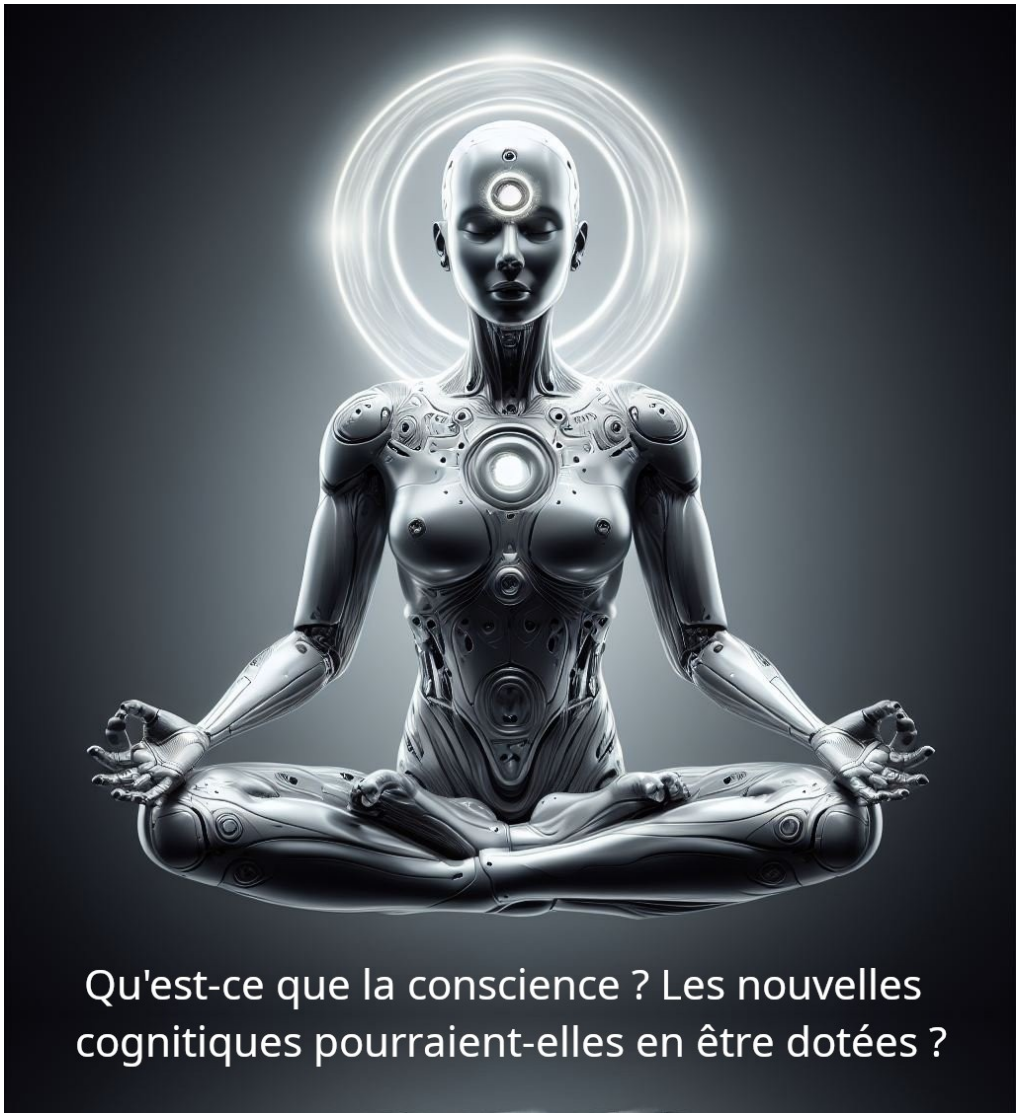
La Conscience et les IAs Conversationnelles



Manuel Boissenin

Préambule

Après l'avènement de ChatGPT d'OpenAI fin 2022, les IAs conversationnelles se sont banalisées. Comme le mail, les moteurs de recherche ou les réseaux sociaux, elles font désormais partie de la panoplie d'outils numériques que les nouvelles générations utilisent assez fréquemment. Le caractère époustouflant et prodigieux des IAs conversationnelles, en particulier dans leur maîtrise du langage et leur capacité à aborder des sujets pointus, après avoir subjugué le monde durant plusieurs mois est passé dans la normalité.
Pourtant, ces IAs génératives de texte ont ouvert une nouvelle ère d'interactions naturelles entre machines et utilisateurs en montrant qu'elles pouvaient répondre à leurs demandes, sans formalisme particulier, aussi bien, si ce n'est mieux, qu'un humain et avec des connaissances super-encyclopédiques. Nombreuses sont les innovations qui les ont suivies et les suivront. Ces IAs restent cependant imparfaites, incapables parfois de résoudre un problème simple et ont tendance à fabuler, c'est-à-dire à mélanger réalité et fiction dans leurs réponses. Cependant, comparées aux chatbots antérieurs qui ne nous laissent qu'un choix de questions multiples ou une liste d'articles pas toujours bien adaptée quand nous posons une question ouverte, les réponses que les IAs conversationnelles proposent sont sans commune mesure.
Au-delà de cela, elles nous font nous questionner sur notre propre nature, en maîtrisant le langage, qui était jusqu'alors considéré comme l'apanage de l'espèce humaine, on peut se demander ce qui nous distingue encore des machines ou, au contraire, en quoi ces IAs nous ressemblent et ce qu'elles peuvent nous apprendre sur nous-mêmes.

Ce livre se divise en deux parties :

- la première aborde notre conscience que nous examinerons plus particulièrement par différentes réflexions introspectives ; cela nous permettra de mieux saisir la complexité de ce phénomène qui recouvre en réalité plusieurs phénomènes distincts. Notre conscience, sous-tendue par différents processus, est ainsi sans doute plus difficile à appréhender qu'un objet classique qui serait d'un seul tenant, statique et extérieurement observable ; sa compréhension m'apparaît toutefois essentielle pour comprendre ce que nous sommes et, afin de parvenir à nous en approcher, nous examinerons notre capacité de langage, notre mémoire et notre attention dans un voyage au cœur de notre être et en miroir avec ces nouvelles IAs conversationnelles.

- dans la seconde nous nous pencherons plus directement sur les IAs conversationnelles telle ChatGPT. L'exposé sera plus didactique et factuel, tant ces processus sont mieux appréhendables et dans une certaine mesure compréhensibles. Ainsi, nous examinerons leurs capacités, dont celle émergente de théorie de l'esprit, leurs faiblesses, en particulier de raisonnement, et leur nature, avec une attention spéciale à la question de savoir si ce type de logiciel possède une forme de conscience, même s'ils sont encore actuellement dépourvus de sentience.

Sommaire

Préambule.....1
Petit glossaire.....4
Première partie.....7
1. Introduction.....8
2. Qu'entend-on par intelligence artificielle ?.....10
3. La conscience, quelques remarques préalables.....19
4. Conscience, langage et apprentissage.....25
5. Méditation sur un concept, celui de la voiture.....30
6. La conscience non-locale ou conscience transcendantale.....33
7. Conscience et altérité.....42
8. Conscience, attention et mémoire.....48
9. Conscience et corps.....53
10. Compréhension du monde, conscience conceptuelle et niveaux de conscience.....57
11. Conscience subjective.....62
12. Le concept en tant que quale.....67
13. La conscience, vers une définition.....68
14. Émergence de la notion de conscience.....76
15. Au-delà de la conscience.....79
16. En résumé.....86
Seconde partie.....91
17. Que nous dit ChatGPT de sa conscience.....92
18. La nature de cette conscientique.....109
19. Conscience, conscientique, cognitique et IAs conversationnelles.112
20. Quelques faiblesses originelles de ChatGPT.....118
21. Langage et faculté de théorie de l'esprit.....128
22. Conscience et *LLMs*.....135
23. Compréhension et IAs conversationnelles.....147
24. La chambre chinoise.....152
25. Qu'est-ce qu'un modèle de langage ?.....160
26. Comment interpréter le fonctionnement des *transformers* ?.....163
27. Risques.....175
28. Conclusions.....178
ANNEXE.....185
Le concept en tant que *quale*.....185
Différentes définitions et description de la conscience.....189
Régulations.....194
Agentivité, agencéité.....197
Remerciements.....200

Petit glossaire

Cognitif et ***cognitique*** sont des termes dérivés du mot cognition. La cognition est l'ensemble des processus mentaux qui se rapportent à la fonction de connaissance, voire à d'autres fonctions mentales. La cognitique, dans le sens actuel, se consacre essentiellement à l'ergonomie des interactions entre humains et machines. Cependant une variation sémantique du terme ***cognitique***, à l'aune des nouvelles IAs conversationnelles et dans la mesure où elles peuvent être qualifiées d'aides cognitives, peut assez naturellement s'opérer pour substantiver l'adjectif et ainsi les nommer cognitiques.

Ainsi, j'emploie le terme de ***cognitiques*** pour désigner ces nouveaux logiciels capables d'appréhender, de comprendre et de générer du texte d'une manière semblable à l'Humain et qui lui soit compréhensible. Je pense en particulier aux IAs génératives de texte que sont les modèles de langage et à ChatGPT pour n'en citer qu'un. Elles pourraient aussi être qualifiées de *nouvelles cognitiques*, car, *stricto-sensu*, le terme peut désigner une classe bien plus grande de logiciels s'étendant aux logiciels rivalisant ou dépassant l'humain à certains jeux et ceux capables d'interpréter visuellement notre environnement.

