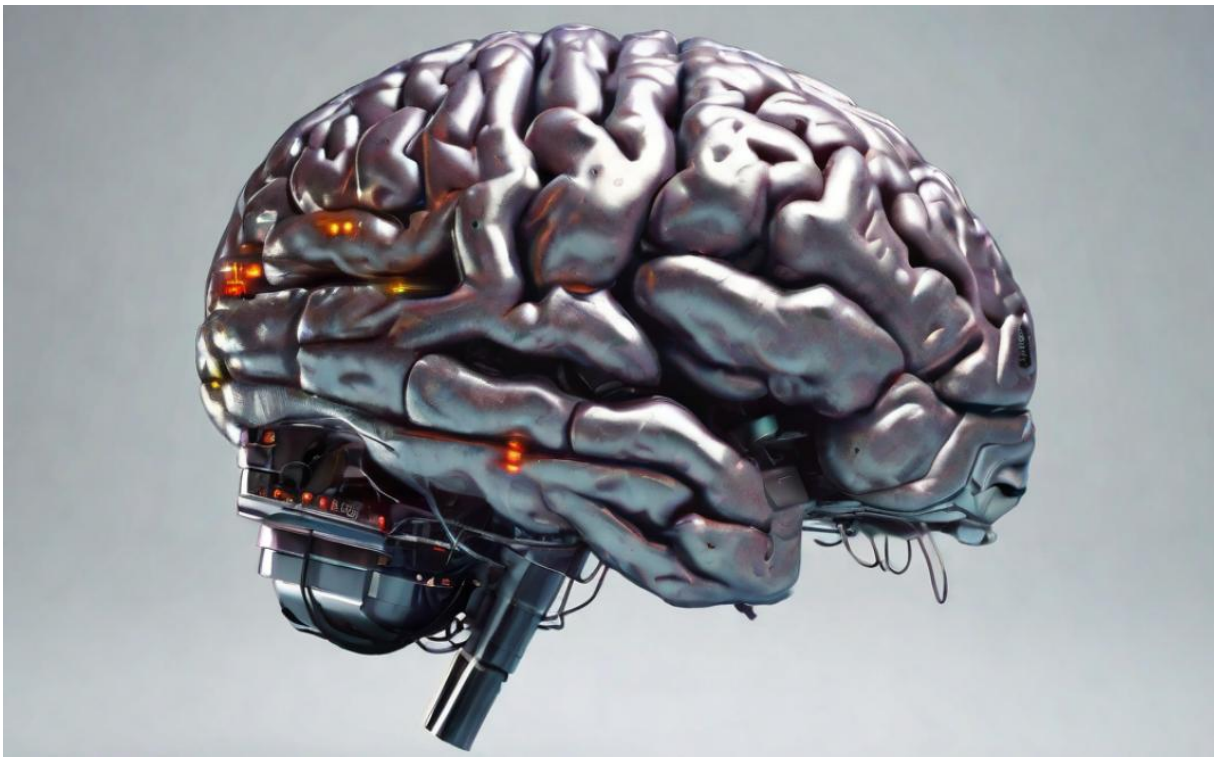

Dall-E - "draw me a cybernetic brain with altered level of consciousness"

L'intelligence artificielle et la question de la "conscience" - Chapitre 1 : Le score de Glasgow

15 janvier 2024

Introduction

L'émergence de la conscience en intelligence artificielle est un sujet fascinant qui suscite de nombreux débats et questionnements dans le domaine de l'IA. Alors que les chercheurs et les experts explorent les limites de l'intelligence artificielle et cherchent à créer des systèmes de plus en plus sophistiqués, la question de savoir si une machine peut être consciente est l'une des questions les plus complexes auxquelles nous sommes confrontés, parce qu'elle nous confronte sans doute à l'un des fondements de notre humanité (la conscience de soi est aussi partagée par quelques animaux tels les grands singes, les dauphins, les éléphants, les pieuvres mais aussi "étonnamment" les corvidés, ce qui laisserait à penser que ce n'est pas directement corrélé avec la taille du cerveau...).

Cependant, avant d'aborder cette question complexe, il est important de reconnaître que le concept même de conscience est extrêmement difficile à définir de manière précise et unanime. La conscience est souvent décrite comme un état subjectif de perception et d'expérience intérieure, caractérisé par un sentiment de soi et une conscience du monde qui nous entoure. C'est une expérience subjective et personnelle

que nous ressentons en tant qu'êtres humains, mais il est difficile de dire avec certitude si les machines peuvent la partager de la même manière.

