé, alors que l'humain ne s'y tromperait pas.

Dans le cas du *data shift*, on change légèrement la forme de données d'entrée, que ce soit via la formulation, le contexte culturel,…) et les performances des LLM chutent, même si le sens

---

[83] *Ibid.*, p. 436.

[84] Daniel Andler, *Intelligence artificielle, intelligence humaine : la double énigme*, Paris, Gallimard, « NRF Essais », 2023, p. 131.

reste le même. Il y a donc un décalage entre les données d'entrée et son utilisation pendant l'entraînement.

Pour ce qui est de l'*underspecification*, plusieurs modèles différents peuvent expliquer les mêmes données d'entraînement, mais ces derniers se comportent différemment dans des situations nouvelles. Les modèles vont alors se focaliser sur des éléments de l'image qui ne sont pas strictement en lien avec ce qui leur est demandé d'identifier. Par exemple, lorsqu'un hôpital utilise le modèle pour identifier si un organe est touché ou non par une maladie, celui-ci va porter son attention sur d'autres informations présentes sur l'image, mais qui sont hors-sujet par rapport à la demande initiale (comme le logo de l'hôpital, etc.).

En somme, bien qu'issu du contexte de l'IA symbolique, l'argument de la chambre chinoise de Searle reste encore d'actualité à l'ère du *deep learning*. Les architectures connexionnistes, malgré leur puissance statistique, sont encore incapables d'accéder à la signification réelle des données qu'elles traitent. Leurs performances restent tributaires de l'intervention humaine qui leur est externe. En effet, leurs performances qui restent fondées sur des opérations syntaxiques sans sémantique propre. Si nous pouvons parler d'un « sens », nous dirions qu'il s'agit surtout de celui qui se dégage à partir de statistiques probabilistes et corrélatives.

II.2.6. L'argument neuro-biologique

Dernier argument phare de Searle dans lequel il nous invite à réaliser une expérience à la première personne[85] : le lecteur doit se pincer et ensuite, il doit s'interroger sur ce qui vient de se produire chez lui. Il donne trois grilles de lecture à cette expérience : la première, c'est celle des neurobiologistes qui expliquent le parcours de la sensation à travers ses récepteurs sensoriels et son circuit nerveux pour finalement atteindre dans le thalamus. La deuxième, c'est simplement le fait d'avoir ressenti une douleur dont il serait inutile de convoquer un tiers professionnel pour nous expliquer de quoi il s'agit, car cette sensation se manifeste sous le mode d'existence à la première personne. Searle associe donc cette sensation à ce qu'il appelle un *qualia*. Enfin, troisièmement, le lecteur a acquis une nouvelle disposition comportementale qui lui permet de témoigner que celui-ci a senti un léger pincement.

[85] John R. Searle, *Le mystère de la conscience*, Paris, Éditions Odile Jacob, 1999, p. 104.

Searle précise que la douleur, bien qu'elle soit un exemple frappant, n'est pas représentative de toutes les expériences subjectives. Le cours de la pensée consciente, par exemple, ne se manifeste pas toujours de la même manière. Néanmoins, toutes partagent un point commun : une subjectivité ontologique. Le douleur que je ressens est celle qui m'appartient : le cerveau cause des états conscients qui se manifestent sous forme de qualités subjectives internes.

Searle insiste sur le fait que sans l'apport la subjectivité, la science n'aurait de support pour produire des recherches[86] Or, la douleur est un sentiment qui bien que subjectif, est expérimentée par d'autres personnes. Dès lors, nous ne pouvons pas dire qu'il s'agit d'un phénomène isolé qui échappe à la science. La méthodologie pratiquée par la science vise bien entendu l'objectivité épistémique, mais cela ne veut pas dire qu'elle nie pour autant l'objectivité ontologique de la matière étudiée par cette dernière. Le « donné » est là, que l'on veuille ou non, et il sort du cadre des idées pré-conçues.

Ainsi, la science doit expliquer le monde à partir de phénomènes qui, bien que subjectifs ontologiquement, sont parfaitement intégrables à travers une méthode rigoureuse. Écarter cette possibilité reviendrait à renoncer à comprendre l'expérience humaine et à réduire toute connaissance à un point de vue externe, désincarné et abstrait. Or, éprouver une douleur est un fait réel, palpable, dont nous pouvons témoigner, et qui appartient pleinement à la réalité du monde vécu.

Searle défend donc une approche neuro-biologique de la conscience. Le calcul à lui seul ne peut tenir compte de ce qui s'y passe. Il cherche à réhabiliter la conscience comme un phénomène biologique réel. Et pour qu'un cerveau artificiel soit mis sur le même pied d'égalité qu'un cerveau organique, il faudrait que le premier produise de la conscience à travers des processus causaux identiques.

Dennett et Searle expriment chacun un désaccord profond quant à la nature de la conscience. Si l'on veut comprendre ce clivage, il faut revenir un instant sur les *qualia*, ces expériences subjectives, internes et vécus à la première personne. En fait, l'opposition entre Dennett et Searle

[86] *Ibid.,* p. 120.

sur la possibilité d'une machine pensante repose en grande partie sur leur désaccord à propos des *qualia*, ces contenus subjectifs de l'expérience consciente.

Chez Dennett, ces *qualia* n'existent pas réellement. Dans *La conscience expliquée*, il les décrit comme un concept flou et problématique[87]. Il reprend notamment l'idée développée par Ornstein et Thompson en 1984 selon laquelle la couleur n'existe pas en soi, mais résulte d'une interaction entre le cerveau, l'œil et l'environnement. Ainsi, la qualité d'une couleur dépend entièrement du regard de l'observateur[88]. Sans observateur, il n'y a pas de couleur « en soi ».

Searle, au contraire, voit les *qualia* comme constitutifs de la conscience. Pour lui, une machine peut imiter des comportements, mais sans vécu subjectif, il n'y a pas de pensée authentique. Toute IA relève de la syntaxe, là où l'esprit humain donne du sens dans la sémantique. La conscience ne se réduit pas à une simulation : elle est causée par des processus biologiques concrets. C'est un phénomène subjectif et enraciné dans un corps vivant.

Malgré tout, Searle estime que le domaine de la neuroscience ne dispose pas d'une théorie unificatrice qui permette de nous expliquer tout ce qu'il se passe dans le cerveau. Il conclut que le mystère de la conscience ne sera dissipé que lorsque nous disposerons d'une biologie de la conscience aussi robuste que la biologie de la vie elle-même[89]. Cette approche permettrait de dépasser le vieux clivage entre mécanistes et vitalistes : les premiers cherchent à expliquer la vie par la chimie et la mécanique, tandis que les seconds postulent un élan vital mystérieux. Pour Searle, la conscience n'a pas besoin d'un "élan vital", mais d'une explication biologique rigoureuse qui encore à construire.

Cet argument neuro-biologique, Searle le maintient encore dans son dernier article paru en 2015. Il ajoute que que des tentatives ont été faites pour simuler un cerveau humain, mais il rappelle qu'il ne faut pas mettre sur le même plan la simulation et la duplication. Pour illustrer son propos, il écrit que simuler la digestion sur un ordinateur ne permet pas de digérer un repas, de même qu'un modèle numérique de neurones ne produit pas une pensée réelle. Il maintient également que le

---

[87] Daniel C. Dennett, *La conscience expliquée*, Éditions Odile Jacob, 1993, p. 457.

[88] *Ibid.,* p. 471.

[89] John R. Searle, *Le mystère de la conscience*, Paris, Éditions Odile Jacob, 1999, p. 208.

cerveau est un organe biologique qui est régi par des principes causaux particuliers que la simple simulation computationnelle ne saurait reproduire.

En ce sens, Andler identifie que l'argument neuro-biologique fait partie des arguments classiquement avancés pour s'opposer à la reproduction des processus cognitifs par des algorithmes[90]. Toutefois, il tempère cette objection : le cerveau peut bel et bien produire des algorithmes, notamment lorsqu'on effectue une opération mathématique complexe comme une multiplication. Il souligne également que les technologies connexionnistes sont capables de simuler certains processus mentaux déjà présents dans le cerveau humain. En effet, comme nous l'avons déjà évoqué dans la première partie de ce travail, l'IA connexionniste fonctionne sous un mode d'apprentissage par expérience qui, par la suite, renforce des connexions neuronales. Dans notre cerveau, ce phénomène se traduit par le fait que certaines voies synaptiques sont renforcées au profit d'autres en fonction de notre propre expérience.

Ainsi, ces technologies ont une configuration sous forme de réseaux de neurones artificiels qui imitent justement les neurones que l'on retrouve dans un cerveau organique. Qu'il s'agisse aussi par exemple de reconnaître des motifs, de catégoriser des objets ou de s'adapter progressivement à son environnement, le cerveau humain (tout comme l'IA connexionniste) s'inscrivent dans un processus d'apprentissage par essai-erreur ou bien par un système de renforcement.

En effet, nous savons que le cerveau humain est particulièrement doué pour identifier des invariants à travers des expériences multiples et variées : il peut reconnaître un chat malgré la multitude de races qui modifient leur apparence. Malgré la diversité des apparences, l'IA parvient elle aussi à dégager de grandes tendances à mesure qu'elle s'entraîne à partir d'un grand stock d'images.

Cependant, Andler souligne que les projets actuels visant à reproduire les processus humains cérébraux sont encore balbutiants et leur résultat incertain[91]. En revanche, ce type de projet nous interroge pour ce qui est de savoir s'il nous permettra d'acquérir une meilleure connaissance du cerveau, de l'intelligence et s'il constitue ou non des systèmes d'AI générale.

---

[90] Daniel Andler, *Intelligence artificielle, intelligence humaine : la double énigme*, Paris, Gallimard, « NRF Essais », 2023, p. 216-218

[91] *Ibid.,* p. 312.

D’autres penseurs comme Landgrebe et Smith[92] repris par Fjelland sont en revanche beaucoup plus catégoriques quant à la question d’émuler un cerveau humain de manière artificielle. Selon eux, il est vrai qu’il existe des modèles descriptifs de sous-systèmes biologiques ou bien des modèles prédictifs de tout petits sous-systèmes, mais que, même de nos jours, nous ne sommes pas en mesure de concevoir un modèle mathématique du plus vieil organisme vivant comme l’*archaeum*, qui est un micro-organisme unicellulaire procaryotes. Ceci s’explique par le fait ce système biologique fait entrer en jeu plus de 100000 biomolécules. À coté de cela, la simulation artificielle d’un cerveau humain demeure pour eux techniquement hors de portée.