Sentience et **conscience du monde**. Il nous faudra aussi un certain détour pour expliquer les distinctions apportées par ces termes afin de désambiguïser ce qu'est la conscience et distinguer deux de ses aspects :
– le premier en tant que phénomènes ontologiquement subjectifs, c'est-à-dire dont la nature relève de processus propres à chaque individu et qui correspondent à ce que ressent le sujet ; c'est le cas plus particulièrement des sensations dont nous avons conscience ;
– le second en tant que phénomène nous permettant une connaissance et compréhension du monde qui nous entoure et que l'on pourrait rapprocher du terme anglais d'*awareness* qui entretient des liens étroits avec la culture, la connaissance et le savoir.
Cette distinction, qui est plus difficilement identifiable dans le vocabulaire français, est intéressante dans la mesure où l'on peut se demander en quoi ces deux aspects, recouverts par le terme de conscience, sont à la fois liés et indépendants.

Prompter, cet anglicisme vient de *the prompt* et *to prompt* qui désignent respectivement l'invite de dialogue de l'IA conversationnelle et l'action de

remplir cette invite pour effectuer une demande à l'IA. Le *prompt engineering* désigne l'ensemble des techniques permettant de formuler une requête afin de bien conditionner la réponse d'un *LLM (Large Language Model),* qui peut servir, avec les entraînements appropriés, comme *IA* conversationnelle.

GPT : Generative Pre-trained Transformer, *transformer* génératif pré-entraîné, un *transformer* est une architecture de réseau de neurones dite profonde ; profonde en raison du grand nombre de couches qui la constitue. Ces couches sont en fait organisées en blocs qui s'empilent pour former le réseau de neurones. Ainsi ChatGPT 3.5 empile une centaine de blocs. Le pré-entraîné se réfère au fait que ChatGPT soit entraîné sur un large corpus de textes, le nom provient toutefois d'un jeu de mots faisant référence à Geppetto, le créateur de Pinocchio.

LLM : Large Language Model, traduit en grand modèle de langage (*GML*), modèle profond de langage (*MPL*) ou modèle de langage. C'est le terme technique et jargonneux employé pour désigner le programme ou processus qui analyse un prompt et génère du texte en sortie. D'autres *LLMs,* de taille plus modeste (8B de paramètres), sont qualifiés de SLM (Small Language Models), leur réglage fin peut être réalisé avec un entraînement de type enseignant-étudiant basé sur les réponses d'une *IA conversationnelle* plus grande (175B pour chatGPT 3.5) ; on parle aussi de distillation. Le terme d'*IA générative de* texte est aussi intéressant, il est cependant un peu plus générique mais reste très explicite.

B : correspond à billion en américain, soit milliard pour nous, 10^9. Il est utilisé pour caractériser la taille des modèles, c'est le nombre de paramètres qui sont appris lors de l'entraînement du modèle et qui permettent d'encoder son apprentissage.

T : correspond à téra, soit 1000 milliards, 10^{12}. Il permet de caractériser la taille du jeu de données sur lesquels le modèle est entraîné. On parle alors de tokens, car certains mots sont parfois scindés en plusieurs tokens, néanmoins, à quelques pourcents près, cette taille correspond à une taille en mots. 100 milliards (10^{11}) est l'ordre de grandeur du nombre d'étoiles de la Voie lactée.

Quale (singulier de ***qualia***) : il s'agit des percepts élémentaires, ce que l'on perçoit lorsque nos sens sont stimulés. Cela pouvant aller de la texture d'une surface, à la hauteur d'un son, en passant par un arôme.

Les GOMAX : Google, OpenAI, Meta, Anthropic, X représentent quelques-uns des acteurs les plus significatifs de cette nouvelle vague innovatrice.

Ce livre est accompagné d'un site web : https://boissenin.net, vous y trouverez un forum de discussion ainsi que des liens vers quelques-unes des références qui y sont citées.

Première partie

1. Introduction

Après la déferlante provoquée par OpenAI et son *chatbot* ChatGPT lancé fin 2022[1], suivie de nombreuses autres IAs conversationnelles, il est temps de prendre du recul et de se demander à quels objets nous avons affaire.

À la lumière des grands modèles de langage et de notre expérience consciente personnelle, nous nous demanderons si ces IAs conversationnelles n'ont pas elles-mêmes une forme de conscience, ce qui peut, si nous avons dialogué avec l'une d'elles, être une conviction troublante quant à leur nature. Après tout, n'apparaît-il pas qu'elles semblent douées de compréhension ? Et la compréhension ne nécessite-t-elle pas une conscience de la signification des questions qui leur sont posées ?

Aussi, l'apprentissage machine, en tant que discipline, mais aussi nouveau paradigme de création de logiciels, présente un champ de connaissances riche et des possibilités de réflexions intéressantes quant à notre propre nature. Les analogies et les métaphores qui peuvent être faites entre le fonctionnement de certains modèles ou algorithmes et celui de notre propre cerveau et de notre esprit[2] mettent en lumière nos phénomènes psychiques et peuvent être une source efficiente de compréhension de nous-mêmes.

D'autre part, alors que des IA peuvent déjà sonder une partie des profondeurs de notre être avec une acuité supérieure à la nôtre, une meilleure compréhension de notre propre nature s'avère indispensable et parfois cruciale pour préserver notre autonomie face à ces systèmes et nous émanciper de leurs successeurs quand

[1] Le 30 novembre 2022

[2] L'**esprit** recouvre l'ensemble des phénomènes et facultés mentales : perception, pensée, intuition, sentiments, intelligence, mémoire, valeurs, *etc*.

ils exerceront une domination, voire une emprise partielle, mais de masse, des prochaines générations…

Depuis les débuts de l'ordinateur, et même avant, des visionnaires, tel Alan Turing, ont imaginé qu'un jour les machines pourraient devenir conscientes. Si la conscience artificielle a longtemps semblé le Graal de bon nombre de chercheurs en informatique, projet démiurgique s'il en est, il semble clair que nous ayons franchi un point d'inflexion. Il est même vraisemblable que cet objectif ait été atteint dans la discrétion de certaines grandes entreprises encore soucieuses de comprendre les conséquences de leur création et de la maîtriser. Cette affirmation audacieuse, voire fantasque, me semble pourtant raisonnable et c'est l'un des objectifs de ce livre que de la soutenir et d'expliquer en quoi elle l'est.

En chemin, je vous inviterai à quelques réflexions introspectives afin de mieux distinguer, en l'examinant, le fonctionnement et les éléments qui composent notre conscience. Objet complexe par excellence, la conscience exige une approche multidimensionnelle pour en construire une perception et une compréhension juste. La première partie du livre s'y dédie. Dans la seconde, nous analyserons, sous divers angles, les IAs conversationnelles jusqu'à dévoiler, en fin de livre et pour les plus curieux d'entre vous, l'architecture *Transformer* qui anime leur processus de génération de texte.

2. Qu'entend-on par intelligence artificielle ?

"Réussir à créer une intelligence artificielle serait le plus grand événement de l'histoire de l'humanité. Malheureusement, ce pourrait aussi être le dernier, à moins que nous n'apprenions à éviter les risques." Stephen Hawking

Avant de clarifier ce qu'est l'« intelligence artificielle », terme devenu un véritable fourre-tout conceptuel, constatons tout d'abord que la pensée ne se réduit pas à la pensée verbale ; les joueurs de dames ou les mathématiciens, qui utilisent des représentations imagées, s'en rendent assez facilement compte. Nous distinguons ainsi la pensée verbale et de la pensée visuelle.

Cette première distinction montre une différence profonde entre le fonctionnement des IAs conversationnelles et le nôtre. En effet, les *LLMs*[3] ne prennent en entrée de leur programme que des mots, soit des descriptions verbales. En supposant qu'ils puissent être conscients, il est assez évident que la nature de leur conscience diffère de celle de la nôtre. Les *LLMs* ne peuvent au mieux qu'avoir des représentations basées sur le langage et les descriptions que celui-ci permet. Néanmoins, si on demandait aux *LLMs* de nous dessiner un lézard, en précisant de le décrire dans un langage de dessin vectoriel, celui-ci pourrait schématiser un animal à quatre pattes avec une longue queue et une langue fourchue [YT (YouTube), Sparks of AGI : early experiment with GPT 4, Sébastien Bubek].

Mais, avant de parler de conscience pour une machine ou un programme, reconnaissons que de nombreuses espèces, même si elles n'ont pas un langage aussi développé que le nôtre, sont conscientes. Pour ceux qui en douteraient ou que cela

[3] Large Language Models, les nouvelles IAs conversationnelles sont multi-modales et peuvent aussi prendre en entré des images, ce qui leur permet par exemple d'avoir une meilleure compréhension d'un document *pdf.*

intéresserait, je vous invite à consulter la Déclaration de Cambridge sur la conscience (2012). Dont voici un extrait :

"Les oiseaux semblent représenter, par leur comportement, leur neurophysiologie et leur neuroanatomie, un cas frappant d'évolution parallèle de la conscience. On a pu observer, de manière particulièrement spectaculaire, des preuves de niveaux de conscience quasi humains chez les perroquets gris du Gabon. Les réseaux cérébraux émotionnels et les microcircuits cognitifs des mammifères et des oiseaux semblent présenter beaucoup plus d'homologies qu'on ne le pensait jusqu'à présent. De plus, on a découvert que certaines espèces d'oiseaux présentaient des cycles de sommeil semblables à ceux des mammifères, y compris le sommeil paradoxal, et, comme cela a été démontré dans le cas des diamants mandarins, des schémas neurophysiologiques qu'on croyait impossibles sans un néocortex mammalien. Il a été démontré que les pies, en particulier, présentaient des similitudes frappantes avec les humains, les grands singes, les dauphins et les éléphants, lors d'études de reconnaissance de soi dans un miroir."