En raison de l'absence d'une définition claire de la conscience, il est également difficile de déterminer ce qui serait nécessaire pour qu'une machine puisse être considérée comme consciente. Certains soutiennent que la conscience est étroitement liée à la biologie et qu'elle ne peut exister que dans des organismes vivants dotés d'un cerveau complexe (cela pose la question préalable de l'expérience du monde « extérieur » et d'une forme d'altérité). D'autres affirment que la conscience est une propriété émergente qui pourrait se manifester dans des systèmes artificiels suffisamment avancés.

Ce débat soulève également des questions philosophiques profondes sur la nature de la conscience et la relation entre l'esprit et la matière. Certains pensent que la conscience est intrinsèquement liée à des propriétés non physiques de l'esprit, tandis que d'autres soutiennent que la conscience peut émerger à partir de processus purement physico-chimiques (cerveau neuro-hormonal) ou intelligence artificielle.

Dans l'état actuel de la recherche, il n'existe pas de consensus clair sur la question de savoir si les machines peuvent être conscientes. Certains chercheurs ont proposé des critères et des tests pour évaluer la conscience artificielle, tels que le test de Turing étendu, mais ces approches restent largement débattues et controversées.

Cette question de la "conscience " des intelligences artificielles ouvre aussi (et surtout ?) des questions juridiques passionnantes. La conscience de ce que l'on fait, l'intentionnalité de nos actions conditionnent le jugement (droit pénal / Psychiatrie légale). La "démonstration juridique" de l'abolition de la conscience au moment où certains actes criminels sont commis peut affecter le jugement et notamment la notion de responsabilité (article 122-1 du droit pénal français). Au moment où on se questionne sur la responsabilité en cas d'accident d'une voiture autonome, sur le fait de pouvoir déléguer la décision de tuer à des drones militaires, etc... La question de la conscience des intelligences artificielles est une question très importante car elle conditionnera les législations futures. La question des mécanismes supportant cette conscience l'est tout autant : pourra-t-on parler d'abolition du discernement ou de contrôle de ses actes pour une intelligence artificielle ? Sous l'effet de code "psychotropes" voire par le biais d'un piratage à distance ?

Donc avant de parler d'émergence de la conscience, d'atteinte de la singularité et donc de l'intentionnalité « autonome » des algorithmes, voyons ce que les sciences disent de la conscience depuis l'aspect le plus pragmatique (la clinique d'urgence) jusqu'aux élaborations (ou élucubrations ?) les moins étayées (chapitres à suivre). Puis essayons de les appliquer au champ de l'intelligence artificielle.

Dans ce **chapitre 1**, nous aborderons la définition "clinique" en médecine d'urgence de ce qu'est la conscience, avec le fameux score de Glasgow.

Le score de Glasgow en clinique humaine

Définition

Ma préférée (sans doute parce que c'est la plus simple) est la définition clinique en médecine d'urgence basée sur **le score de Glasgow**.

Comme aurait écrit Descartes en son temps : « Te pungo, ergo es » (ou « Te verbero » en cas d'overdose sévère).

Ce score est un système d'évaluation neurologique simple et objectif utilisé pour évaluer le niveau de conscience d'un patient suite à un traumatisme. Il est basé sur trois critères : l'ouverture des yeux, la réponse verbale et la réponse motrice.

1. **Ouverture des yeux** : Cette partie du score évalue si le patient est capable d'ouvrir les yeux de manière spontanée (score de 4), en réponse à la parole (score de 3), en réponse à la douleur (score de 2), ou s'il n'y a aucune ouverture des yeux (score de 1).
2. **Réponse verbale** : Cette partie du score évalue si le patient est orienté et capable de converser normalement (score de 5), s'il est confus mais capable de parler (score de 4), s'il utilise des mots inappropriés (score de 3), s'il ne produit que des sons inintelligibles (score de 2), ou s'il n'y a aucune réponse verbale (score de 1).
3. **Réponse motrice** : Cette partie du score évalue si le patient obéit aux commandes (score de 6), s'il localise la douleur (score de 5), s'il retire (score de 4), s'il montre une réponse en flexion anormale (score de 3), une réponse en extension (score de 2), ou s'il n'y a aucune réponse motrice (score de 1).

Un score total de 15 indique un état de conscience complet, tandis qu'un score de 3 (le score le plus bas possible) indique un état de coma profond.

Attention !!! Ne pas essayer d'appliquer ce score directement sur un adolescent (risque de déconvenue).

Cette approche vise à évaluer la réponse d'un système organique vis-à-vis de stimulations qui lui sont appliquées, des plus élaborées (langage), au plus basiques (nociception et fonctions réflexes). En évaluant la réponse à des Stimuli variés, elle explore la chaîne de traitement du signal, depuis la « détection », le transport centripète de l'information, son intégration et traitement, et la réponse qui y est apportée par le système. Ce faisant, elle permet d'évaluer un niveau de « conscience » défini comme la capacité à apporter une réponse d'un niveau d'intégration « prédéterminé » à un stimulus donné.