Andler estime pour sa part que seule cette branche de la recherche pourrait apporter un éclairage sur les mécanismes biologiques à l’oeuvre lorsque l’intelligence est sollicitée. Cette branche peut constituer une piste dans la création du IA forte de type prométhéenne. Comme pour Searle, tout deux affirment que notre état de la connaissance en matière biologique est encore insuffisant pour cerner correctement le phénomène de la conscience. À ce jour, personne n’est encore parvenu à reproduire artificiellement un cerveau. Par conséquent, leur raisonnement consiste à dire que puisque comme la conscience émerge d’un processus neuro-biologique complexe et qu’un cerveau artificiel n’a pas encore été conçu, les machines ne peuvent donc pas être conscientes.

Conclusion intermédiaire

Relecture des critiques philosophiques à l’aune du *deep learning*

Quelles leçons pouvons-nous tirer à ce stade de notre enquête ? Tout d’abord, il devient nécessaire d’interroger la pertinence des critiques philosophiques adressées à l’intelligence artificielle à l’aune du *deep learning*. D’autre part, est-ce que les performances actuelles de cette IA permettent-elles de dépasser les limites soulignées par ces critiques ?

À bien des égards, Turing se montre précurseur en évoquant l’objection de Lady Loveface qui consistait à dire une machine ne pourrait rien produire qu'on ne lui aurait préalablement ordonné. Cette objection perd aujourd’hui de son poids face aux capacités démontrées par les IA

[92] Fjelland, Ragnar, « Computers Will Not Acquire General Intelligence, but May Still Rule the World », *Cosmos+Taxis*, vol. 12, no 5-6, 2024, p. 63.

connexionnistes. L'exemple d'*AlphaGo* illustre cette mutation : l'agent n'a pas simplement reproduit des stratégies humaines, mais en a trouvé de nouvelles, issues de son propre apprentissage.

De plus, nous pouvons reconnaître à Turing comme à Dennett le mérite d'avoir anticipé l'apparition de l'IA connexionniste à travers leurs expériences de pensée. Le premier l'a exprimé sous le terme de learning machine, tandis que le second l'a imaginé à travers le robot WUNDERKIND. Leur objectif était de critiquer la vision selon laquelle l'être humain posséderait une faculté supérieure inaccessible aux machines

Nous remarquons également que les limites d'hier associées à l'IA et relevées par Dennett — à savoir l'ambiguïté linguistique et la connaissance du monde — tendent à être dépassées à mesure que la technologie se développe. Les études mises en évidence montrent que les derniers modèles de LLM sont de plus en plus performants pour surmonter la difficulté posée par des phrases ambiguës. Il n'est donc pas insensé d'imaginer qu'un entraînement accru rapproche encore ces systèmes des compétences humaines en la matière.

De même, les systèmes-experts sur lesquels Dennett s'exprimait ne disposaient pas à l'époque du formidable puits de connaissances dont héritent aujourd'hui les IA grâce au *Big Data*. C'est sur ce point qu'il revient dans son article de 2019 : les IA peuvent désormais acquérir un stock gigantesque d'informations, les traiter et en retirer des données exploitables. Naturellement, se pose ensuite la question de l'usage que le pouvoir peut faire de ces informations, dimension éthique que nous traiterons dans la dernière partie de ce travail.

Les arguments de Searle sont-ils encore pertinents ?

Sur le plan philosophique, nous avons vu en quoi l'argument de la chambre chinoise, visant l'IA symbolique, pouvait être remis en question. En effet, la personne isolée dans la chambre ne connaissait pas le chinois, mais Dennett et Chalmers font remarquer que l'analyse de Searle se focalise sur cette personne, et non sur le système dans lequel elle opère. Leur changement de perspective invite à reconsidérer le présupposé de l'absence de connaissance sémantique attribuée à cette personne.

Cependant, cet argument a constitué un pivot de la thèse de la cécité sémantique des machines et conserve une part de pertinence : l'IA connexionniste, comme nous l'avons vu avec Andler, montre des limites structurelles liées à la sémantique, telles que les attaques adversariales, le data shift ou l'underspecification. En somme, il subsiste toujours un décalage entre les données traitées par cette IA et leur application dans notre monde, complexe et changeant, encore difficile à appréhender pour la machine. L'exemple le plus parlant reste celui des hallucinations, où l'IA générative produit une réponse inventée de toutes pièces mais présentée comme un fait, sans « savoir » pourquoi elle se trompe.

En outre, un autre argument phare de Searle était de dire que les informations stockées dans l'IA symbolique dépendent de l'œil de l'observateur et n'avaient de sens que pour ce dernier alors que la machine suivait « aveuglément » les règles qu'on lui avait encodées. Cet argument possède encore une certaine actualité dans le cas de l'IA connexionniste, car nous avons montré qu'au début de son entraînement, elle a nécessairement besoin du guidage humain car elle se trompe souvent dans un premier temps. En conséquence, c'est encore l'humain qui donne du sens aux opérations exécutées par la machine, malgré le fait que cette dernière apprend par la suite d'elle-même. Plus proche de nous en 2023, Landgrebe et Smith[93] reconnaissent la pertinence actuelle de l'argument de Searle en précisant que ce qu'il se passe dans une machine n'a pas de sens pour un observateur non-humain : le sens n'existe qu'à travers le regard de l'observateur humain.

Par la même occasion, nous notons que les IA génératives restent dépendantes des données sur lesquelles elle s'entraîne et donc du choix des ingénieurs d'inclure telle ou telle source. Ce qui a pour conséquence que si leur base d'entraînement contient des données incomplètes ou biaisées, cela aura évidemment un impact néfaste sur ses résultats. Ainsi, elle est incapable de produire de nouvelles connaissances authentiques qui sortent du cadre de sa base d'entraînement. Nous pouvons dire que sa créativité n'est en fait qu'une recréation des corpus auxquels elle a accès, que ce soit du texte ou des images.

Enfin, pour ce qui est de l'argument neuro-biologique amené par Searle, il est pertinent de faire remarquer qu'il n'existe à ce jour encore aucun système artificiel qui soit parvenu à reproduire la complexité du cerveau. Bien qu'il s'agisse selon Andler d'une branche de recherche qui pourrait

---

[93] Landgrebe, Jobst, et Barry Smith, « Why Machines Do Not Understand: A Response to Søgaard », arXiv.org, juillet 2023. DOI : 10.48550/arXiv.2307.04766.

nous apporter davantage de compréhension quant aux mécanismes biologiques à l'oeuvre lorsque nous sollicitons notre intelligence, ce dernier précise tout de même que la reproduction des processus cérébraux humains reste encore un domaine incertain même si les chercheurs ambitionnent de fabriquer une « super-intelligence » qui surpasserait nos capacités.

Un mode de fonctionnement davantage basé sur des corrélations statistiques que sur la causalité

Cela dit, il ne faudrait pas non plus sous-estimer les capacités de l'IA générative à produire des contenus bluffants à partir des régularités qu'elle parvient à dégager sur base de son entraînement. Elle peut en effet reconnaître des motifs, des corrélations ou générer du langage, mais sans nécessairement en comprendre le sens. Par rapport à ce point, nous avons besoin de rappeler ce qui fait la spécificité de ce type d'IA : son mode de fonctionnement se fonde sur une génération par probabilité conditionnelle, c'est-à-dire qu'elle prédit la suite la plus probable d'un texte à partir d'un contexte donné. Par conséquent, ces modèles génèrent effectivement une apparence de compréhension et donnent l'impression à son interlocuteur humain qu'elle « comprend » ce qu'elle dit. Ce constat est également partagé par Dennett dans son article de 2019[94] : les IA génératives basées sur le *deep learning* parviennent à exploiter une somme d'informations considérable sans avoir besoin de la comprendre.

Fjelland[95] rappelle que bien que les machines soient performantes lorsqu'il s'agit de repérer des régularités statistiques, elles n'apprennent uniquement que par corrélation. Effectivement, comme il le fait remarquer, la corrélation peut par exemple nous aider à constater certaines choses, comme par exemple nous aider à prédire une épidémie de grippe sur base des recherches google effectuées par des personnes qui en sont atteint ou qui pensent l'être, mais nous n'aurions pas d'explications sur la cause de l'épidémie lorsqu'il s'agit de savoir de quelle manière le virus s'est propagé.

---

[94] Dennett, Daniel C., « What Can We Do? », in J. Brockman (dir.), *Possible Minds: Twenty-Five Ways of Looking at AI*, New York, Penguin Press, 2019.

[95] Fjelland, Ragnar, « Why General Artificial Intelligence Will Not Be Realized », *Humanities & Social Sciences Communications*, 7(1):10, 17 juin 2020, p. 5-6.

Pour donner un autre exemple cette fois-ci tiré de la prévention de la criminalité[96], les corrélations peuvent effectivement attirer et guider les autorités sur les zones à risques dans lesquelles il serait judicieux d'intervenir. Cependant, cet outil ne permettrait pas d'étudier les causes de cette criminalité. Il n'est pas possible pour les autorités de se passer des renseignements issus du terrain sur base notamment de l'enquête de voisinage. Sans cette dimension ancrée dans la réalité, les autorités pourraient passer à côté d'informations déterminantes et utiles qui permettraient de recourir par la suite à d'autres stratégies de prévention.

Fjelland dénonce ainsi la vision de certains ingénieurs qui pensent que l'accumulation de données rend inutile l'explication d'un phénomène. Il y voit un appauvrissement épistémologique, qui écarte la dimension explicative et compréhensive de la science. Comprendre le « pourquoi » est essentiel. Les IA génératives ne construisent pas de représentations causales du monde : elles restent au niveau de la corrélation statistique, là où l'humain raisonne causalement.

Plusieurs positions philosophiques à distinguer

Pour conclure cette deuxième partie, nous proposons une classification des philosophes de l'esprit selon leur position sur l'IA faible et l'IA forte. Aucun d'entre eux ne soutient que les machines actuelles sont conscientes. Mais certains défendent la thèse de l'IA forte, c'est-à-dire la possibilité future d'une conscience issue de l'informatique.

Nous distinguons ainsi trois grandes visions sur la question de l'IA et de son développement. Pour commencer, nous associons Turing, Dennett et Chalmers dans la catégorie des penseurs qui estiment que bien que l'IA de leur temps présentait des limites, celles-ci sont avant tout liées au développement de notre technologie. À leur sens, ce développement va potentiellement aboutir à l'apparition d'un esprit conscient et que nous serons amenés à reconsidérer notre conception de la conscience.

Cette pensée s'explique de plusieurs façons : tout d'abord, Turing suppose qu'une équivalence fonctionnelle comportementale conversationnelle suffit pour démontrer de l'intelligence. Si une machine montre qu'elle est capable d'imiter le comportement humain de sorte

[96] *Ibid.*, p. 2.