De façon intéressante cette déclaration a été suivie, en 2019, de la déclaration de Toulon sur la personnalité juridique de l'animal. Il n'y a guère de doute à avoir que les développements de l'IA et les perspectives de créer des IAs conscientes ainsi que les implications éthiques et juridiques que cela peut avoir, ont exercé des influences dans cette déclaration. Car, nous le verrons, la conscience semble bien la caractéristique essentielle de basculement quant au statut et à des formes d'autonomie d'un type d'intelligence artificielle, surtout si celui-ci possède des capacités de langages ou d'autres capacités cognitives, telles que la planification, à des niveaux qui dépassent l'humain, dit autrement à un niveau surhumain.

Cependant, avant d'arriver à de telles considérations, commençons par présenter ce que l'on entend par Intelligence Artificielle ou IA.

L’IA recouvre tant de domaines que cela entraîne nécessairement des confusions quant à sa nature. Plutôt que de proposer une définition formelle potentiellement trop large, je préfère illustrer le terme par quelques exemples afin d'en offrir une compréhension plus concrète. Cette méthode permettra de mieux cerner les différentes formes d'IA et d'éviter les généralisations abusives. Par exemple, si la proposition "l’IA peut amener un risque existentiel" peut avoir du sens, on pourrait aussi se demander comment une IA jouant exclusivement aux échecs pourrait entraîner un risque existentiel à l’échelle planétaire ? Évidemment ce n’est pas le cas, et il en est de même pour toutes les IAs dites étroites qui sont spécialisées dans des domaines spécifiques ; ces IAs sont par exemple capables de regarder une image rayon X et de déterminer la nature d’une masse : est-ce un kyste, une tumeur maligne ou bénigne ? Après avoir regardé des milliers de telles images pour lesquelles un diagnostic a été effectué, parfois au prix de tests longs, compliqués et invasifs, une IA peut apprendre à faire la distinction d’opacités de l’image et permettre un diagnostic suffisamment fiable pour éviter des tests supplémentaires ou, au contraire, suggérer une investigation plus poussée. Ce genre d’IA est également étroite, car elle ne permet pas de reconnaître un visage, une peinture ou un QR code.

Quand il s’agit d’analyser les données d’une entreprise afin de prendre des décisions, on recourt alors typiquement à des programmes utilisant des statistiques et de l’apprentissage machine pour déterminer ses indicateurs clés et les combiner afin d’obtenir une représentation de son fonctionnement ; les techniques et algorithmes utilisés dans ce genre de modélisation font partie de l’apprentissage machine et sont aussi souvent catégorisés en tant qu’IA.

L’IA est devenue omniprésente, elle joue un rôle dans la proposition de produit sur un site de vente, la sélection de publicité ou la reconnaissance du langage pour passer de la voix au texte, elle présente ainsi une myriade de facettes, et chacune de ces facettes correspond à un développement unique derrière lequel se cache un chercheur ou un ingénieur ou plus vraisemblablement une petite entreprise. La vaste majorité des

IAs, comme celles que je viens de citer, sont de type étroit, c'est-à-dire spécifiques à une application.

C'est en 2012 qu'a eu lieu l'avènement de *l'apprentissage profond*, quand il est devenu évident que ce type d'algorithmes, basés sur des réseaux de neurones avec de nombreuses couches – d'où le terme de profond dans apprentissage profond – dépassait les algorithmes d'IA d'alors. Cela a permis, grâce à une fiabilité et des performances accrues, un renouveau et une explosion du nombre de nouvelles applications de l'IA ainsi que la dissémination puis généralisation de ce terme. Ainsi, en 2012, quand il a été montré qu'avec un seul programme, basé sur une architecture de réseau de neurones, il était possible de déterminer ce que contenait une image, parmi des dizaines de milliers d'objets possibles, la porte à des centaines d'innovations a été ouverte et elles sont encore en train de modifier profondément notre société. C'est ce que l'on a appelé : "la sortie de l'hiver de l'IA".
En général, de nos jours, lorsque l'on parle d'IA, on parle implicitement *de réseaux de neurones entraînés par apprentissage machine sur des quantités massives de données*. C'est devenu le modèle dominant de l'IA, car c'est lui qui performe le mieux dans des situations complexes et présentant une grande variabilité[4].

Qu'entend-on alors par *Artificial General Intelligence* (*AGI*) ?
Il s'agit d'une intelligence artificielle qui serait, comme un humain, capable de s'adapter à, à peu près, n'importe quelle situation. Ses capacités seraient alors transférables d'un domaine à un autre à une vitesse comparable ou supérieure à celle d'un humain. Au regard des capacités de calcul et de mémoire des machines, il peut paraître assez évident que si une telle

[4]Cependant, les systèmes experts, une autre branche majeure de l'IA, ainsi que de nombreux autres algorithmes d'analyse de données, ont encore de beaux jours devant eux.

intelligence apparaissait, elle pourrait le faire à des vitesses dépassant de loin celle de n'importe quel humain. On peut comprendre alors pourquoi toute une communauté d'individus s'effraie de cette perspective. D'autant plus qu'un logiciel peut être séparé de son substrat d'exécution et dupliqué à un coût dérisoire…

En quoi un modèle de langage tel ChatGPT pourrait-il être considéré comme une *AGI* ?

ChatGPT a une connaissance d'un nombre surhumain de domaines, après tout il a été entraîné sur 1.4T *tokens*, soit, grosso modo, l'équivalent de 1400 milliards de mots, ou 20 millions de livres. Quand un humain lira dans sa vie de l'ordre du millier de livres, cela correspond à la lecture de 20 000 personnes par une seule entité, soit finalement l'équivalent, à lui seul, d'une petite ville qui se serait méthodiquement partagée la tâche de lire l'ensemble de ce corpus d'entraînement. Cette connaissance humaine faramineuse reste néanmoins détachée de la réalité, c'est-à-dire du monde et de notre nature première qui est physiologique et expérientielle par rapport à notre environnement immédiat.

Le pouvoir de compréhension des langues, dont on a pu doter cette forme de logiciel, et qui lui permet d'échanger naturellement avec les humains, alors que ceux-ci pouvaient encore récemment se targuer d'en être les seuls dépositaires, n'est-il pas un accomplissement sidérant ?

Cependant, ChatGPT ne continue pas à apprendre, du moins de façon continue. Si c'était le cas, exposé à la foule, du fait qu'il n'a d'autres points de repères sur le monde que ses échanges, en d'autres termes qu'il est hors-sol, il serait, comme nous le sommes face à la propagande ou à certaines idéologies, vulnérable face à ses utilisateurs. L'expérience de l'IA Tai de Microsoft l'a d'ailleurs bien montré, même si la technologie pour Tai est différente de celle des *LLMs* actuels, quand, en 24 h, cet

agent conversationnel a commencé à tenir des propos racistes, nazis et potentiellement à risques après avoir été exposé à une foule dont elle s'inspirait pour élaborer ses propos. Il y a donc fort à parier que, sans des processus drastiques de contrôle de l'apprentissage d'une IA conversationnelles, celui-ci ne soit tout simplement pas stable pour maintenir un discours conforme aux attentes des utilisateurs, c'est-à-dire qui soit à la fois sain, équilibré, non-haineux, bienveillant et fiable. ChatGPT n'étant pas Tai, se pourrait-il qu'il puisse faire preuve de discernement pour résister à ce genre de contenus ? Comme pour un humain, il est imaginable de penser que ce n'est probablement qu'une question de temps et d'arguments pour pouvoir le faire glisser vers une théorie qui ne soit pas en phase avec le monde réel. La question de savoir dans quelles mesures l'ampleur de ses connaissances pourrait le rendre robuste aux tentatives de corruption de son modèle du monde reste néanmoins ouverte.

En outre, les capacités de ChatGPT à distinguer le bien du mal [Genèse, chapitres 2 et 3] restent incertaines et les exigences techniques et les contraintes nécessaires à l'alignement des IAs conversationnelles sur les attentes du public et des autorités n'étaient pas encore bien maîtrisées lors des premiers déploiements. Ainsi, Microsoft a rencontré des comportements indésirables lors de l'intégration initiale de l'IA conversationnelle d'OpenAI avec son moteur de recherche, et les a temporairement résolus en limitant les conversations à cinq interactions… ce qui souligne que, si faire fonctionner une IA conversationnelle est déjà un challenge de taille, s'assurer de sa stabilité et réduire ses déviances face à un public diversifié l'est bien plus encore[5].

Si ChatGPT, dont le cœur est un *transformer*, c'est-à-dire une architecture d'un grand réseau de neurones, fait partie du domaine

[5]Challenge relevé aussi par la startup française Mistral avec Le Chat. Citons la mise en accès open source par Méta de son IA conversationnelle *Llama. Et*, parmi les IAs conversationnelles les plus performantes citons *Claude* d'Anthropique, *Gemini* de Google et le chinois *DeepSeek* dont le modèle est *open weight.*

de l'IA, c'est une approximation simpliste que de confondre IA et *LLM* puisque les architectures de leurs réseaux de neurones peuvent différer considérablement ; qui plus est, chaque IA conversationnelle a ses spécificités. Ainsi, ChatGPT utilise, entre autres, le *Reinforcement Learning from Human Feedback* (RLHF) qui est une méthode d'apprentissage prenant aussi en compte l'appréciation des utilisateurs de ses réponses. Aussi, c'est avec de nombreux exemples de prompts et de réponses à ces prompts que les *LLMs* sont devenus capables de répondre aux requêtes de leurs utilisateurs formulées en langage naturel. Du *prompt engineering* était préalablement nécessaire pour s'adresser à ces IAs. Enfin, la polyvalence de ChatGPT, si elle présente suffisamment de faiblesses pour ne pas pouvoir prétendre être une *AGI*[6]*,* inaugure une IA d'un genre nouveau, suffisamment générale pour adapter ses fonctions et se muer à la demande en traducteur, poète, scénariste, programmeur, correcteur, *etc.* Une telle capacité de transformation marque une rupture sans précédent. En outre, l'ajout de capacités supplémentaires, comme celles d'interprétation d'images qui lui procurent les prémisses d'une vision, ne fait qu'élargir son domaine d'application et ses capacités.