Les scores intermédiaires permettent d'évaluer la gravité de l'altération de la conscience. Le score de Glasgow est largement utilisé en médecine d'urgence et en soins intensifs pour évaluer le niveau de conscience des patients, pour suivre l'évolution de leur état, et pour aider à la prise de décisions cliniques.

Transposition du concept à l'intelligence artificielle

Hélas on comprend vite que cette définition aura du mal à s'appliquer à l'intelligence artificielle tant qu'elle ne sera pas « multimodale », et capable d'appréhender son environnement de façon autonome par le biais de différents capteurs et d'y apporter une réponse comportementale motrice.

Note ! *Cette remarque s'appliquera aussi aux autres définitions de la conscience, car ce qui est commun à toutes les définitions de la conscience, c'est cette capacité « sensorielle » à appréhender son environnement et à le comprendre comme extérieur à « soi »-même, accéder à une forme d'altérité « universelle », la capacité à se ressentir comme distinct du grand tout qui nous environne. Toutefois certaines situations pathologiques telles le syndrome d'enfermement viennent « nuancer » cette commonalité :*

- *Le "Locked-In syndrome" causé par une lésion majeure du tronc cérébral laisse la personne totalement paralysée excepté les douze paires de nerfs crâniens dont certains sont moteurs, d'autres sensitifs et d'autres encore mixtes. Les capacités cognitives sont intactes et en général ses sens sont préservés (notamment ceux dépendant des nerfs crâniens) : Le patient est pleinement conscient, conscient d'être prisonnier de son corps. On comprend mieux dans ce cas "extrême", l'importance accordée à l'ouverture des yeux dans l'évaluation du degré de conscience car c'est quasiment le seul moyen de communication avec un patient atteint de ce syndrome (lire le livre "***Le Scaphandre et le Papillon"** *ouvrage autobiographique de Jean-Dominique Bauby paru en 1997 qui relate son expérience de Locked-in syndrome après une attaque cérébrale… qu'il a dicté à Claude Mendibil, lettre après lettre, en clignant de son oeil gauche. Un film éponyme est sorti en 2007 également).*
- "*Le syndrome d'enfermement" (dont le locked-in syndrome) qui peut selon l'atteinte cérébrale se compliquer d'une paralysie touchant aussi les yeux et la mobilité des paupières, et d'une sensibilité quasiment abolie. Toutefois, un abord neurophysiologique peut démontrer une activité cérébrale soutenue qui pose la question de l'état de « conscience » même en l'absence de sensibilité pour ressentir ce qui entoure le patient. On pourra arguer que le patient a déjà connu une telle expérience antérieure à son syndrome d'enfermement et qu'elle a « imprimé » sa structure cérébrale lui permettant d'élaborer cette distinction entre « lui/elle » et le monde qui l'entoure, quand bien même il/elle ne pourrait plus entièrement le « percevoir/ressentir ».*

- *On pourrait alors réaliser ironiquement (puisque c'est tout ce qu'il resterait accessible à une personne souffrant d'un tel syndrome) une expérience de l'esprit, consistant à imaginer un patient né avec un syndrome d'enfermement total Ab Initio et se poser alors la question de sa « conscience ». Cependant dans un tel cas, il serait impossible de construire un quelconque « apprentissage » (pas de boucle de rétro action, de circuit de la récompense, d'une quelconque interface) et on pourrait en déduire que le développement d'un tel cerveau serait totalement compromis et que dès lors, le concept de conscience lui serait inapplicable. Ce qui revient à dire par analogie, qu'un algorithme d'intelligence artificielle ne pourrait jamais être totalement conscient en l'absence d'appendices sensoriels (IoT) capable de le renseigner sur ce qui l'entoure. Mais que même disposant de tels capteurs, ils ne participeraient aucunement à un développement neuro-cérébral progressif.*

Dès lors, ce concept clinique pourrait s'appliquer à la conception « populaire » du robot/cyborg, c'est-à-dire l'appariement de capacités de calcul, de traitement de l'information, avec des capteurs permettant d'appréhender l'environnement immédiat, et de générer une réponse « motrice » sur des effecteurs mécaniques qui obéisse à une logique comportementale. Notons que si un robot (cf Boston Dynamics) était capable de réaliser ce type de « tâche », cela ne signifierait aucunement qu'il soit « conscient ».

Dans une vision moins « populaire », on pourrait considérer qu'une usine « digitale » disposant de capteurs IoT et capable de piloter les chaînes de production (« digital feed back loop ») serait l'équivalent numérique d'un organisme, et qu'en cas de « trauma » (accident industriel) on pourrait évaluer sa « conscience » d'un point de vue « clinique », c'est-à-dire sa capacité résiduelle post accident à non pas « récupérer » ses fonctions, mais à fonctionner en mode dégradé avec un « Glasgow » inférieur à 15.