à tromper son interlocuteur en chair et en os, alors il considère que la puissance de calcul pourrait profondément changer notre conception de la conscience et que de ce fait, il n'y aura potentiellement plus une spécificité ou une supériorité humaine face aux machines.

Dans le même ordre d'idée, Dennett estime à travers sa vision computationnelle de l'esprit qu'une machine peut être considérée comme consciente si elle simule correctement les dynamiques cognitives humaines. Sa vision de la conscience, basée sur le modèle des versions multiples, ouvre la porte à la conception selon laquelle la conscience est assimilée à un programme d'ordinateur. Par conséquent, il défend l'idée selon laquelle les machines pourraient un jour développer une conscience. Sa vision est entre autres rendues possible par sa critique des *qualia,* ces expériences subjectives, internes et vécus à la première personne qui ne sont pas selon lui la preuve d'une conscience.

Enfin, nous avons Chalmers, qui, à travers sa démonstration de l'invariance organisationnelle, défend l'idée que les fonctions cognitives peuvent être reproduites dans une machine. Une même organisation fonctionnelle aboutit aux mêmes états mentaux et ce, peu importe leur substrat de base (biologique ou non). La conscience artificielle est donc en principe possible à condition que l'IA respecte les mêmes structures fonctionnelles que l'humain. Comme nous l'avons vu, l'IA connexionniste s'inspire effectivement de l'organisation neuronale de notre cerveau, mais selon Chalmers, nous ne pouvons pas encore affirmé que les LLM qui bénéficient de cette technologie soient actuellement pourvus d'une conscience artificielle. Il leur manque encore certaines composantes et il estime qu'il est possible qu'elles soient atteintes dans un futur proche.

Du côté de Searle, nous l'associons à ceux qui estiment qu'il est et restera impossible en tout temps que la machine soit dotée d'une conscience. Pour lui, les *qualia* sont indissociables de la conscience. La machine peut imiter des comportement humains, mais elle ne manifeste pas un vécu subjectif. La conscience ne peut pas se résumer à une simple simulation, elle est au contraire le fruit de processus biologiques concrets qui sont trop complexes pour que la machine puissent les reproduire.

Comme il le dit dans son article de 2014[97], pour comprendre ce qu'est une démangeaison, il ne suffit pas d'en avoir une description, mais de savoir ce que ça fait réellement d'en avoir une. Cette réalité est de l'ordre de ce qui est « ontologiquement subjectif ». De plus, comme les machines ne manifestent pas un vécu subjectif, elles ne possèdent pas non plus de croyances, de désirs ou de motivation qui pourraient les guider dans leurs choix. Par conséquent, il estime qu'il n'y a pas de raison d'imaginer qu'une entité informatique autonome puisse un jour nous dominer.

Son article est aussi une occasion pour lui de réaffirmer que peu importe leur complexité, les machines ne pensent pas en l'absence de conscience. L'apparence ne suffit pas à simuler la conscience, car celle-ci est une réalité vécue, intérieure, irréductiblement subjective. Les partisans de l'IA forte, selon lui, commettent une erreur fondamentale en confondant la simulation d'un processus avec le processus lui-même.

Dans la troisième partie de ce travail, nous allons tenter de développer une vision qui propose plutôt une troisième voie. C'est-à-dire une voie qui reconnaît des limites constitutives et structurelles à la technologie actuelle, mais qui ne ferme toutefois pas la porte à un autre type de technologie qui dépasserait ces dernières à l'avenir. En effet, l'IA connexionniste ne fut pas anticipée du temps de l'IA symbolique et elle est apparue avec son lot de surprises. Enfin, nous estimons que l'avancée technologique se poursuivra immanquablement et que les ingénieurs qui y travaillent auront toujours à l'esprit et comme horizon de concevoir une technologique tentant de se rapprocher de l'intelligence humaine.

PARTIE 3 : Est-ce que les machines pensent ? Corps, monde et jeu du langage

Une fois tous ces arguments posés et discutés, nous pouvons désormais esquisser une réponse à l'une de nos interrogations de base. Il est indéniable qu'à la fois les machines utilisant l'IA symbolique et l'IA connexionniste démontrent une pensée. De fait, comme il a été rappelé, nous recourons tous les jours à la déduction ou l'induction pour conduire nos vies.

---

[97] Searle, John R., « What Your Computer Can't Know », *The New York Review of Books*, vol. 61, no 15, 9 octobre 2014.

Comme pour les machines, dans notre vie de tous les jours, nous sommes amenés à résoudre des problèmes. Lorsque nous souhaitons par exemple organiser un voyage, il nous faut d'abord réfléchir à toutes les possibilités qui s'offrent à nous et à sélectionner celles qui nous conviennent le mieux. Ainsi, plusieurs facteurs entrent en compte : le budget alloué, le moyen pour parvenir à telle localisation, le nombre de vêtements à prendre en tenant compte du nombre de jours passé sur place, le choix d'un logement, etc. Cette façon de procéder, il est possible de l'encoder dans un programme d'ordinateur et l'IA symbolique fait parfaitement l'affaire.

Dans le cas de l'IA connexionniste, qui elle, fonctionne à travers l'induction, nous recourons aussi tous les jours à ce mode de pensée. L'un des exemples les plus évidents à citer est bien celui par exemple de l'apprentissage social. Tout au long de nos vies, nous apprenons que certains types de comportements sont plus adéquats que d'autres dans tel type de situations sociales. Cet apprentissage est naturellement le fruit de notre expérience à travers nos essais, nos succès et bien sûr nos erreurs.

De même, sur base de que nous voyons, nous avons appris à pouvoir distinguer un chat d'un chien et inversement. Nos expériences de vie nous ont offert cette connaissance à force d'en croiser et d'interagir avec. Là où l'IA connexionniste est effectivement plus puissante que nous, c'est parce qu'elle dispose de bases de données considérables qu'un seul être humain ne pourrait emmagasiner à lui seul.

Mais est-ce que cela suffit à mettre sur le même plan l'intelligence humaine et l'intelligence artificielle, bien que celle-ci emprunte des mécanismes issus de la première ? Certes, l'IA connexionniste, à l'aide de son réseau artificiel neuronal, a eu l'ingéniosité d'imiter un mécanisme cérébral connu, mais est-ce que ça veut dire que nous avons réussi à reproduire ce qui se passe dans le cerveau humain ?

En tout cas, à l'heure actuelle, personne n'est encore parvenu à reproduire un cerveau artificiel, car l'architecture organique est encore trop complexe. Même s'il s'agit toujours d'un projet sur lequel des scientifiques se sont engagés. Idem, lorsque l'on essaye de jumeler une IA dernier cri avec un corps robotique, il est manifeste que l'IA possède un corps, mais qu'elle n'est pas son corps. Ainsi, plusieurs éléments indissociables de la corporéité lui échappent : d'une part,

les contraintes liées à l'anatomie, d'ordre physiologique, et de l'autre, la spontanéité, l'exploration libre de l'espace et du mouvement[98].

C'est comme si l'IA dans ce cas-ci servait d'un centre de commandes qui donnait des instructions à une autre entité qui n'est pas la sienne. Il y a comme une réactivation d'une vision mécaniste et dichotomique où l'on retrouve d'un côté l'esprit sous forme d'IA, et de l'autre, son corps. Comme le robot ne possède pas de véritable « soi », nous doutons comme Andler qu'il ne suffit pas de lui injecter des capacités mobiles ou motrices pour en faire une entité autonome.

À l'instar de ce que dit Andler, nous pensons qu'il ne faut ni être naïf, ni sous-estimer les progrès en IA. Est-ce que les premiers chercheurs en IA dans les années cinquante s'attendaient-ils à l'apparition des LLM ? L'IA connexionniste a en tout cas ravivé la flamme de l'intérêt scientifique et public pour le domaine de l'intelligence artificielle à travers ses performances impressionnantes.

Tous ces progrès n'auraient pas été possibles sans l'intervention de l'ingéniosité humaine qui peut dans certains cas sortir du cadre et proposer un produit révolutionnaire. Cependant, cet argument, comme le dit Andler[99], peut être aussi à double-tranchant dans le sens où cette spécificité humain pourrait aussi amener à la création d'une super-intelligence de type prométhéen, flexible, autonome et surpassant les capacités humaines. Nous ne sommes pas à l'abri d'un tel exploit, mais force est de constater que nous n'y sommes pas encore.

C'est pourquoi il nous paraît important à présent d'explorer une troisième voie qui se situe entre la thèse de l'IA faible et celle de l'IA forte. Nous pensons effectivement que les avancées de l'IA ont permis de ré-interroger des arguments philosophiques que nous avons exploités plus haut, mais nous pensons aussi qu'il y existe au moins trois dimensions propres à l'intelligence humaine auxquelles l'IA actuelle ne peut avoir accès.

Cette partie du travail sera donc consacrée à expliciter et explorer ces trois dimensions que sont : la corporéité de l'intelligence, la notion de monde (et d'*Umwelt*) et le jeu du langage. Toutes

---

98 Daniel Andler, *Intelligence artificielle, intelligence humaine : la double énigme*, Paris, Gallimard, « NRF Essais », 2023, p. 182.

99 Ibid., p. 327-328.

ces dimensions renvoient chacune à la question du rapport au monde que nous entretenons et que les machines ne peuvent atteindre.

Pour nous aider dans le développement de cette partie, nous ferons entre autres appel au maître à penser d'Andler qui n'est d'autre qu'Hubert Dreyfus, particulièrement connu pour ses ouvrages qui explorent la question de savoir ce que les ordinateurs ne peuvent pas faire. Nous verrons que certains de ses arguments seront repris et reformulés par son fils spirituel qu'est Andler. De même, en filigrane de cette partie, il est également question de faire appel à la phénoménologie ainsi qu'à des éléments issus de la philosophie du langage à travers Wittgenstein.

En somme, s'il est possible de reconnaître que les machines reproduisent certains processus cognitifs que l'être humain mobilise dans sa vie quotidienne, il n'en demeure pas moins que la pensée proprement humaine s'inscrit quant à elle dans une dynamique incarnée. Celle-ci implique un corps vécu, vecteur de notre compréhension du monde, de notre rapport aux autres ainsi que de notre capacité à nous adapter à des situations concrètes et parfois imprévisibles.

III.1. L'IA n'a pas de corps propre

Assez intéressant pour le faire remarquer, dans son article de 2019[100], Dennett ne parle pas du corps à proprement parler, mais il évoque le fait que contrairement à d'hypothétiques agents conscients artificiels, nous ne partageons pas avec eux notre vulnérabilité ainsi que notre mortalité. Or, ces deux composantes sont essentielles si l'on veut être assimilé à un agent moralement responsable. En effet, comment pourrait-on faire signer un contrat contraignant à une machine invulnérable ? Cette idée de contrainte est rendue possible pour nous car les conséquences nous impactent personnellement, que ce soit être condamné à payer une amende ou à être enfermé dans une prison. La technologie des machines actuelles les rend donc insensibles aux conséquences de leurs actes. Par conséquent, elles ne sont pas traversées par des affects tels que l'amour ou la haine ni dotées d'une personnalité particulière.