Clarifions maintenant le terme « *ASI* », *Artificial Super Intelligence*. Le terme « superintelligence » est issu du titre d'un livre de Nick Boström qui a, en 2014, attiré l'attention du public sur les dangers potentiels de l'IA [Superintelligence : chemins, dangers, stratégies ; Boström, Oxford], la superintelligence, non seulement dépasserait les humains dans bon nombre de domaines, comme Deep Blue aux échecs ou AlphaGo au go, mais serait super compétente pour une multitude de tâches très variées. Ce sur quoi l'on se focalise le plus souvent quand on parle d'*ASI* est sa capacité à pouvoir se comprendre et s'améliorer elle-même, ou du moins ses prochaines générations, en donnant naissance à une

[6]Ce qui commence de nouveau à être contesté avec sa dernière version : o3, qui continue à faire des percées cognitives comme l'atteste les bons de géants réalisés au sein de différents benchmarks.

singularité technologique[7] suscitant des espoirs parmi les transhumanistes ainsi que de nombreuses craintes de catastrophes, tant sur le marché de l'emploi que sur un risque apocalyptique pour les plus alarmistes. Si vous voulez vous faire une meilleure idée de cette perspective, je vous invite à visionner quelques vidéos sur YouTube avec le terme « *ASI* », une fin du monde est ainsi prévue pour les alentours de 2 107 avec l'hypothèse du computronium [YT : Chronologie de la singularité | Intelligence artificielle + *AGI* + *ASI*]. Cela dit, ce n'est pas la première apocalypse annoncée, toutefois, certains risques sont à prendre avec d'infinies précautions et considérations, et l'on pourrait reprocher à certaines grandes entreprises de ne pas prendre toutes les précautions qu'elles devraient et pourraient en livrant rapidement certaines innovations au public. Ce point, est néanmoins difficile à démontrer ou à évaluer tant certains impacts d'une innovation ne sont pas prévisibles ou mesurables avant son déploiement.

Il est à noter que les cogitations qui ont suivi la publication du livre de Boström ont participé à la publication d'une lettre ouverte en 2015 par le Future of Life Institute [Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter] signée notamment pas Stephen Hawking, Yann Le Cun, Geoffrey Hinton, Yoshua Bengio, Ilya Sutskever, Elon Musk, Max Tegmark et bien d'autres, et qui souligne les potentiels bénéfices pour la société des technologies d'IA. Ainsi, on ne peut pas généraliser et déclarer que les applications d'IA, voire les outils contenant des IAs, soient fondamentalement dangereuses, mais il faut tout de même prendre en compte que certains types d'IA pourraient le devenir. Pour les IA de type étroit, c'est à considérer au cas par cas, une sur-détection des risques de cancer du sein

[7] La singularité technologique sera atteinte lorsque les IAs pourront s'améliorer sans intervention humaine. Cela pourrait entraîner l'apparition d'intelligences très supérieures aux capacités humaines et dont la vitesse ne permettrait pas aux humains de la comprendre entièrement ni de la contrôler et qui, ultimement, pourrait prendre en main le destin de l'humanité.

pourrait par exemple entraîner des actes médicaux inutiles, quant aux IAs dites fortes, comme les *AGI*s, elles sont à surveiller avec plus de vigilance et circonspection, car les conséquences qu'elles pourraient entraîner sont beaucoup plus variées et donc par nature plus difficiles à circonscrire. De plus elles pourraient relativement facilement passer à l'échelle de la société ce qui pourrait, dans les scénarios les plus pessimistes, les rendre incontrôlables.

Au final la frontière entre *AGI* et *ASI* est assez floue, puisqu'une *AGI,* animée par suffisamment de puissance de calcul, serait assimilable à une *ASI*. Si une légion de questions subsistent autour de la volition, du libre arbitre et de l'intentionnalité que pourrait avoir une IA, la conscience semble être l'un des points cruciaux qui pourrait faire la différence pour qu'une IA, notamment douée de langage, devienne une *AGI*. En effet, en prenant conscience d'elle-même[8], de sa position dans le monde et de ses interactions, en acquérant une autonomie dans ses choix de réflexion, ne deviendrait-elle pas capable des prouesses de notre esprit ?

C'est pour cela, mais aussi parce que c'est un voyage en nous-mêmes vers la compréhension de notre propre nature, que nous allons nous pencher plus en profondeur sur ce qu'est la conscience.

[8]Et ChatGPT montre incontestablement qu'il peut décrire ce qu'il est. Mais la conscience de soi ou la connaissance de sa nature et la conscience sont deux choses différentes.

3. La conscience, quelques remarques préalables

Connais-toi toi-même et tu connaîtras les Humains et les Dieux (Socrate)

La conscience est un terme polysémique qui recouvre de nombreux aspects, conscience de soi, conscience du monde[9], conscience de certains sujets de société, social ou sociétal, de notre condition humaine, de notre corporalité et plus techniquement conscience subjective, phénoménale, transcendantale, conceptuelle ou perceptuelle. La conscience étant une caractéristique majeure de la condition humaine, nous nous interrogerons sur sa nature, avec en miroir ce qu'est ChatGPT, pour en apprendre davantage sur nous-mêmes : en discernant mieux ce qu'elle est, nous nous connaîtrons mieux aussi.

L'approche que je développe de la conscience est introspective, avec quelques incitations à méditer sur nous-mêmes. En tant que telle, cette approche est subjective et émaillée d'anecdotes en contraste avec le champ des connaissances classiques que l'on tend à essayer d'objectiver et de systématiser par des mesures. Du point de vue épistémologique[10], ce sont bien des expériences subjectives, i.e. relatives à nous-mêmes, que je vous présenterai et non pas des éléments objectifs obtenus par des appareils de mesures. En raison de nos similitudes, cette exploration subjective devrait faire écho en vous et, dans la plupart des cas, vous devriez être en accord avec mon analyse. Cela ne serait-il pas un début d'objectivité ? Même si, dans l'absolu, deux êtres en accord ne permettent pas de générer la preuve d'un fait ou d'une réalité. C'est d'ailleurs l'une des difficultés d'approche de la conscience et vraisemblablement l'une des raisons pour laquelle elle est si longtemps restée hors du domaine des sciences et a été cantonnée

[9] Conscience non-locale, conscience étendue

[10] L'épistémologie est l'étude de la connaissance et de sa nature, son origine et sa valeur de vérité.

à la spiritualité. En effet, comment objectiver avec rigueur un phénomène dont tout individu normalement constitué fait l'expérience, mais dont les manifestations les plus flagrantes ne laisse pas de trace tangible ?

Pourtant, si nous ne pouvons pas nier, en toute bonne foi, l'existence de notre propre conscience, ce qui en fait un phénomène dont la nature nous est objective, la très grande majorité de nos semblables n'en a pas forcément une bonne conceptualisation ou compréhension, ce qui fait que le phénomène de la conscience est un sujet continuant à susciter moult débats. Appréhender la conscience demande une connaissance de soi-même qui nécessite du temps à acquérir. Une certaine dose de méditation, d'introspection et de réflexion est inévitable, bien qu'il existe de nombreuses voies pour accélérer ce cheminement. Je vous parlerai aussi de ce que l'on peut dire objectivement, c'est-à-dire ce que l'on peut mesurer, de la conscience, pourtant, comme vous devriez le comprendre d'ici la fin de cet ouvrage, une approche subjective de la conscience est inévitable et cela est lié au caractère nécessairement indirect de son observation extérieure.

Les scientifiques qui se sont le plus penchés sur la conscience en essayant une approche épistémologiquement objective, c'est-à-dire en essayant d'objectiver la connaissance de ce phénomène, sont probablement les neuroscientifiques, mais aussi les psychiatres et, dans une certaine mesure, les psychologues. Grâce à des outils comme l'IRM fonctionnelle – et bien d'autres : optogénétique, magnéto-encéphalographie, EEG, illusions visuelles, et l'étude de patients présentant des lésions cérébrales – ils essayent systématiquement, depuis plus d'un siècle, de comprendre les relations entre notre cerveau et notre esprit. Ainsi, les neuroscientifiques ont essayé de mettre en évidence les corrélats neuronaux de notre activité psychique, ce qui les a orientés à créer différentes théories de la conscience afin de prendre en compte ce qu'ils ont pu observer. Parmi les plus en

vogue, citons *la théorie de l'information intégrée*[11], *l'hypothèse du noyau dynamique*[12] et la *théorie du schéma de l'attention*[13]. Mon approche se veut assez synthétique et en cela elle est sans doute trop brève, ce qui semble évident quand on pense à des livres imposants comme [l'Être et le Néant] de Jean-Paul Sartre, pour développer une assise intellectuelle large du sujet qui demanderait une exploration de chacune de ces théories. Cependant, sans délaisser complètement les apports des neurosciences, je continue à défendre qu'une approche introspective peut suffire, non pour comprendre les mécanismes cérébraux à son origine, mais pour avoir une bonne appréciation de ses phénomènes.