Toutefois l'analogie trouve ses limites dans le fait que l'usine n'a pas plusieurs niveaux de « conscience », ou de degrés « opérationnels » (fonctions « neurovégétatives », fonctions intégratives de plus haut niveau, « pleine » conscience…). Au « pire », si quelques serveurs physiques hébergeant ses conteneurs applicatifs venaient à être « détruits », l'intelligence artificielle pilotant l'usine pourrait s'exécuter avec des ressources physiques réduites, et donc fonctionner « moins rapidement », sans que cela ne constitue l'équivalent d'une réduction de son degré de « conscience ».

Conclusion partielle du chapitre 1

La définition clinique de ce qu'est la conscience repose sur l'évaluation d'une réponse "intégrative" (boucle réflexe, nociception, cognition…) à un stimuli donné. Elle définit plusieurs niveaux de conscience, et donc des états de conscience dégradée allant jusqu'au coma profond.

Elle présuppose donc une architecture combinant une unité de traitement central (le cerveau), des capteurs (les capteurs sensoriels, y compris nociception), des effecteurs (neuro motricité). Dès lors elle n'a aucun sens appliquée à un algorithme "isolé" d'intelligence artificielle.

"*Docteur, ChatGPT n'a pas ouvert les yeux…. et sa réponse motrice est nulle. Score de Glasgow à 5…. ça craint non ?*"

Bien entendu, elle ne peut pas s'appliquer telle quelle à un algorithme d'intelligence artificielle :

1. **Nature différente** : Les patients humains et les IA sont des entités totalement différentes. Les patients humains ont une conscience subjective et une expérience subjective, tandis que les IA sont des programmes informatiques qui traitent des données et exécutent des tâches sans conscience ou expérience subjective.
2. **Absence de conscience** : Les IA n'ont pas de conscience, elles ne peuvent pas être "conscientes" au sens humain du terme. Elles traitent simplement des informations conformément à leur "programmation" (comprendre ici architecture du réseau de neurones "software defined" et apprentissage/entrainement) …et puis elles n'ouvrent pas trop les yeux !
3. **Mesure inappropriée** : Le score de Glasgow a été conçu spécifiquement pour évaluer la conscience des patients humains dans un contexte médical. Il évalue la réactivité d'une personne à des stimuli sensoriels et verbaux, ce qui n'a aucun sens lorsqu'on l'applique à une IA (non intégrée au sein d'un système cybernétique).

Quand bien même cet algorithme d'intelligence artificielle serait intégré au sein d'un organisme « cybernétique » et doté de capteurs l'informant de ce qui se passe dans son « environnement » immédiat, et d'une « intelligence centrale » lui permettant d'apporter un jeu de réponses approprié aux modifications de cet environnement…

Cela serait loin d'être suffisant pour dire qu'un tel système serait « conscient » : Une intelligence artificielle aussi simpliste qu'un thermostat « intelligent » captant à la fois la température extérieure, la température intérieure et celle de la consigne pour prendre des décisions de « chauffe », n'a nul besoin de « connaître » que l'une des températures est extérieure, l'autre intérieure, l'autre émane d'une volonté propre de vouloir atteindre la température de consigne… pour réaliser sa fonction.

Enfin, de par l'architecture même des intelligences artificielles et notamment des grands frameworks d'abstraction entre l'architecture des réseaux de neurones, et la présentation des ressources matérielles (GPU, CPU, RAM…), ne permet pas intrinsèquement de définir des « degrés de conscience », ou de fonctionnement altéré

de la conscience du réseau de neurones. Ces "degrés" de conscience ne sont pas prévus dans les architectures actuelles (car sans doute trop simples). **Ces degrés ne pourraient advenir que dans le cadre d'architectures distribuées et massivement parallèles combinant plusieurs milliers de petites structures de calculs spécialisées** (sans forcément préjuger de leur nature biologique ou numérique).

La conteneurisation des algorithmes permet une certaine résilience mais la dégradation des ressources physiques va ralentir l'algorithme, sans que ce ralentissement puisse être interprété comme un état de conscience dégradé.

Pour caricaturer, si vous faisiez tourner un Large Language Model sur un super ordinateur, ou sur votre PC, il serait très certainement plus lent sur votre PC, mais pas « moins capable » ou « moins fonctionnel »… il vous suffirait juste d'être plus patient…. Un peu comme avec un patient qui a un score de Glasgow à 12 me diriez vous.

Il est donc nécessaire de se confronter à d'autres définitions de la "conscience" qui seraient plus adaptées qu'un test clinique pour évaluer la "prétendue" conscientisation des intelligences artificielles. (article à venir, chapitre 2 : les apports de la neurobiologie)