[100] Dennett, Daniel C., « What Can We Do? », in J. Brockman (dir.), *Possible Minds: Twenty-Five Ways of Looking at AI*, New York, Penguin Press, 2019.

De son côté, comme nous l'avons vu, Chalmers estime dans son article de 2023[101] que les sens et l'incarnation ne sont pas nécessaires pour manifester une conscience ou de la compréhension. Selon lui et en principe, un système désincarné et sans perception sensorielle pourrait néanmoins avoir une forme de conscience cognitive, comme par exemple raisonner sur les mathématiques, réfléchir à sa propre existence ou émettre des croyances sur le monde. En d'autres termes, il effectue un distinguo entre la conscience sensorielle et corporelle encore absente chez les machines et la conscience cognitive qui pourrait se manifester chez des modèles purement linguistiques.

De plus, face à l'objection de Searle et de la chambre chinoise telle que nous l'avons vue, Chalmers suggère que le texte lui-même sur lequel s'entrainent ces systèmes provient du monde sur base d'expériences humaines, de descriptions perceptives, etc. De cette façon, le philosophe australien redonne une légitimé à la réalité virtuelle face à la réalité physique.

En tout cas, comme nous avons pu le voir avec les défenseurs de la thèse de l'IA forte ainsi que leurs arguments, nous pouvons dire qu'il y a une tentation de concevoir l'intelligence humaine comme un ensemble de capacités formalisables et simulables par une machine. Face à cette position, de nombreuses critiques ont émergé et c'est l'occasion pour nous d'introduire l'apport d'Hubert Dreyfus et en particulier sa thèse qui consiste à dire que tout comportement intelligent est intrinsèquement lié au corps, au contexte et surtout d'un rapport direct avec le monde que les machines ne peuvent actuellement pas reproduire.

Dans son ouvrage intitulé *What Computers Still Can't Do*, Dreyfus défend sa thèse selon laquelle un comportement intelligent humain ne peut être reproduit par un ordinateur numérique, car il repose avant tout sur des capacités incarnées qui ne sont pas réductibles à des opérations symboliques ou à des règles explicites. Ces capacités engagent un corps vivant qui habite et évolue au sein d'un monde et non via une entité désincarnée recevant passivement des données.

Dreyfus associe le modèle computationnel à un héritage cartésien du désincarné. Selon lui, l'IA repose sur l'idée selon laquelle le corps peut être exclu du raisonnement et de ce fait, cette conception renvoie à une tradition philosophique qui remonte à Platon jusqu'à Descartes. Même si

[101] Chalmers, David J., « Could a Large Language Model Be Conscious? », *Boston Review*, 9 août 2023.

d'un autre côté, Descartes avait justement fait remarquer que la raison humaine, contrairement aux machines, est capable de s'adapter à un nombre indéfini de situations. En effet, il souligne qu'il existe chez les machines une insuffisance essentielle liée à ses états finis.

> « A brain in a bottle or a digital computer might still not be able to respond to new sorts of situations because our ability to be in a situation might depend, not just on the flexibility of our nervous system, but rather on our ability to engage in practical activity. After some attempts to program such a machine, it might become apparent that what distinguishes persons from machines, no matter how cleverly constructed, is not a detached, universal, immaterial soul but an involved, situated, material body »[102].

Même si les ordinateurs modernes sont dotés d'un nombre mirobolant d'états possibles — et ce n'est pas ces récents développements qui lui donneront tort — ils ne peuvent malgré tout pas faire face à des situations nouvelles, car ils ne sont pas incarnés à travers un corps matériel, ni ancré dans la réalité :

> « With the aid of concepts borrowed from phenomenology, I shall try to show how pattern recognition requires a certain sort of indeterminate, global anticipation. This set or anticipation is characteristic of our body as a "machine" of nerves and muscles whose function can be studied by the anatomist, and also of our body as experienced by us, as our power to move and manipulate objects in the world. I shall argue that a body in both these senses cannot be reproduced by a heuristically programmed digital computer-even one on wheels which can operate manipulators, and that, therefore, by virtue of being embodied, we can perform tasks beyond the capacities of any heuristically programmed robot »[103].

---

[102] Hubert Dreyfus, *What Computers Still Can't Do : A Critique of Artificial Reason*, Cambridge (Mass.), MIT Press, 1992, p. 236.

[103] *Ibid.,* p. 237.

Ainsi, grâce à l'apport de la phénoménologie en lien avec la théorie de la reconnaissance des formes, Dreyfus va donc démontrer que l'humain est capable d'accomplir des tâches qui échappent à n'importe quel robot programmé. C'est à partir de ce constat qu'il envisage d'interpréter l'humain comme à la fois une machine constituée de nerfs et de muscles, mais aussi comme un corps qui relève du vécu.

Les phénoménologues ont montré que notre reconnaissance d'objets spatiaux ou temporels tirés de notre vie quotidienne se fonde sur notre perception de l'ensemble. Pour donner un exemple, Dreyfus porte à notre connaissance le fait que l'on apprécie une mélodie non pas à partir des notes de manière isolées, mais bien en les reconnaissant comme des parties intégrantes d'un tout[104].

Autrement dit, la signification des détails est avant tout déterminée par l'ensemble. Or, les programmes informatiques n'ont pas été capables de reproduire cette interdépendance entre les parties et le tout. Là où les ordinateurs vont par exemple analyser une maison en partant des détails pour aboutir à un « tout », l'humain va au contraire projeter un « tout » puis s'attarder sur les détails.

Dreyfus fait appel ici aux notions d'horizon interne et externe. Par horizon externe, nous entendons l'ensemble des éléments qu'une personne juge pertinents en vue d'une action donnée. Là où par exemple nous voyons une chaise comme un objet pour s'asseoir, la machine l'interprètera comme un amas de pixels qui comprend des mesures et qui rentre dans certaines catégories. Cet horizon externe nous permet ainsi de donner du sens dans l'environnement dans lequel nous vivons.

Pour ce qui est de l'horizon interne, nous entendons tout acte mental qui est dirigé vers un but ou un objet. Or, comme les machines ne disposent pas de subjectivité, ni d'intentionnalité vécue, elles planifient ou prennent des décisions non pas à travers leur propre expérience vécue, mais selon les règles du programme ou d'un apprentissage à partir de données issues de l'expérience humaine.

[104] *Ibid.,* p. 238.

De même, il note que lorsque nous acquérons d'autres compétences telles que conduire, savoir danser ou parler une langue étrangère[105], Dreyfus reconnaît qu'au début, nous commençons par suivre « bêtement » des règles. Mais à force de pratiquer, il se produit un phénomène où nous pouvons agir de manière presque automatique sans trop y penser. C'est donc à ce moment-là que ces règles apprises se logent dorénavant dans notre inconscient et qu'elles font partie intégrante à ce qu'il appelle un *gestalt* musculaire qui fait en sorte que notre comportement manifeste une nouvelle fluidité. En d'autres termes, il s'agit de l'incorporation d'un schème corporel global, souple et adaptatif, mais qui n'est pas formalisable.

Cette idée est également reprise plus tard en 2020 par Fjelland qui cite Polyani[106]. Dans notre vie quotidienne, sans nous en rendre spécialement compte, nous appliquons des règles. Par exemple, de nombreux nageurs ne sont pas capables d'expliquer ce qui les fait rester à la surface au lieu de couler. L'explication existe, il s'agit du contrôle qu'ils exercent sur les poumons, mais peu d'entre eux sont capables de l'expliquer avec précision une fois que la pratique est ancrée. Il en va de même lorsque l'on fait du vélo : beaucoup de personnes sont capables de rouler et de tenir en équilibre dessus, mais ils auraient du mal à expliquer clairement la manière dont ils s'y prennent. Et même si on leur expliquait les phénomènes physiques qui entrent en jeu, cela ne les aiderait pas davantage dans leur pratique car ils savent déjà comment s'y prendre.

Un autre exemple encore plus évident de cette connaissance tacite celui est la marche. La plupart d'entre nous sommes des experts dans cette discipline, mais quand il s'agit de devoir décrire et décomposer les phénomènes musculaires qui entrent en jeu, nous sommes face à une difficulté. Dès lors, comment une machine pourrait-elle reproduire ce phénomène ? Car nous remarquons que tout n'est pas formalisable dans les activités que nous menons dans notre vie quotidienne et parfois même sans que nous nous en rendons compte.

Ce phénomène, Dreyfus l'associe à l'acquisition de la compétence perceptive déjà développée chez Merleau-Ponty. Ce dernier avance que pour apprendre à sentir une matière

---

[105] *Idem.*

[106] Fjelland, Ragnar, « Why General Artificial Intelligence Will Not Be Realized », *Humanities & Social Sciences Communications*, 7(1):10, 17 juin 2020, p. 3.

particulière comme de la soie[107], il faut d'abord apprendre à déplacer sa propre main d'une certaine façon quand on la parcoure. Si nous savons que c'est de la soie, c'est parce que nous la touchons de manière particulière, la sentons et la voyons. Nous nous y prenons pas de la même manière, par exemple, lorsqu'il s'agit de parcourir et comprendre ce qu'est une planche en bois.

C'est de cette façon que nous pouvons par la suite faire la différence entre les différentes matières. Même si avant cela, nous n'avons pas encore des sensations claires, mais bien confuses. Le corps joue donc ici un rôle indispensable dans les sens tels que le goût, l'ouïe ou le toucher et il est aussi vecteur de significations. Cet ensemble de compétences corporelles nous permettent à la fois de reconnaître des objets en particulier, mais aussi de les mettre en relation avec nos compétences exploratoires représentées par nos sens.

Ce rapprochement avec Marleau-Ponty renvoie particulièrement à sa notion du « corps propre » qui postule que l'intelligence humaine est une manière d'habiter le monde par le corps. Il n'est donc pas à confondre avec un simple objet et il traduit d'un être-au-monde particulier[108]. C'est à travers notre corps que nous pouvons percevoir et réaliser des actions immédiates. Or, dans le cas des machines, elles ne disposent pas d'un corps « propre ». Leur savoir est externe et il est issu d'un monde auquel elles ne participent pas. En conséquence, elles ne s'inscrivent pas dans une perpective incarnée. En outre, ce corps propre exerce une équivalence intersensorielle immédiate à travers laquelle nous relions nos perceptions.