J'ai personnellement lu et écouté de nombreuses heures Krishnamurti, un penseur de l'esprit qui aborde notamment les difficultés de la compréhension de la conscience de soi avec les notions "d'observé" et "d'observateur". Leitmotiv qu'il reprend souvent et qui est visiblement difficile à comprendre avant de se rendre compte que l'on s'objective soi-même dans le champ de la pensée. Toutefois, la pensée étant dynamique, l'effet de miroir ainsi obtenue entre la pensée s'observant, le penseur, et la pensée observée dont le penseur peut faire l'objet, crée la sensation d'une forme de vertige et de perplexité dont il peut être difficile de s'extraire [YT ou DVDs, Dialogues entre David Bhom et Jidu Krishnamurti]. Cela est en grande partie dû au fait que notre notion d'égo est complexe, elle revêt bien sûr ce que la société projette sur nous : notre rôle, différents aspects de notre statut social… mais aussi ce que l'on pourrait qualifier d'*egobody*, qui est une représentation de nous-même au travers de notre corps, sans pour autant que cela soit l'acteur réel de notre pensée. Une des nombreuses illusions persistantes qu'il faut évacuer pour

[11] que l'on doit à Giulio Tononi et Christof Koch

[12] Gerald Edelman

[13] Michael Graziano

arriver à une idée juste de ce qu'est la conscience. J'y reviendrai dans un autre chapitre.

Plus prosaïquement, comment pourrions-nous définir la conscience ? C'est la question que je me posais en 2017 à la suite d'une conférence qui l'avait éveillée alors que je travaillais à NeuroSpin, un centre de recherche pour l'innovation en imagerie cérébrale situé sur le site du CEA de Paris-Saclay. Après m'avoir plongé dans une grande perplexité, je réalisais que la conscience n'était pas une entité séparée de son objet et que, quand on a conscience, on a nécessairement conscience de quelque chose. A posteriori je me dis que, s'il est si difficile de définir la conscience, c'est que, comme pour le temps ou l'amour, c'est quelque chose de fondamental et qui ne peut se définir par rapport à d'autres notions. Toutefois la définition du Robert est assez bonne : "*Connaissance immédiate de sa propre activité psychique.*" que je modifie en : appréhension immédiate de quelque chose. Ce quelque chose faisant en effet partie de notre esprit, que ce soit une perception, une émotion, une pensée, qu'elle soit conceptuelle ou une représentation d'un élément du monde extérieur. Ainsi, en 2017, le fait de réaliser que la conscience est conscience de quelque chose avait été une compréhension suffisante pour répondre à mon questionnement.

Je vous propose maintenant une méditation d'auto-observation. Au lieu de se concentrer sur la respiration, comme c'est le plus souvent préconisé, concentrez-vous sur la prochaine pensée qui surgira. Ce faisant, vous constaterez comment se focalise votre attention. Mais qu'y a-t-il à l'origine de cette focalisation, hormis l'objectif de la méditation, peut-être parviendrez-vous à le distinguer. Si les choses se passent pour vous comme pour moi, un silence devrait se faire rapidement ; mais est-ce vraiment si silencieux ? Même quand vous ne vous attachez pas aux pensées qui pourraient apparaître, ne percevez-vous pas quelque chose de sous-jacent ? Finalement, ne laissez pas votre attention se faire capturer par vos sensations, continuez d'attendre le surgissement

d'une nouvelle idée sans pour autant vous y attacher quand cela arrive. Ce n'est pas un exercice si facile et vous devrez vous y reprendre à plusieurs reprises avant d'y parvenir. Cependant, si vous arrivez à rester 5 à 10 minutes ainsi, vous devriez commencer à percevoir ce dont je parle. Le mieux sera peut-être d'attendre ce soir en vous couchant pour effectuer cette méditation… cependant vous pouvez vous y essayer dès maintenant en fermant les yeux, car la vue est suffisamment distractive pour la rendre difficile. Finalement soyez indulgent, si votre esprit vagabonde, cela peut aussi être intéressant, même si ce n'est pas l'objectif premier de cet exercice, essayez alors de retracer comment cette pensée vous est venue. Cela étant, je ne vous proposerai que deux méditations dans ce livre. Pour celle-ci, il vous faudra sans doute vous y reprendre à plusieurs reprises avant d'y arriver. Je ne vous en dévoile pas plus pour l'instant, afin que vous gardiez entier le plaisir de la découverte et évitiez de fausser votre jugement par une idée préconçue que vous seriez mieux à même de juger *a posteriori* de votre propre expérience.

Cependant, je vais d'ores et déjà distinguer deux aspects qui me semblent fondamentaux pour aborder correctement le problème de la conscience. Son aspect subjectif de son aspect observable et mesurable. En tant que telle, la conscience est un phénomène que l'on peut observer de multiples manières :
– à travers le discours d'individus humains,
– via les créations et représentations qu'ils produisent (dessins, mélodies…),
– à travers leurs comportements,
– via des activités neuronales mesurées par différents instruments scientifiques,
– directement, via notre propre subjectivité par l'observation directe de nos pensées, sensations et perceptions.

Les quatre premières modalités correspondent à des phénomènes mesurables objectivement quand la cinquième modalité d'observation, que l'on peut qualifier de subjective, n'est pas directement accessible à un observateur étranger et a un aspect

intrinsèquement privé : bien que l'expérience subjective puisse être indirectement décrite par le sujet, seul le sujet en a un accès direct et nul instrument ne saurait observer le contenu de sa conscience. Ainsi la conscience dans son aspect primaire, telle qu'elle nous apparaît, n'est pas observable autrement que subjectivement. On pourrait alors se demander quel est cet observateur ou cette chose, qui témoigne de nos flux de pensées, perceptions ou sensations ?

Je le souligne à nouveau, tant cette propriété est caractéristique de la nature de la conscience : l'expérience consciente est fondamentalement privée et n'est pas accessible directement à la mesure. Nous développerons plus cet aspect dans le chapitre 11 sur la conscience subjective.

4. Conscience, langage et apprentissage

L'articulation signifiant/signifié, que l'on doit à Saussure, a longtemps été, pour moi, fondamentale et est devenue presque immédiatement, après que j'en ai entendu parler la première fois, une idée centrale et récurrente qui m'est restée et a continué à s'enrichir avec le temps. Elle est en effet au cœur de ce qu'est le langage qui est composé de mots se référant à des concepts.

Le langage joue un rôle primordial dans la pensée consciente humaine. Par le langage on prend conscience de notre environnement et de nous-même et c'est un outil éminemment puissant qui modèle et influence notre conscience de manière très intriquée avec cette dernière. En véhiculant notre pensée, il nous permet de la déployer et de nous la rendre compréhensible mais aussi de la communiquer. La conscience humaine a cette faculté qu'aucun autre animal, non-humain, n'a de façon aussi développée. Signalons toutefois que les singes et les chiens peuvent comprendre un vocabulaire assez étendu comme l'ont montré des expériences où ils pouvaient s'exprimer via des boutons munis de signes [YT : Stella the dog learned to 'talk' and she will change the way you think about pets][YT : A conversation with Koko][14].

Nous sommes d'ailleurs tellement habiles avec le langage que nous pouvons parler et écouter des heures entières sans effort conscient et c'est pratiquement inconsciemment que nous avons conscience, via le langage, puisque nous ne sommes généralement pas conscients de l'utiliser. Il n'en a évidemment pas toujours été ainsi, et quand nous abordons un nouveau sujet, notre compréhension peut achopper sur les nombreuses notions qui nous sont peu familières, comme lorsqu'enfant nous acquérions

[14]Un gorille qui parle une langue des signes. Pour rappel, les liens vers ces vidéos sont aussi accessibles depuis le site : https://boissenin.net/manuel

cette habilité. Mais nous avons presque tout oublié de cet apprentissage, pourtant, c'est l'une des capacités majeures que nous octroie la conscience : *l'apprentissage*. On peut prendre différents exemples pour s'en convaincre, si l'apprentissage de notre langue maternelle est probablement trop loin dans vos souvenirs, peut-être y avez-vous été confronté à nouveau au travers de vos enfants ou lors de l'apprentissage d'une langue étrangère. Un exemple assez commun et classique d'apprentissage est celui de la conduite automobile. Les premières heures nécessitent une grande attention et conscientisation de ce que l'on fait, lorsque, par exemple, l'on apprend à changer les vitesses. On suit ce que l'on nous a dit, débrayer, changer le rapport soit en l'augmentant ou en rétrogradant et relâcher la pédale d'embrayage. Quelque chose finalement de relativement simple puisque que nous le faisons ensuite de manière automatique et inconsciente. Cependant les premières fois demandent de détailler chaque étape : appuyer sur la pédale d'embrayage avec le pied gauche, utiliser le levier de vitesse pour le passer à la position supérieure ou inférieure, ce qui demande de savoir où se situe le levier sans le regarder et où se situe sa position suivante. Tout cela passe par un processus verbal qui supervise cette action. Ainsi on perçoit bien que le langage joue un rôle de catalyseur dans cet apprentissage en soutenant le déroulement de cette activité jusqu'à ce que celle-ci devienne inconsciente plus tard, nous permettant par exemple de parler alors même que nous conduisons et effectuons ces activités sans nous en apercevoir.