Le corps est donc une condition *sine qua non* de sa perception et c'est par son intermédiaire que l'on agit dans un monde. De plus, il contient en lui-même toute une séries d'habitudes qui font sens au sein du même monde. Peut-on alors dire que la machine dotée d'un corps est-il le sien ? Dans le cas de la robotique, l'IA a effectivement un corps, mais elle n'est pas son corps. Elle agit au contraire comme un centre de commandes qui envoie des instructions à une structure qui ne lui appartient pas. Si l'on reprend le notion de Merleau-Ponty sur le corps propre, nous pouvons affirmer que l'IA n'en a pas.

---

[107] Hubert Dreyfus, *What Computers Still Can't Do : A Critique of Artificial Reason*, Cambridge (Mass.), MIT Press, 1992, p. 249.

[108] Maurice Merleau-Ponty, *Phénoménologie de la perception*, Paris, Gallimard, coll. « Bibliothèque des idées », 1945. Rééd. Paris, Gallimard, coll. « Tel », 1976, p. 175

Ce point permet donc à Dreyfus d'affirmer que notre corporéité permet de contourner l'analyse formelle propre aux ordinateurs, dans le sens où ces derniers, face à un objet donné, ils devraient analyser ce dernier à partir de caractéristiques communes pré-enregistrées contenues dans une liste. Or, le corps, quant à lui, fonctionne comme un système synergique : je suis capable, par exemple, de reconnaître une texture avec ma main, mon pied ou mon regard. Et cette configuration si particulière, il n'est pas possible de la formaliser.

Ce faisant, Dreyfus reconnaît que la reconnaissance de formes est une tâche aisée pour les machines à partir du moment où il existe des traits saillants et identifiables[109]. Toutefois, lorsqu'il s'agit de la reconnaissance de formes complexes, cette méthode échoue. C'est pourquoi un phénoménologue comme Merleau-Ponty a pu déterminer quels mécanismes étaient à l'oeuvre lorsque l'humain se trouvait face à des formes complexes. Selon lui, il est question de l'intervention d'un corps actif, organiquement interconnecté et prêt à réagir face à son environnement.

Il est vrai qu'actuellement, si l'on soumet une image connue contenant une illusion d'optique, il y a fort à parier que si elle fait partie de la base de données de l'IA, celle-ci parviendra à la reconnaître et y apporter les différentes grilles de lecture. Mais imaginons maintenant qu'on lui soumette une illusion d'optique inédite, parviendra-t-elle à la reconnaitre et à être trompée de la même façon qu'un humain ? Nous pensons que la seule façon pour elle de la reconnaître doit se faire à partir d'indices visuels pris isolément sur lesquels elle s'est préalablement entrainée. Si c'est le cas à partir de la connaissance que nous avons de son fonctionnement, ce point donne d'autant plus de poids au propos de Dreyfus : l'IA ne parvient donc pas à saisir la réalité à partir de l'ensemble, mais elle tient plutôt sa compréhension une fois les détails analysés isolément puis additionnés ensemble. En tout cas, une expérience comme celle-ci mériterait d'être menée, car nous pourrions dès lors observer la manière dont elle réagit face à des illusions d'optique inédites.

En conclusion, ce chapitre insiste sur le caractère profondément enraciné du comportement intelligent humain dans un corps qui habite le monde et qui y interagit sans devoir à formaliser toutes les données. De même, le corps dispose d'une transférabilité sensori-motrice dans le sens où il a la capacité de transférer une compétence d'un sens à un autre.

---

[109] *Ibid.,* p. 250.

III.2. La compréhension contextuelle échappe à l'IA

Une fois la corporéité posée comme première étape de notre rapport au monde, Dreyfus insiste également sur le rôle fondamental de la situation, ce qui va lui permettre de s'attaquer à l'un des grandes thèses de l'IA forte : que tout comportement humain intelligent serait gouverné par des règles explicites. Si l'on part du principe que chaque fait ne possède pas une signification fixe, il est absolument nécessaire de recourir au contexte pour clarifier les choses dans des situations où celles-ci nécessitent de l'être.

Bien que comme nous l'avons vu, l'IA symbolique peut être très redoutable lorsqu'il s'agit de battre un champion mondial d'échecs, il n'en demeure pas moins que cette compétence reste éloignée de ce que Dreyfus appelle les « problèmes ouverts du quotidien ». Il va donc démontrer en quoi le comportement peut être régulier sans pour autant être gouverné par des règles formalisables, contrairement à ce que prétendent les défenseurs de l'IA forte comme Minsky[110].

Ainsi, par « problèmes ouverts du quotidien »[111], Dreyfus montre qu'il existe trois types de difficultés que l'on ne retrouve pas dans de simples jeux ou tests. Premièrement, ces problèmes nous demandent de déterminer quels faits sont potentiellement pertinents. Dans un deuxième temps, nous identifions lesquels sont effectivement pertinents et enfin, nous distinguons ceux qui sont essentiels de ceux qui sont accessoires. Et lorsqu'il s'agit d'effectuer cette dernière opération, Dreyfus remarque qu'elle ne peut être décidée à l'avance, ni indépendamment d'un moment précis et dans une situation déterminée.

Si l'on estime que des faits ne sont pas pertinents en soi, cela signifie que nous les catégorisons comme tels en fonction de buts et besoins humains et que ceux-ci varient en fonction de chaque situation. Cette évaluation de la situation dépend précisément de notre immersion dans une situation concrète. Sauf que, à la différence de nous, la machine n'est pas immergée dans une situation. Par conséquent, elle doit traiter constamment tous les faits comme potentiellement pertinentes, ce qui pousse les chercheurs en informatique à soit stocker une infinité de faits (ce qui

---

[110] Marvin Minsky (1927-2016) est un scientifique cognitiviste qui est l'un des plus grands partisans de la thèse de l'IA forte. Il estimait que la conscience humaine pourrait être un jour simulée par des machines suffisamment complexes.

[111] *Ibid.,* p. 257.

est limité par la mémoire de travail), soit à exclure arbitrairement certains faits potentiellement pertinents. Dans le deuxième scénario, cela implique donc que certains faits deviennent hors de portée pour le programme.

De nouveau, pour illustrer toute la complexité d'une situation et des problèmes ouverts qu'elle sous-tend, Dreyfus évoque l'exemple des paris hippiques[112] emprunté à Charles Taylor. Comme nous pouvons l'imaginer, ces paris font entrer en jeu un nombre illimité de faits potentiellement pertinentes.

Naturellement, l'esprit humain à lui seul ne peut appréhender tous ces cas de figure. C'est pourquoi il privilégiera certaines données au détriment d'autres qu'il juge pertinentes telles que l'âge du cheval, ses performances passées, etc. Mais il existe cependant d'autres éléments auxquels un parieur pourrait ne pas penser, mais qui sont tout de même cruciaux. Par exemple, le cheval est-il allergique à l'herbe qu'il foule ?

Comment procéderait une machine dans le cas d'une course hippique ? Son approche consisterait sans doute à trier selon un ensemble de données préalablement codées et étiquetées en se basant entre autres sur des documents provenant de la médecine vétérinaire. Néanmoins, elle ne pourrait forcément pas mobiliser tous les faits pertinents, car ils découlent de la situation qu'elle ne maîtrise pas étant donné qu'elle n'y est pas immergée. De plus, si elle devait emmagasiner une liste des faits du monde réel, elle ferait face à une infinité que sa mémoire ne peut raisonnablement stocker.

Ce développement permet ainsi à Dreyfus d'avancer l'une de ces thèses principales : le monde humain, tel que nous le connaissons, est pré-structuré[113] en fonction des buts et des préoccupations humaines. C'est-à-dire que les objets qui le peuplent ont une signification déterminée par ces mêmes préoccupations. En revanche, du côté des machines, cette structuration du monde ne peut être reproduite par celles-ci, car elles traitent les objets indépendamment de tout contexte. Dreyfus fait quand même remarquer que ces préoccupations peuvent être codées dans une machine, mais qu'il s'agit avant tout de l'intervention humaine.

---

[112] *Ibid.,* p. 258-259.

[113] *Ibid.,* p. 261.

Le champ de notre expérience n'est pas du tout neutre de notre point de vue, car il est structuré en fonction de nos intérêts. Les objets apparaissent donc en fonction de notre intérêt dominant du moment. Et comme ce champ d'action a été construit par nous, il est donc normal que nous y retrouvons ce qui est pertinent pour nous. Il y a donc d'un côté, ce qui nous ignorerons, et de l'autre, ce que nous intégrerons comme essentiel. Et nous placerons notre curseur en fonction de nos propres préoccupations et de notre savoir-faire passé.

> « When we are at home in the world, the meaningful objects embedded in their context of references among which we live are not a model of the world stored in our mind or brain; they are the world itself. This may seem plausible for the public world of general purposes, traffic regulations, and so forth »[114].

De manière plus large, l'idée que défend Dreyfus est de dire qu'il s'oppose à une conception mécaniste de l'humain qui recevrait des données issus de monde et dénues de sens et à partir desquelles il devrait y accoler une signification sur base d'un stock de connaissances décontextualisées. Autrement dit, le cerveau humain ne trie pas des faits hors contexte[115], car l'être humain est incarné dans un monde qui a du sens pour lui. Et ce sens ne peut découler sous forme de règles ou de représentations, car ce serait occulter notre capacité à agir adéquatement sans pour autant suivre de prescription explicites.

En somme, c'est la situation, le monde vécu, le corps engagé, la compréhension d'éléments contextuels qui nous permettent d'agir sans tout formaliser et d'en saisir le sens. C'est précisément cette dimension incarnée, située et pré-réflexive que l'IA ne parvient pas à reproduire.

Pour donner un autre exemple de l'importance de l'élément contextuel, nous citons l'expérience de pensée de Theodore Roszak reprise par Fjelland[116]. Nous nous imaginons en train d'observer un psychiatre chevronné à son travail. Lorsqu'il passe dans la salle d'attente, il aperçoit

---

[114] *Ibid.,* p. 265-266.

[115] *Ibid.,* p. 271.

[116] Fjelland, Ragnar, « Why General Artificial Intelligence Will Not Be Realized », *Humanities & Social Sciences Communications*, 7(1):10, 17 juin 2020, p.8.

ses patients qui manifestent un grand été de détresse : certains ont des pensées suicidaires, d'autres manifestent des hallucinations, etc. Le praticien prend le temps de l'écouter, mais le situation ne s'arrange pas malgré son professionnalisme dont il fait preuve. Si nous nous limitons à ces informations, nous ne pouvons pas vraiment comprendre pourquoi le praticien peine autant dans son travail.