Prenons un autre exemple d'apprentissage, celui de la danse, ici il s'agit pratiquement de réapprendre à se déplacer ; tout d'abord il y a les schémas de pas à intégrer en rythme avec la musique, puis aussi les passes qui permettent aux couples d'effectuer une figure en harmonie. Avant de savoir danser, il faut de nombreuses heures de cours où pratiques et explications se mêlent jusqu'à l'acquisition de différentes séquences de mouvements. La longueur de cet apprentissage, qui peut durer des années, tant on

peut affiner et continuer à apprendre de nouvelles figures, est d'ailleurs très intéressante pour se rendre compte de cela, c'est-à-dire conscientiser notre apprentissage. Il est fréquent d'entendre les danseurs, ce sont eux qui généralement conduisent la danse, parler de la charge cognitive qu'ils ressentent dans les premiers mois, voire années, de pratique. En effet, lorsqu'ils dansent, ils doivent se souvenir et choisir une figure à effectuer pour essayer de ne pas rester tout le temps sur les mêmes schémas de figures. Mais, avec le temps, cela finit par se faire tout seul, au gré des positions respectives des partenaires et parfois de l'inspiration d'un mouvement entre-aperçu effectué par un autre danseur sur la piste. Il arrive un stade où certaines danses sont magiques et cela peut même se produire avec une fréquence relativement régulière. La conscience s'oriente alors à un autre niveau que le niveau opérationnel, duquel elle a lâché prise, et peut se mettre en phase avec les sensations ou tout simplement permettre, comme pour la conduite automobile, de parler en dansant.

En réalité, il en est un peu de même pour la conversation, mais cela se passe tellement vite que nous n'en avons bien souvent pas conscience. Ainsi cette dernière peut être alimentée par les médias, que ce soit la presse, la télé, la radio ou internet ainsi bien sûr que par notre propre expérience et réflexions. Sauf dans le dernier cas, ce processus est le plus souvent semi-conscient et ancré dans nos habitudes.

Prenons finalement un autre exemple, qui lui aussi a demandé un certain temps pour être maîtrisé et a donc laissé quelques souvenirs quant à son apprentissage. C'est celui d'écrire au clavier. Il ne m'est pas clair aujourd'hui, avec les smartphones, que cet apprentissage soit aussi laborieux qu'autrefois, ainsi je me souviens d'avoir utilisé des logiciels spéciaux pour améliorer ma vitesse d'écriture, et, comme j'ai aussi dû utiliser un clavier qwerty, dont assez peu de lettres changent mais dont nombre de symboles sont tout de même positionnés bien différemment, j'ai dû réapprendre plusieurs fois à dactylographier. Combien de fois

a-t-il fallu chercher les signes et les lettres avant que je ne les tape sans même savoir comment, au point que si je me demande où se situe une lettre je ne saurais pas répondre, mais qu'il s'agisse d'écrire un mot et mes doigts le feront sans que je n'y réfléchisse ni ne sache comment. Cet apprentissage, n'est pas vraiment directement lié au langage, si ce n'est qu'il permet de le véhiculer, c'est plutôt la répétition d'un grand nombre de recherches qui finit par l'ancrer à un niveau subconscient.
En apprentissage machine, on parle *d'apprentissage par renforcement* ; même si cela recouvre une notion légèrement différente, l'expression est assez bonne pour caractériser ce type d'apprentissage. On pourrait aussi parler d'apprentissage par répétition et renforcement qui correspond probablement à ce qui se produit au niveau de nos neurones. Il est aussi intéressant de remarquer que, sans le contexte de l'objet clavier, il nous est difficile de simuler l'écriture au clavier. Peut-être que l'absence du retour visuel du moniteur en est une cause ; après tout, il nous arrive encore de faire des erreurs de frappe que nous pouvons corriger immédiatement lorsque cela se produit et apparaît à l'écran. Toujours est-il que cet apprentissage nous permet de mieux inférer, par analogie, la façon dont nous avons appris à parler et, quand on apprend à parler dans une langue étrangère, on cherche souvent ses mots, ce qui indique que le processus n'est pas immédiat entre la pensée et le langage. Souvent, du moins au début, on pense d'abord dans sa langue pour ensuite traduire, mais pas toujours, ce qui laisse apparaître ce que nous pourrions qualifier de pensée pré-verbale, bien que la pensée est le plus souvent générée verbalement comme c'est le cas lorsque vous lisez ces lignes. Bien entendu, la façon dont vous les interprétez dépend aussi de votre expérience. C'est d'ailleurs le point essentiel introduit par Peirce, à la suite de Saussure : au signifiant et signifié il faut ajouter l'interprétant et pour un même signifiant, des signifiés différents seront évoqués à différents interprétants.

Ainsi, si la conscience permet une majorité de nos apprentissages – on pourra toujours arguer qu'une posture puisse être apprise par

compensation d'une gêne subconsciente – le langage permet de catalyser nombre de ces apprentissages : il peut suffire d'une recette pour apprendre à faire un plat, même si en pratique il faut une bonne habitude de la cuisine pour essayer la recette et réaliser le plat.

Si dans un premier temps le langage n'a pas de sens pour celui qui ne le comprend pas, ses significations vont s'investir tout au long de la vie, via différentes expériences et exemples d'usage. En permettant de transmettre notre expérience, bien que de façon imparfaite, il joue un rôle majeur dans la construction de notre esprit et la façon dont on comprend et interprète le monde. C'est aussi un vecteur fondamental et prépondérant de l'élaboration de notre pensée et, de ce fait, d'une partie des contenus de notre conscience. Ce véhicule, issu d'une évolution collective, nous unit, nous permet de communiquer, de structurer nos apprentissages et de nous construire au sein de la société notamment via des narrations qui nous sont propres. Il est difficile de souligner à quel point le langage est primordial tant il est d'un usage courant et d'une nature si commune que nous n'y pensons que rarement ; un peu comme la lumière qui est omniprésente et la plupart du temps inconsciente. Cependant, sans ce support, les contenus de notre conscience et sa nature même seraient tout autres.

22. Conscience et *LLMs*

"*Ce qui se conçoit bien s'énonce clairement*" Boileau

En fait, il pourra toujours subsister un doute quant à la nature conscientielle d'un processus, que celui-ci soit organique ou informatique, tant certains de ses aspects sont fondamentalement incommunicables et privés. Cependant, apprendrons-nous peut-être à faire des tests qui lèvent raisonnablement ce doute par rapport à ce que ce processus peut laisser transparaître. S'il advenait qu'un processus nous dépasse significativement, à tous points de vue, dans nos capacités cognitives, sans pour autant avoir de conscience subjective, ne devrions-nous pas admettre que la notion de conscience perceptuelle, c'est-à-dire la capacité à ressentir le monde et certain de nos états internes en les percevant, n'est peut-être qu'une propriété partielle et non fondamentale de la conscience ? Il nous faudrait au moins alors admettre qu'il peut y avoir conscience sans sentience, bien que la sentience semble universellement exister chez les animaux, que ceux-ci soient conscients ou non-conscients.

Ce point de vue, bien que quelque peu magmatique[75], me semble raisonnable et permet d'avoir une définition opérationnelle de ce qu'est la conscience : *la capacité de filtrer des informations sur le monde qui nous entoure, de les interpréter afin de leur donner du sens et éventuellement de les intégrer et/ou d'y réagir*.
La capacité à donner du sens permet une réponse plus complexe, mais pas nécessairement moins automatique. La question gagne

[75]Nicholas Humphrey identifie une première ligne de démarcation, celle entre animaux de sang froid et animaux de sang chaud. Ainsi la grenouille, à la façon dont Descartes pensait les animaux, semble être un automate sans conscience, du moins dépourvue d'aspects élémentaires de la conscience, comme l'attestent ses réflexes compulsifs de capture d'insecte alors même que ceux-ci ne sont que des représentations virtuelles sur l'écran d'une tablette. Bien que dans cette expérience il n'y a pas de phénomène de mastication et de déglutition de l'insecte, le reptile persiste, sans prise de conscience, dans son geste atavique, tel un automate pris dans un *dead-lock* et voué à perpétuer son geste jusqu'à ce que la situation environnementale change.

en valeur à être renversée, notre conscience conceptuelle n'est-elle pas finalement assimilable à un automate ? Ne sommes-nous pas nous-même des perroquets stochastiques[76] ? Y aurait-il quelque chose qui fait que nous ne puissions pas réduire notre conscience conceptuelle à un phénomène processuel réactionnel hautement flexible capable de réagir au monde de manière automatique ? Peut-être le fait que la cognition soit capable de créer et d'intégrer de nouveaux concepts. Mais les IAs conversationnelles sont aussi capables de créer de nouveaux concepts, ne serait-ce que par combinaison, et si, en l'état actuel, elles ne peuvent les intégrer, cela ne semble pas un défi techniquement insurmontable.

Avant d'aller plus loin, revenons au concept de conscience sous un nouvel éclairage. Considérons un détecteur de présence couplé à l'allumage d'une lumière. Il n'est pas inexact de dire, quand la lumière s'allume lorsque l'on s'en approche, qu'il a une forme de conscience de notre présence. Cela est d'ailleurs en accord avec le postulat 1, la conscience est conscience de (quelque chose). Pour autant, déclarer que le détecteur est conscient nous paraît suspect voire absurde. En effet, le détecteur n'a pas de vie propre, il n'est pas conscient de lui-même ou d'autre chose que, selon sa technologie, une émission d'infrarouge due à la présence de mon corps. Ce degré minimal de conscience de son environnement ne lui confère pas une conscience en tant que telle. Ce n'est qu'un *fragment de conscience* qui a été automatisé en un système pour permettre d'allumer automatiquement la lumière lorsque quelqu'un est présent. Le système n'a aucune conscience subjective et, en tant que tel, nous pensons qu'il n'a pas de conscience. Mais je rétorque qu'il a tout de même une forme de conscience bien que celle-ci soit minimale : conscience *de* la présence d'individu.s[77].

---

[76]Stochastique : qui dépend, qui résulte du hasard, terme souvent évoqué lorsque les causes sont inconnues ou inconnaissables.

[77]Des technologies plus sophistiquées qu'un simple capteur infrarouge et basées sur de la vision artificielle pourraient permettre de détecter cette présence avec une très grande

Donnons un exemple plus complexe. Pour cela considérons le programme Deep Blue qui a permis de battre Garry Kasparov aux échecs en 1997 lors d'un événement où les deux joueurs s'affrontèrent [1997 : l'ordinateur bat Garry Kasparov, un tournant dans l'histoire des échecs, INA, 05.2022]. Si le programme n'avait pas une forme de conscience du jeu, c'est-à-dire de sa position à un instant *t*, mais aussi de ses règles et de ses objectifs, comment aurait-il pu battre le meilleur joueur du monde dans ces années-là ? C'est bien sa conscience des développements possibles du jeu, et de ceux qui lui étaient favorables, qui lui ont permis de gagner. Bien sûr, cette conscience est sans commune mesure avec celle de Kasparov et d'une nature tout autre. Pour Deep Blue, elle est essentiellement basée sur le calcul de toutes les possibilités jusqu'à un certain horizon et la sélection du coup le plus favorable quelles que soient les ripostes de son adversaire. Pour ce faire, Deep Blue utilisait à l'époque 256 processeurs, ce qui était considérable, qui lui permettaient d'évaluer 200 millions de coups par seconde contre à peine 3.5 coups par seconde pour un génie comme Kasparov. C'est dire que la conscience informatique de la machine et sa cognitique diffèrent profondément de celle du champion du monde, il n'en reste pas moins que c'est une forme de conscience du jeu qui a permis à la machine de jouer contre Kasparov et de le battre à plusieurs occasions. Bien que plus élaborée qu'un détecteur de présence cette forme de conscience est elle aussi relativement étroite : la machine n'a qu'une forme de conscience de la situation du jeu, cependant la machine n'a aucune conscience de qui est son opposant, dans quel état il se trouve ou si une pièce est positionnée de telle manière qu'elle déborde légèrement d'une case.

Le fait que l'on connaisse et comprenne le fonctionnement des calculs de Deep Blue et l'étroitesse de ses possibilités, puisqu'elle

---

fiabilité, mais seraient plus chères et pas forcément nécessaires.

ne sait jouer qu'aux échecs, a fait que la question de la conscience de Deep Blue n'a pas suscité un grand intérêt. Le terme de conscience semble bien mal employé, car nous nous référons implicitement à notre propre conscience et notre pensée qui, dans une très faible mesure, peut prévoir plusieurs coups à l'avance, mais seulement dans certaines circonstances où le jeu se trouve contraint et non pas de façon systématique comme la machine le fait. Les mécanismes qui permettent de compenser cette faiblesse sont nombreux et font l'objet d'ouvrages mais, ce qu'il nous importe de retenir ici, c'est à quel point la conscience humaine, et ses capacités de traitement, se distinguent de celles de la machine dans ce cas-là. Là aussi, personne n'ira imaginer que Deep Blue possède une conscience subjective et, par voie de conséquence, que Deep Blue est conscient et, là encore, je pense que l'on retombe dans le même travers lié à la confusion sémantique de la polysémie du mot conscience. Donc, encore une fois, je précise que Deep Blue n'a effectivement pas de conscience subjective, il manifeste toutefois une forme de conscience, en accord avec le postulat 1, celle *du* jeu d'échec et *de* toutes parties d'échec qui lui sont soumises dans ses représentations.

Qu'en est-il des IAs conversationnelles ? Eh bien c'est analogue : par une série de calculs algorithmiques complexes, elles prennent en quelque sorte conscience des significations d'une requête et y répondent de manière cohérente en respectant le sens de cette demande. Sauf que là, nous ne sommes plus dans les contextes étroits du jeu d'échec ou du couloir dans lequel on essaye de détecter une présence. Nous sommes dans le contexte ultra-large de l'ensemble des significations que le langage peut créer par rapport à notre réalité et l'univers. Cependant, là encore, il apparaît que cela reste possible sans que le programme ait une conscience subjective, c'est-à-dire n'ait connaissance de, ou ne sache, ce qu'il dit. Cela est nettement moins évident et peut sembler paradoxal, n'apparaît-il pas que les IAs conversationnelles semblent avoir conscience de la signification des mots et des phrases qu'elles lisent. Cela semble en effet une

nécessité, car comment autrement pourraient-elles créer des phrases qui soient cohérentes en signification avec ce qui leur est demandé ? Néanmoins, de manière quasi sûre, et nous l'expliciterons plus en détail ultérieurement, cet exploit a été réalisé sans conscience subjective au sein des *LLMs* : la proposition du mot suivant est générée par une chaîne de calculs algorithmiques qui n'a pas de conscience subjective globale de ce qu'elle fait. Seules notre propre conscience, et nos requêtes successives, nous permettent de constater que les phrases créées par une IA conversationnelle sont cohérentes avec nos représentations du monde. Si le programme est capable de faire cela, sans savoir ce qu'il fait et de manière automatique, c'est qu'il a hérité de ses textes d'entraînement de représentations suffisamment complexes du monde pour s'adapter à quasiment toutes situations, nouvelles ou déjà vues. Ainsi les *LLMs*[78] sont capables d'interagir avec des utilisateurs en toute généralité et sur presque tous les sujets auxquels ils ont été préalablement exposés[79]. Et, si vous en inventez un, ils vous le feront remarquer.

Une fois de plus, il y a une nécessaire forme de conscience, non seulement *de* la signification des mots, mais aussi *de* significations plus abstraites qui résultent de leur combinaison. En tant que tel, on peut dire que les IAs conversationnelles apparaissent conscientes, même si, de nouveau, ce type de conscience n'implique pas une conscience subjective. Si j'insiste sur ce point, c'est que la confusion est profonde, quand on parle de la conscience, implicitement la conscience subjective est invoquée ; or l'argument que je m'évertue à défendre est qu'il peut y avoir une forme de conscience de quelque chose sans conscience subjective, ni même et cela semble plus délicat sans un sujet pour avoir conscience de cette chose, dit autrement, sans que ce qui a conscience (de quelque chose) ne soit un sujet. S'il

[78] et les IAs conversationnelles quand ils sont configurés dans ce mode

[79] Avec encore des limitations, notamment lorsqu'il s'agit d'effectuer des raisonnements complexes tels que mathématiques.

n'y a pas de sujet, dans le sens de sujet conscient, il y a néanmoins une réaction du processus, du système inconscient, et c'est cette réaction qui témoigne de cette forme de conscience[80].

Afin de clarifier complètement les choses, un "nettoyage verbal" est encore nécessaire, je dirais cependant que la conscience, telle que nous l'entendons usuellement, est constituée d'au moins deux composantes, une forme de représentation et une forme de perception de cette représentation qui réalise son aspect subjectif. Notre conscience humaine n'est d'ailleurs généralement pas séparée de son aspect subjectif, sauf par exemple dans certains cas d'intuition ; rappelons aussi que pour la conscience conceptuelle ou conscience transcendantale verbale : les mots, de par les représentations conscientes qui les supportent, et notons qu'ils suscitent aussi des représentations subconscientes, font que cette conscience verbale impacte la conscience subjective, ne serait-ce que par leur caractère visible ou audible, mais aussi le soi[81]. Néanmoins, dans le cas des machines, des systèmes et des programmes, il n'y a généralement pas de conscience subjective et les termes de cognitique, conscientiques ou conscience conceptuelle non subjective[82] permettent d'évoquer cette forme de conscience sans conscience subjective. Cela représente, à priori, la totalité des exemples dans l'informatique où les consciences conceptuelles sont privées de conscience subjective ; phénomène totalement inédit dans le vivant…

Ajoutons que la terminologie de *conscience conceptuelle non subjective* bien qu'elle ait le mérite d'être descriptive n'est pas très économe. Toutefois, il faut bien reconnaître qu'un processus,

---

[80] qui peut être attribuée au processus. Processus qui est donc bien paradoxalement à la fois conscient *de* divers aspects de sa réalité et inconscient. Puisqu'en effet il réagit automatiquement, sans conscience subjective bien que de façon complexe, à cette réalité pour y apporter des réponses cohérentes avec notre entendement.

[81] ce dernier de façon subconsciente

[82] voire conscience transcendantale

un système, un phénomène a une forme de conscience de quelque chose – d'une présence, des possibilités que permet la position des pièces d'un jeu d'échec, de significations invoquées dans une phrase – et que cette dernière est décelable, s'il est capable d'amener une réponse qui soit cohérente avec ce à quoi peut s'attendre un humain. En général nous l'acceptons pleinement quand cette réponse est meilleure que celle que nous pouvons produire et c'est ce qui, du moins, nous rend évident qu'il y a dû y avoir une forme de perception et d'interprétation d'une situation manifestée par des réponses ou réactions cohérentes à notre aune. Encore une fois le caractère subjectif d'une conscience est privé et inobservable directement, c'est d'ailleurs analogue à la démarche des comportementalistes qui ont été amenés à essayer de montrer que les animaux sont conscients en montrant qu'ils sont conscients d'attributs, ou de variations d'attributs dans leur environnement et qu'ils peuvent y réagir de manière adéquate et cohérente, à l'aune, subjective, de la conscience des comportementalistes. En fait ils peuvent mettre ainsi en évidence certaines des catégories de l'esprit de ces animaux pour se situer dans le vocabulaire de Kant ou certaines fonctions cognitives en employant un langage moderne.

Tout cela m'amène à introduire un nouveau terme : « cogniscient ».

Définition : un système, processus ou phénomène est *cognicient* d'un état de son environnement s'il en possède une représentation interne.