Une fois ce décor posé, Roszak propose d'imaginer un contexte plus large qui entoure la situation. Nous apprenons que le psychiatre travaille en réalité dans un bâtiment qui lui-même se trouve à proximité d'un camp de concentration. Les patients sont donc des prisonniers des camps de concentration. Sans cet élément de contexte crucial, nous ne pourrions pas comprendre l'état de détresse des patients. La situation fait soudain sens, même si comme nous n'avons pas nous-même expérimentés cette situation, il nous est difficile de nous mettre à la place des patients. Néanmoins, nous pouvons comprendre la situation dans une certaine mesure, car nous partageons le même monde dans lequel les machines ne sont pas immergées. Nous ressentons et nous sommes capables de nous mettre « à la place » des autres car nous partageons un même monde.

Autre exemple cette fois-ci tiré de Langrebe et Smith et repris par Fjelland[117] au sujet de l'importance de la compréhension du contexte est celui de la durée d'une pause pendant une conversation. Comme chacun sait, sa durée varie en fonction du contexte et implique une charge émotionnelle en fonction des personnes avec qui nous nous trouvons. Ainsi, un entretien d'embauche est perçu comme un contexte certainement plus stressant que le fait de se retrouver entre amis autour d'une table en train de déguster un délicieux repas. Les enjeux sous-jacents ne sont évidemment pas les mêmes. Dès lors, comment une IA pourrait-elle emmagasiner toutes ces situations complexes, avec leurs multiples niveaux de lecture ?

Récemment, ChatGPT a proposé une fonctionnalité permettant de dialoguer avec elle. Si l'on se penche sur la gestion du silence telle que citée dans le paragraphe précédent, nous remarquons que l'IA ne manifeste pas d'hésitations naturelles comme nous pourrions le faire habituellement lorsque l'on dialogue avec autrui. De même, elle ne marque pas de blancs significatifs, ni de suspensions de parole pour laisser place à la réflexion ou à l'émotion. Cependant, nous pourrions sans doute la prompter de sorte à ce qu'elle insère des silences durant une

---

[117] Fjelland, Ragnar, « Computers Will Not Acquire General Intelligence, but May Still Rule the World », *Cosmos+Taxis*, vol. 12, no 5-6, 2024, p. 62.

conversation, mais cela resterait avant tout une mise en scène contextuelle détachée d'intentionnalité ou de stratégie de silence implicite. Preuve en est qu'elle ne fait pas partie de notre monde.

Ainsi, la thèse que défend Dreyfus et que nous reprenons à notre compte, c'est que l'intelligence n'est pas une conception purement abstraite, mais bien une entité incarnée. Elle ne se résume pas à des règles ou à des représentations, mais dans une compétence pratique et située à l'aide d'un être corporel engagé dans un environnement. La corporéité est ce qui ancre l'intelligence dans le réel et de ce fait, elle rend sensible au contexte, aux besoins, aux buts et aux interactions.

III.3. L'IA n'a pas d'*Umwelt* ou l'argument bio-psychique

Dans le même ordre d'idées, Daniel Andler reprendra à son compte plusieurs notions héritées de Dreyfus. Il insiste sur l'importance de la dimension biologique de l'intelligence — déjà présente chez Aristote, Ernst Haeckel et Hans Jonas — qui constitue souvent l'angle mort des débats entourant l'intelligence artificielle[118]. Cette dimension repose sur le fait que, tout au long de leur vie, les êtres vivants expérimentent des processus mentaux comme les pensées, les émotions ou encore la conscience. Or, ces processus émergent de phénomènes biologiques nombreux et complexes, liés notamment aux neurones, à la génétique ou à la chimie cérébrale.

Nous pouvons noter qu'Andler, tout comme Searle, soulignent tous deux l'importance de la dimension biologique. D'un côté, le premier note son rôle crucial lorsqu'il s'agit de comprendre le comportement du vivant, tandis que le second le cite en lien avec nos sensations qu'il désigne comme les *qualia* et qui relèvent de ce qui est « ontologiquement subjectif ». Comme Searle l'expliquait dans son article de 2014, l'IA peut nous dire ce qu'est une démangeaison, mais elle ne l'a jamais ressentie et donc elle ne sait pas ce que ça fait d'en avoir une. Sans être dotée d'une machinerie biologique aussi complexe que la nôtre, elle ne peut, par conséquent, pas ressentir la sensation.

---

[118] Daniel Andler, *Intelligence artificielle, intelligence humaine : la double énigme*, Paris, Gallimard, « NRF Essais », 2023, p. 255.

La dimension biologique est donc un critère de comparaison fondamental qui permet de comprendre pourquoi le vivant diffère de la machine. La machine de Turing, par exemple, est capable de résoudre une multitude de problèmes algorithmiques avec une grande efficacité, mais elle est limitée à ce seul type de problèmes[119]. Pour reprendre une distinction d'Aristote entre plantes, animaux et êtres humains, seuls les deux derniers types d'êtres vivants évoluent dans un environnement ouvert, dans lequel ils se déplacent et où ils rencontrent une infinité de configurations nouvelles auxquelles ils vont devoir s'adapter pour pouvoir survivre.

C'est donc à partir de l'argument bio-psychique qu'Andler présente la notion de « *Umwelt* » d'abord énoncée par le biologiste et ethnologue Jakob von Uexküll puis repris plus tard par le médecin philosophe Georges Canguilhem[120]. L'*Umwelt* désigne le monde subjectif propre à chaque être vivant. Chaque organisme perçoit et interagit avec son environnement à travers ses capacités sensorielles et ses besoins spécifiques. Ce monde n'est donc pas objectivement le même pour toutes les espèces. Ainsi, Andler cite le psychologue américain James Gibson qui prend l'exemple du singe et du sanglier : face à un même arbre, le singe y voit un support pour grimper, tandis que le sanglier le perçoit différemment, en l'assimilant comme un obstacle qui gène son déplacement et qu'il va s'employer à contourner.

L'*Umwelt* est donc indissociable de la constitution propre de l'animal et de son rapport singulier à son environnement. Ces éléments, combinés ensemble, permettent à Andler de dépasser une définition strictement technique de l'intelligence comme simple capacité à résoudre des problèmes.

Dès lors, peut-on dire qu'un robot (c'est-à-dire une IA dotée d'un corps mécanique) peut vivre dans un *Umwelt* ? Si l'on accepte la définition précédemment donnée, la réponse est non : l'environnement de cette IA ne constitue pas un *Umwelt* au sens fort, car elle ne partage pas la même histoire évolutive avec celui-ci[121]. En effet, cette machine n'aura pas le même type d'autonomie qu'un animal : elle est programmée pour agir selon des scénarios précis dans des

---

[119] *Ibid.,* p. 256.

[120] *Ibid.,* p. 257.

[121] *Ibid.,* p. 266.

contextes définis. L'animal, lui, possède des capacités cognitives qui lui permettent de s'adapter de manière souple à des situations imprévues.

Autrement dit, le monde dans lequel nous vivons est façonné par nous et il a du sens pour nous. Ensuite, par un deuxième mouvement, le monde nous façonne et influence ainsi notre perception de ce dernier. La machine, quant à elle, se situe en dehors de cette double influence, car ce n'est pas elle qui a construit le monde tel que nous le connaissons aujourd'hui.

Sans *Umwelt*, les machines ne peuvent pas faire face à une situation en agissant de la manière la plus ajustée possible. Et ces situations sont marquées par une grande variabilité et traversées par un flux de croyances et d'informations auquel l'être vivant doit répondre. Il existe donc deux conceptions de l'intelligence : l'une comme simple faculté de résoudre un problème, l'autre comme capacité à faire face à une situation qui fait entrer en jeu une multitude d'informations différentes.

Pour illustrer cette distinction, Andler décrit sept scénarios mettant en scène un personnage nommé André, qui vient de terminer sa journée de travail[122]. Chaque scénario introduit une variation, avec des implications différentes, révélant ainsi des configurations diverses du rapport à la situation. Voici une réécriture légèrement simplifiée de ces scénarios :

1) André, après avoir consulté divers moyens de transports, décide finalement d'attendre un autobus à l'arrêt le plus proche afin de rejoindre son ami pour aller dîner chez lui.

2) André rejoint son arrêt et décide d'engager la conversation avec un inconnu qui attendait aussi l'autobus afin d'éviter de devoir tenir la jambe à un de ses collègues qui l'épuise avec ses lamentations.

3) André attend pour prendre le prochain autobus qui vient. Selon cette configuration, il n'y aura aucun obstacle qui viendra entraver son déplacement. Entre-temps, il discute avec un inconnu qui attend lui aussi son autobus.

[122] *Ibid.,* p. 271-272.

4) André arrive finalement chez son ami qu'il avait perdu de vue, mais il avait appris entre-temps qu'il était lié à une affaire embarrassante. André le salue et s'excuse de son retard en disant qu'il est aussi gêné que son ami lors de son affaire embarrassante.

5) Dans une autre version de la situation précédente, André se déleste de son imperméable et il va s'installer en face de son ami en attendant que ce dernier lui offre un verre, puis il engage la conversation.

6) Dans une autre version de la première situation, André estime qu'arriver avec un retard de 45 minutes serait inacceptable, alors il décide d'informer son ami qu'il ne viendra pas chez lui.

7) Dans encore un autre version de la première situation et après ses observations sur les transports en commun, André décide de rallumer son ordinateur pour ouvrir le dossier d'une affaire qu'il avait précédemment mise de côté.

Andler classe ensuite ces situations en deux types :

- Type 1 : Situations comme le scénario A, où la transition de la situation à un problème à résoudre est relativement directe, sans nécessiter de médiation importante. L'action est fluide et se déroule un peu comme si l'on était sur un unique rail.

- Type 2 : Situations où cette transition n'est pas évidente, c'est-à-dire tous les autres scénarios. Plusieurs problématisations peuvent être envisagées, ou aucune, et il n'existe pas une solution unique ni claire.

Dans toutes ces situations de type 2, les préférences d'André (expérience, humeur, habitudes, etc.) jouent un rôle central[123]. Ces préférences préexistent à la situation, orientent son interprétation, et réduisent le champ des perspectives envisageables. La situation de type 1, uniquement représentée par la scénario numéro 1, impose une problématisation immédiate, tandis que les situations de type 2, c'est-à-dire tous les autres scénarios comme nous venons de le voir, appellent une diversité d'interprétations, sans qu'aucune ne s'impose d'emblée.

[123] *Ibid.,* p. 279.