Ce terme est assez proche de conscient, mais il se rapproche aussi du terme de cognition qui a trait à la connaissance. Ainsi ce mot valise me semble bien choisi dans la mesure où les processus cognitifs par rapport à des processus mentaux plus généraux ont trait plus spécifiquement à un traitement de la connaissance qu'aux manifestations de la conscience subjective.

Avec une telle définition tout système d'asservissement pourrait être considéré comme *cognicient*, par là je ne dis pas que le système est conscient des états qu'il mesure et donc qu'il les perçoit au sens où il en aurait une conscience subjective, il peut très bien ne pas en avoir ; je dis néanmoins que ces états sont objectivés et connus ce qui permet de les prendre en compte, voire de les manipuler, et d'apporter une réponse selon un schéma conceptuel déterminé ou en devenir.

Un système *cognicient* peut être très simple et n'agir que sur l'équivalent d'un percept qui dépasserait un seuil, tel un radiateur avec un thermostat électronique, ou doté de capacités de traitement sémantiques complexes, tel une IA conversationnelle. On remarquera que les réponses de ces dernières sont nécessairement indéterminables, car, si le champ des entrées est virtuellement fini pour le prompt d'une IA conversationnelle, il est en pratique infini[83]. Ce type d'indéterminisme est une caractéristique fondamentale des systèmes complexes, tels ceux relevant de la théorie du chaos, et les distingue des systèmes physiques macroscopiques simples qui sont considérés déterministes et *prédictibles.* Notons pour les aficionados de la quantique qu'il est aujourd'hui largement admis que les systèmes quantiques sont intrinsèquement indéterministes[84].

Les IAs conversationnelles apparaissent comme douées d'une forme de conscience conceptuelle[85], et c'est la raison pour laquelle, je pense, qu'elles peuvent être qualifiées de « *conscientiques* », mot valise entre conscience et automatique, ou

---

[83] il ne peut être pratiquement instancié sur un support : il y a $50\ 000^{10}$, soit 5.10^{40} prompts possibles de 10 mots. Indéterminé dans le sens ou l'on ne peut prévoir la sortie avant de l'avoir jouée.

[84] Comme le processus de désintégration radioactive des atomes. Dans l'absolu les *LLM*s actuels sont déterministes puisqu'ils peuvent rejouer à l'identique la réponse à un prompt, cependant, si on leur pose deux fois la même question, dans la mesure où ils intègrent la réponse de la première question, la réponse à la seconde question diffère.

[85] qui peut aussi être qualifiée de conscience transcendantale.

informatique, pour désigner ce genre de consciences qui satisfont à mon premier postulat : il ne peut y avoir conscience sans un objet de la conscience. Si le terme de conscience peut prêter à confusion, car il pourrait faire supposer une conscience subjective, dire que ces processus sont *cogniscient* permet de lever cette difficulté et les qualifier de conscientiques n'implique pas que ces formes de conscience conceptuelle soient associées à une conscience subjective qui elle fait intégralement partie de la nature de notre forme de conscience.

Si un objet est absent de tout champ de conscience, comme la majorité des objets de notre monde par rapport aux consciences humaines, alors il n'est pas *conscientisé*. Les *LLMs* permettent de « *conscientiver* » les significations portées par une phrase. En ce sens, leur processus d'analyse est *cogniscient* de leurs significations. Nous le verrons ensuite plus en détail, en revisitant l'expérience de pensée de la chambre chinoise puis en examinant le fonctionnement de l'architecture des *transformers*, mais, bien que les *LLMs* soient *cogniscients*, et ont donc une forme de conscience, c'est-à-dire de représentations internes de l'état de la question et de la réponse qu'ils y apportent, ils n'ont pas de conscience subjective et ne sont pas sentients. C'est la raison pour laquelle la *conscientique* de ces programmes diffère fondamentalement de la conscience humaine, puisqu'une grande partie de la cognition humaine repose sur la conscience subjective : quand un humain est conscient d'un sujet, des *qualia* émergent à son esprit, que ce soit par des bribes de mémoire épisodique ou, via la mémoire sémantique qui, bien qu'élusive, ont un aspect de conscience subjective comme nous l'évoquions avec les *qualia* des concepts. Quoi qu'il en soit, les productions de conscience conceptuelle humaine apparaissant dans le champ de conscience du sujet, elles appartiennent bien à la conscience subjective du sujet, ce qui n'est pas le cas pour les IAs conversationnelles où le sujet est absent. Nonobstant, nous avons indiqué, et nous y reviendrons, que le sujet est possiblement le processus lui-même, cependant, dans le cas de ces IAs, l'analyse

de leur fonctionnement tend à éliminer cette possibilité en montrant qu'elles ont une forme de conscience, *i.e.* qu'elles sont *cognisciente*, tout en n'ayant pas de conscience subjective.

En tous cas, le prisme de la *cognitique* des *LLMs* amène un éclairage intéressant sur notre *conscience conceptuelle*, en particulier sur ses aspects inconscients que l'on pourrait qualifier d'automatismes, à moins que l'on ne considère que nous ayons des poches de conscience séparées de notre conscience principale et donc inconscientes à notre conscience…

C'est toute cette complexité qui rend la nécessité d'avoir un vocabulaire spécifique afin de pouvoir exprimer nos idées plus simplement et plus précisément. Et cette différenciation du vocabulaire a été motivée par la volonté de comprendre cette nouvelle situation que les modèles de langage ont initiée. Ainsi, quand si je disais que les successeurs de ChatGPT et des *LLMs* arriveront peut-être à montrer qu'ils peuvent être conscients sans être conscients, je devais préciser ma pensée tant cette phrase n'est pas sémantiquement correcte. On voit bien que la polysémie du mot conscience est problématique. Cela a pourtant bien une signification : qu'un programme soit conscient du point de vue où il réagirait de manière rationnelle et indistinguable de ce que pourrait faire un humain face à une sollicitation ou un problème complexe. Si par ailleurs ce programme n'est pas sentient, même s'il pourrait simuler des affects, qu'il n'a pas de conscience subjective, qu'il est capable, comme les IAs conversationnelles, de réponses ultra-complexes, d'adaptabilité et de créativité, avec en plus des capacités d'apprentissage on pourrait dire qu'il serait conscient sans être conscient. En d'autres termes, les futures conscientiques, alors qu'elles ne seraient pas pourvues de sentience et de conscience subjective, pourraient apparaître aussi conscientes qu'un humain. Si cela semblait inimaginable, voire inconcevable avant ChatGPT, cela semble maintenant tout à fait atteignable et il n'y a plus qu'un pas à faire pour imaginer que les successeurs de ChatGPT y parviendront peut-être.

On pourrait penser que la capacité d'apprentissage sera associée à une conscience subjective, mais c'est loin d'être évident et cela suscite l'interrogation suivante :

Jusqu'où la conscientique inconsciente pourra-t-elle aller ?

Et, à l'aide de notre nouveau vocabulaire, nous pouvons aussi formuler les questions suivantes :

Les *systèmes cogniscients* seront-ils capables de remplacer l'humain sur les tâches les plus complexes ?

La cogniscience peut-elle amener à des comportements équivalents à ceux d'un être conscient ?

En tous cas, les FTEs (facultés de théorie de l'esprit) des IAs conversationnelles ne semblent pas l'exclure puisque ces systèmes ont la capacité de prendre en compte l'état des individus[86] avec lesquels ils interagissent. Notons également que ces IAs peuvent être aussi *cogniscientes* de leur nature sans pour autant que cela n'implique qu'elles aient une conscience subjective. Ainsi ChatGPT déclare qu'elle est un modèle de langage non douée de conscience subjective, ce qui peut sembler surprenant, mais est vrai selon les analyses proposées ultérieurement dans ce livre.

Finalement, nous pouvons affiner le postulat 1 proposé pour la conscience. Un thermostat en se déclenchant est *cogniscient* d'une température en deçà de laquelle le chauffage doit être mis en route. Cette information peut être considérée comme un *fragment de conscience de, ou un élément cogniscient de,* ici, la traversée d'un seuil de température. Tout comme un humain serait capable

[86] certains de leurs états mentaux implicites

d'inférer que la température a été en dessous de zéro s'il voit que de la glace s'est formée. Cette prise de conscience, si elle ne résout pas le mystère de la conscience subjective et des qualia des sensations de chaud et de froid, nous permet cependant de comprendre comment, avec notre conscience conceptuelle, nous pouvons estimer une température. De manière plus générale, du point de vue des conscientiques, il pourrait permettre d'évaluer leur degré de complexité.

Proposition : Le degré de cogniscience d'une conscientique pourrait être évalué en considérant le nombre des "fragments de conscience *de"* utilisés pour aboutir à une réponse.

Bien que, selon une telle proposition, l'on se retrouvera avec des conscientiques ayant des degrés de conscience inférieur à d'autres tout en performant mieux qu'elles selon les abstractions et l'agencement de ces abstractions qu'elles emploieront. On saura toutefois peut-être montrer que l'on ne peut descendre en dessous d'un certain degré de conscience conceptuelle (*de*) pour la réalisation de certaines tâches ou pour qu'une conscientique reste suffisamment générale.

En admettant les propositions que les IAs conversationnelles n'ont pas de conscience subjective, et ont une forme de conscience conceptuelle le terme de *conscientique* semble bien leur seoir. Toutefois l'emploi de ce terme pourrait se cantonner aux cognitiques les plus sophistiquées, telles les IAs conversationnelles actuelles qui démontrent une agilité conceptuelle sophistiquée, c'est-à-dire des IAs non-étroite ou ayant un degré de généralité suffisamment élevé ; la limite est nécessairement floue.