Andler fait remarquer que nous avons tendance à associer les situations de type 2 comme relevant du domaine de l'éthique ou de l'intelligence émotionnelle, alors que celle de type 1 est en quelque sorte prémunie de ces perpectives. En effet, celles du type 2 touchent à tous les domaines, mais nous en tenons seulement compte que lorsque nous rencontrons une difficulté inhabituelle[124]. En somme, une même situation peut donner lieu à divers problèmes, dont les solutions ne sont pas nécessairement comparables, car elles mobilisent les préférences du sujet, avant même qu'il ne vive la situation.

Face à cette complexité, Andler s'interroge : si les préférences d'un individu comme André étaient figées, serait-il possible de résoudre ces problèmes par une méthode purement analytique ? La réponse est non. Le passage de la situation au problème mobilise des opérations cognitives d'un autre ordre, où convergent simultanément divers champs tels que les sciences physiques, les sciences humaines, etc[125].

Dans le cas de l'IA symbolique, Andler constate que ses principes sont inopérants pour faire face au monde réel, même si elle contenait une masse d'informations sur des catégories de situations préalablement modélisées dans ce qu'il appelle des micro-mondes. Sans sens commun, ni familiarité avec le monde, cette IA est incapable de s'adapter aux situations humaines caractérisées par leur complexité. De fait, une simple description ne peut pas rendre compte de la richesse d'un *Umwelt* commun.

D'un autre côté, l'IA connexionniste est-elle mieux outillée pour faire face à une situation ? Après tout, contrairement à l'IA symbolique, elle possède un modèle d'apprentissage fondé sur des réseaux neuronaux artificiels. Lorsqu'un cas inédit lui est présenté, elle tente d'apporter une réponse en s'appuyant sur un exemple stocké dans sa base de données, le plus proche possible de la situation présente. Mais est-ce suffisant ?

Il serait illusoire de croire qu'une IA connexionniste, conçue, entraînée et alimentée par des humains, puisse contenir un échantillon suffisamment représentatif de toutes les situations

---

[124] *Ibid.,* p. 280.

[125] *Ibid.*, p. 282

humaines. Pour disposer d'une telle base de données, les ressources physiques nécessaires dépasseraient l'entendement. Toutefois, nous pourrions critiquer cette idée en disant que l'histoire a démontré que la question essentielle de la capacité de la base de données, déjà soulevée par Turing en 1950, est sujette à des rebondissements. Si nous disons qu'elle est limitée à ce jour, nous ne pouvons pas encore anticiper ce qu'il adviendra dans des dizaines d'années.

Toujours est-il que contrairement à la machine, l'être vivant est plongé dans un monde dans lequel il est impliqué et qui l'affecte. L'IA, quant à elle, n'habite pas le monde, la phénoménologie des affects lui échappe. Elle s'informe et s'entraîne sur la vie, mais sans jamais la vivre réellement. Par conséquent, elle ne peut pas saisir ni « comprendre » tous les enjeux auxquels les humains font face dans leur vie.

De son côté, Andler[126] estime que même si l'on parvenait à stocker toutes ces descriptions de situations, cela aboutirait à ce que Nagel appelle *the view from nowhere*. En d'autres termes, l'IA disposerait d'une perspective impersonnelle qui lui ferait atteindre une forme d'objectivité universelle qui serait néanmoins coupée de toute phénoménalité.

III.4. L'IA ne joue pas au jeu du langage

Dans cette partie, nous allons tenter de montrer que l'IA générative ne participe pas au jeu du langage à partir de la notion telle qu'abordée par Wittgenstein dans les *Recherches philosophiques*[127]. Pour reprendre son idée dans les grandes lignes, l'un des buts de son oeuvre est d'opérer une rupture avec la conception référentielle du langage. Contrairement à l'idée selon laquelle les mots ne seraient que des étiquettes apposées à des objets ou à des idées et que la compréhension linguistique reposerait sur une forme de correspondance entre les deux, Wittgenstein nous invite à plutôt privilégier une approche pragmatique. Comme il l'écrit au paragraphe 43, le langage est avant tout le fruit d'une pratique sociale régie par des règles. C'est ce qu'il nomme précisément comme des jeux du langage : comprendre un mot, c'est être capable de l'utiliser adéquatement en fonction du contexte.

---

[126] *Ibid,* p. 283

[127] Ludwig Wittgenstein, *Recherches philosophiques*, traduit de l'allemand par Elisabeth Rigal, Paris, Gallimard, collection « Tel », 2004, p. 27-51.

Les jeux du langages sont donc liés, selon lui, à ce qu'il appelle des « formes de vie », ce qu'il faut interpréter comme des contextes culturels et pratiques qui nous fournissent une cohérence aux règles du langage. Il est donc indispensable d'intégrer une communauté linguistique réelle afin de participer à ce jeu du langage.

En outre, comme il est dit au paragraphe 23, comprendre un mot, ce n'est pas simplement l'associer à quelque chose, mais c'est être capable de l'utiliser à bon escient dans une multitude de contextes. Il se dégage alors une forme de « bon sens » dans l'usage des mots que l'on utilise dans nos pratiques. Lorsque l'on suit une règle, ce n'est pas simplement produire des régularités, c'est aussi être en mesure de reconnaître les écarts, d'expliquer ses usages et de se corriger en cas d'erreur.

À la lumière de cette conception, que pouvons-nous dire avec l'IA générative ? Déjà, elle ne suit pas une règle au sens wittgensteinien du terme : elle reproduit des régularités probabilistes, elle imite l'usage. Certes, elle peut répondre à des questions, générer des poèmes ou tenir une conversation de manière fluide et cohérente, mais elle n'est pas ancrée dans des formes de vie dans laquelle les individus agissent, sentent et font société. Elle manipule des symboles sur la base d'entraînements en lien avec des modèles statistiques appris à partir de vastes corpus textuels. En aucun cas elle n'éprouve ni ne partage aucune des pratiques sociales dans lesquelles ces mots prennent sens.

Cette simulation du langage est rendue possible car l'IA générative est entraînée à prédire la probabilité d'apparition du mot suivant dans une séquence. C'est à partir de milliards de mots extraits de sources comme internet, de livres et de journaux qu'elle identifie des régularités syntaxiques, des structures phrastiques ou des associations lexicales. Si ses textes nous apparaissent cohérents, c'est par le procédé d'une imitation probabiliste, et non par une production de sens ancrée dans un monde de contextes culturels vécus.

Si l'on devait retenir un point essentiel de la thèse wittgensteinienne, c'est celle que le langage n'est pas simplement une structure formelle, mais une pratique ancrée dans la vie humaine. L'IA ne prend pas part à des formes d'engagement, de croyance ou de normativité qui accompagnent les actes du langage. Au mieux, elle produit une simulation du langage, mais sans

être immergée dans un monde. Elle ne participe pas au jeu du langage : elle hérite de traditions humaines sans s'y engager réellement.

Conclusion générale

Arrivés au terme de ce développement, nous pouvons désormais apporter une réponse nuancée à la question de départ. Si l'on suit le raisonnement de Turing, selon lequel une IA peut être dite « pensante » dès lors qu'elle imite le comportement humain de manière indiscernable, alors on pourrait soutenir que des modèles comme ChatGPT, dans certaines situations, remplissent ce critère. D'autant plus qu'il est aujourd'hui possible de simuler des conditions proches du test de Turing, tout en tenant compte des critiques méthodologiques formulées par Dennett, notamment à propos des biais relevés lors des Prix Loebner. Ainsi, l'IA générative contemporaine se montre capable de produire des interactions convaincantes, à condition qu'elle soit correctement « promptée ».

En effet, ces systèmes sont désormais capables de répondre à une grande variété de tâches : interprétation de résultats médicaux, analyses psychologiques basées sur des données fournies, rédaction de poèmes ou de blagues, etc. Les limites autrefois pointées, telles que l'ambiguïté linguistique ou le manque de connaissances sur le monde, tendent à s'estomper. Par conséquent, l'engouement du public pour ces technologies ne faiblira sans doute pas. Toutefois, cet enthousiasme s'accompagne souvent d'une projection erronée de qualités humaines sur ces machines ainsi que d'une surestimation de leurs capacités réelles.

Au-delà de leurs performances, certains arguments en faveur de l'IA faible demeurent pertinents, malgré l'émergence des architectures connexionnistes. Celles-ci peuvent certes apprendre à partir de vastes corpus de données, mais elles restent profondément dépendantes des interventions humaines dans leur conception et leur calibration. Ce sont les concepteurs qui orientent le modèle vers des réponses jugées appropriées, en y injectant sens et finalités. Searle, avec sa célèbre analogie de la chambre chinoise, rappelle que l'IA manipule des symboles sans jamais en comprendre la signification : aussi élaborés soient-ils, ces algorithmes n'accèdent pas à une véritable compréhension. Le sens reste tributaire du regard humain, donc des ingénieurs.

Autre point que nous avons peu abordé dans ce travail et qui va encore dans le sens de l'argument de Searle est celui de l'alignement des valeurs. Les IA génératives actuelles sont conçues de sorte à fournir des réponses qui rentrent dans le cadre de la bienveillance et de ce qui est moralement acceptable. Les concepteurs en sont conscients et l'actualité l'a même montré, le public accorde beaucoup de crédit à ces technologies. L'enjeu pour eux est donc d'éviter tous dérapages dans les réponses fournies par l'IA. De cette façon, les « valeurs » qu'ils ont intégrées dans leur création n'ont de sens que pour nous et les concepteurs. Ainsi, la machine ne « comprend » pas pourquoi elle doit répondre d'une telle façon plutôt qu'une autre. Cette programmation empêche également l'IA d'agir de manière autonome, mais plutôt de manière conditionnée.

Si l'IA parvient à dégager un certain « sens », c'est uniquement grâce à sa capacité à repérer des régularités dans des corpus textuels ou visuels, sans comprendre réellement de quoi elle parle. Son fonctionnement repose sur des statistiques probabilistes et corrélatives : elle détermine le mot le plus probable qui doit suivre dans une séquence. Ce rapport au monde, fondé sur la corrélation et non sur la causalité, limite intrinsèquement sa capacité à appréhender les raisons profondes des phénomènes.

Certes, il est vrai que l'IA connexionniste s'inspire du réseau neuronal déjà présent dans notre propre cerveau et que cette configuration va dans le sens de l'invariance organisationnelle tel que défendue par Chalmers. En revanche, peut-on dire que cela soit suffisant pour attribuer une conscience ou un vécu subjectif aux machines ? Aucun de nos philosophes abordés ne le pensent à l'heure actuelle, mais Dennett et surtout Chalmers estiment qu'il n'est pas impossible que cela se produise et que cela pourrait soulever de nouveaux enjeux éthiques.

Nous soulignons que l'argument neuro-biologique de Searle reste ici encore pertinent : la machinerie biologique dont nous sommes dotés reste encore impossible à reproduire de manière artificielle. Tout comme l'a fait remarquer Andler, personne n'est parvenu à concevoir un cerveau artificiel qui nous permettrait par la même occasion d'en savoir davantage sur les mécanismes de notre conscience. En conséquence, il demeure prématuré d'attribuer une pensée authentique aux machines, en l'absence manifeste de conscience.

Notre analyse a aussi montré qu'il existe des dimensions constitutives de l'expérience humaine qui restent hors de portée des technologies actuelles. En robotique, malgré plusieurs

avancées, l'IA embarquée n'interprète pas son corps comme sien, mais comme un outil externe. Or, Dreyfus a montré que l'intelligence humaine est enracinée dans un corps situé dans un monde. Les machines, même perfectionnées, restent désincarnées et incapables de faire face à des situations inédites nécessitant une intégration de multiples éléments contextuels.

Par ailleurs, étant donné que l'IA n'est pas non plus immergée dans un monde, elle ne peut en saisir les enjeux qui en découlent. Contrairement à nous, elles n'évoluent pas dans un environnement pré-structuré par nos besoins et des intentions. Elles ne partagent pas un monde commun avec autrui, et elles ne sont donc pas en mesure d'éprouver ou de comprendre ce que cela fait d'être à la place de l'autre. Elles ne peuvent pas non plus intégrer la complexité des situations humaines où s'entrecroisent préférences individuelles, émotions, et normes sociales.

En guise de conclusion, même si l'IA actuelle représentée par le connexionnisme parvient à imiter de manière convaincante certains aspects du langage humain ainsi que nos capacités de raisonnement, elle reste cependant fondamentalement dénuée de compréhension, de conscience et d'ancrage dans le monde. Malgré ses performances bluffantes, elle ne « pense » pas au sens humain du terme, car elle n'éprouve ni intentions, ni subjectivité et encore moins de vécu corporel. C'est pourquoi nous estimons que la pensée reste une propriété organiquement humaine.

Mais au delà même de la question de savoir si les machines « pensent », la véritable interrogation qui nous semble importante est celle des intérêts que nous aurions à concevoir une telle entité. Nous le savons, depuis toujours, la technologie a contribué à faciliter notre existence en construisant un monde qui est de plus en plus adapté et accessible à nos besoins. Toutefois, ces avancées se sont souvent accompagnées de critiques. Ainsi, Fjelland[128] nous rappelle que déjà chez Platon, l'apparition de l'écriture — pourtant un progrès technologique majeur — fut critiquée sous forme de mythe, car elle était perçue comme une « technologie de l'oubli » : la transmission orale et la mémorisation, si centrales durant l'Antiquité, risquaient alors de décliner.

Dans notre cas, l'IA et ses avancées, comme nous l'avons montré dans ce travail, continueront à alimenter à la fois engouement et critiques. D'un côté, nous pouvons dire qu'il y a un intérêt pragmatique et scientifique à concevoir des machines qui se rapprochent de l'être humain

[128] Fjelland, Ragnar, « Computers Will Not Acquire General Intelligence, but May Still Rule the World », *Cosmos+Taxis*, vol. 12, no 5-6, 2024, p. 66.

afin d'en savoir davantage sur nous-mêmes, par exemple en mettant à l'épreuve nos théories sur la conscience et l'intelligence, comme ce fut le cas avec le robot Cog. C'était d'ailleurs un des objectifs de Herbert Simon et Allen Newell lorsqu'ils présentèrent leur projet « *Logic Theorist »* à Dartmouth en 1956.

D'un autre côté, nous ne mesurons pas encore tout à fait les coûts que cette avancée pourrait avoir sur notre monde. En effet, de nos jours, beaucoup de personnes accordent beaucoup de crédit aux productions de l'IA générative, qu'il s'agisse de décisions politiques, médicales ou psychologiques. Comme nous l'avons vu, depuis les premières IA conversationnelles et les systèmes-experts, il faut craindre que cette confiance fasse baisser notre propre vigilance ainsi que notre esprit critique. Comme nous l'avons vu, déjà depuis l'émergence des premières IA conversationnelles et des systèmes-experts, il existe une tendance constante à les « anthropomorphiser », malgré leurs limites évidentes. Les nombreux exemples récents de *deepfakes* montrent d'ailleurs que leur usage peut aisément servir à des fins de désinformation.

Tout récemment[129], l'actuel ministre suédois a reconnu recourir à un LLM pour demander un « deuxième avis » sur la manière de gérer son pays. Sa déclaration a évidemment suscité de nombreuses critiques. Parmi celles-ci, nous retrouvons des citoyens mécontents qui estiment qu'ils n'ont pas voté pour une IA, mais pour une personne en chair et en os. Par cette occasion, ils soulignent ainsi clairement la différence entre le traitement humain et celui d'une IA. De plus, des experts en technologie ont alerté sur les risques liés à des bases de données biaisées ainsi qu'à l'apparition d' « hallucinations », phénomène dont nous avons déjà parlé dans ce travail. Il est raisonnable de penser que ces technologies s'invitent également dans d'autres sphères de pouvoir, mais aussi dans le domaine informationnel, et bien au-delà. Quoi qu'il en soit, leur pratique est de plus en plus courante et semble progressivement s'ériger en nouvelle norme.

Notre objectif n'est donc pas de tirer à boulets rouges sur l'IA, qui présente certes des limites dans sa forme actuelle, mais aussi des potentialités intéressantes susceptibles de faciliter et d'améliorer nos vies. Par exemple, elle peut constituer un outil dans le domaine médical pour aider le personnel soignant, rendre des productions écrites et visuelles plus accessibles en facilitant la

---

[129] Roselyne Min, « Le premier ministre suédois utilise ChatGPT. Comment les gouvernements utilisent-ils les *chatbots* ? », *Euronews Next*, 7 août 2025. URL : https://fr.euronews.com/next/2025/08/07/le-premier-ministre-suedois-utilise-chatgpt-comment-les-gouvernements-utilisent-ils-les-ch. Consulté le 7 août 2025.

traduction ou encore servir d'appui en matière de cybersécurité, etc. Dans notre cas, il s'agit surtout de réfléchir à ses usages et d'apprendre à s'en servir correctement, en connaissance de cause. Nous suggérons ainsi de cultiver une véritable transparence : nous pourrions imaginer, par exemple, que les concepteurs et ingénieurs s'impliquent dans la réalisation de vulgarisation scientifique, afin d'expliquer le fonctionnement ainsi que les limites actuelles de l'IA. De même, les programmes pourraient intégrer un tutoriel explicitant leur logique interne, offrant ainsi aux utilisateurs une meilleure compréhension de l'outil qu'ils utilisent.

En définitive, la véritable question n'est pas seulement de s'interroger sur la pensée des machines, mais de déterminer en toute lucidité de la place que nous voulons leur accorder dans notre avenir commun.

Bibliographie

Livres

Daniel Andler, *Intelligence artificielle, intelligence humaine : la double énigme*, Paris, Gallimard, « NRF Essais », 2023.

David J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory*, New York, Oxford University Press, 1996.

Daniel C. Dennett, *La conscience expliquée*, Paris, Éditions Odile Jacob, 1993.

Hubert Dreyfus, *What Computers Still Can't Do : A Critique of Artificial Reason*, Cambridge (Mass.), MIT Press, 1992.

John R. Searle, *Le mystère de la conscience*, Paris, Éditions Odile Jacob, 1999.

Ludwig Wittgenstein, *Recherches philosophiques*, traduit de l'allemand par Elisabeth Rigal, Paris, Gallimard, collection « Tel », 2004.

Articles

Bubeck, Sébastien, et al., « Sparks of Artificial General Intelligence: Early Experiments with GPT-4 », *arXiv.org*, mars 2023. DOI : 10.48550/arXiv.2303.12712.

Chalmers, David J., « Absent Qualia, Fading Qualia, Dancing Qualia », in Thomas Metzinger (dir.), *Conscious Experience*, Paderborn, Schöningh, 1995, p. 309-328.

Chalmers, David J., « Could a Large Language Model Be Conscious? », *Boston Review*, 9 août 2023.

Dennett, Daniel C., « Can Machines Think? », in *Brainchildren: Essays on Designing Minds*, Cambridge (Mass.), MIT Press, 1998, p. 3-29.

Dennett, Daniel C., « Can Machines Think? Deep Blue and Beyond », *Studium Generale Maastricht*, 1997.

Dennett, Daniel C., « What Can We Do? », in J. Brockman (dir.), *Possible Minds: Twenty-Five Ways of Looking at AI*, New York, Penguin Press, 2019.

Edelman, Gerald M., et al., « Synthetic Neural Modeling Applied to a Real-World Artifact », *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no 15, 1992, p. 7267-7271.

Fjelland, Ragnar, « Why General Artificial Intelligence Will Not Be Realized », *Humanities & Social Sciences Communications*, 7(1):10, 17 juin 2020, p. 1-9.

Fjelland, Ragnar, « Computers Will Not Acquire General Intelligence, but May Still Rule the World », *Cosmos+Taxis*, vol. 12, no 5-6, 2024, p. 58-68.

Jones, et Bergen, « Large Language Models Pass the Turing Test », *arXiv.org*, mars 2025. URL : https://arxiv.org/abs/2503.23674.

Landgrebe, Jobst, et Barry Smith, « Why Machines Do Not Understand: A Response to Søgaard », *arXiv.org*, juillet 2023. DOI : 10.48550/arXiv.2307.04766. URL : https://arxiv.org/abs/2307.04766.

Leclercq, Bruno, « Ni fantôme, ni zombie : L'émergence de la conscience subjective dans le flux des expériences », *Bulletin d'Analyse Phénoménologique*, vol. 10, no 3, 2014. URL : https://popups.uliege.be/1782-2041/index.php?id=697.

LeCun, Yann, Bengio, Yoshua, et Hinton, Geoffrey, « Deep Learning », *Nature*, vol. 521, 2015, p. 436-444. URL : https://doi.org/10.1038/nature14539.

Liu, Alisa, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith et Yejin Choi, « We're Afraid Language Models Aren't Modeling Ambiguity », *arXiv.org*, avril 2023. DOI : 10.48550/arXiv.2304.14399. URL : https://arxiv.org/abs/2304.14399.

Lucas, John R., « Minds, Machines and Gödel », *Philosophy*, vol. 36, 1961, p. 112-127.

Ortega-Martín, et al., « Linguistic Ambiguity Analysis in ChatGPT », *arXiv.org*, février 2023.

Searle, John R., « What Your Computer Can’t Know », *The New York Review of Books*, vol. 61, no 15, 9 octobre 2014.

Turing, Alan, « Computing Machinery and Intelligence », *Mind*, vol. 49, 1950, p. 433-460.