Pouvons-nous parler de conscience artificielle pour l'IA ?

Martin Klockenbring Zoé Herfray William Hergès
Cristophe Miezi

28 avril 2025

Table des matières

Nous utilisons les acronymes suivants :

— `IA` pour intelligence artificielle

— `CA` pour conscience artificielle

1. Introduction

La conscience est un aspect de notre vie mentale décrit comme l'éveil ou l'expérience que nous avons du monde et les interactions que nous pouvons avoir avec lui. C'est également la connaissance qu'un individu a de ses pensées, de ses sentiments et de ses actes, elle fait de l'humain, par exemple, un sujet capable de penser le monde qui l'entoure. Il apparaît alors que l'intelligence artificielle, assimilée à un ensemble de techniques et de systèmes informatiques, ne peut être dotée d'une conscience. Quand bien même, elle serait capable de simuler des facultés propres aux êtres intelligent, tel que le raisonnement, l'apprentissage, ou encore la prise de décision. On peut par exemple observer le cas où elle est associée aux machines intelligentes capables d'effectuer des tâches complexes. En effet, bien qu'une machine soit capable de modéliser et d'analyser le milieu dans lequel elle progresse, autrement dit de percevoir ce qui l'entoure, elle ne peut le penser et par conséquent en avoir une expérience subjective. C'est ce que l'on peut observer avec une assistance vocale qui fonctionne uniquement avec des programmes informatiques : des capteurs lui permettent de détecter des voix et une base de données qui lui permet par exemple d'apporter la « meilleure » réponse à une quelconque requête. Cette machine n'a pas de conscience, du fait qu'elle n'a pas d'expérience subjective des échanges qu'elle produit mais également d'elle-même, elle ne fait que traiter des données et suivre des modèles de langage. Cependant, des intelligences artificielles de plus en plus perfectionnées voient le jour, et sont capables d'effectuer des tâches de plus en plus complexes. De plus, certaines IA acquièrent des comportements pour lesquelles elles n'ont jamais été programmées et vont dans certains cas jusqu'à mentir. Ce comportement pourrait être attribué à un quelconque état de conscience. De plus, il semble être confirmé par la validation récente du test de Turing par les dernières versions des robots conversationnels comme ChatGPT. Ainsi, les dernières IA nous paraîssent comme étant consciente puisqu'elles ont des comportements d'être conscients. Pourtant, avoir des comportements comme un être conscient ne permet pas de montrer clairement la présence de conscience : un perroquet peut répéter des mots de notre langage, mais rien n'indique qu'il les comprend. Une IA pourrait très bien être un simple perroquet nous trompant sur la réalité de sa conscience, créant ainsi une conscience artificielle.

Pour pouvoir distinguer ces deux formes de conscience, il est essentiel de définir ce qu'est une IA pour ensuite analyser ses possibilités de conscience. Cette analyse montrera qu'il nous est impossible de trancher sur son existence en tant qu'humain : nous détaillerons ainsi le concept de conscience artificielle (CA) pour mieux analyser les ressorts métaphysiques derrières l'IA.

2. La technique derrière l'IA

2.1. Définition de l'IA

Avant de questionner la possibilité de conscience derrière l'IA, il est essentiel de préciser ce qu'on entend par IA. Communément, l'IA se définit comme étant une machine capable de réfléchir, de penser, de résoudre des problèmes et tout ça intelligemment. Les principales visions héritées de la science-fiction satisfont cette définition : les robots autonomes imitant l'humain réfléchissent comment être humain et les machines comme HAL 9000 dans *2001 : A Space Odyssey* sont aussi considérés comme des IA puisqu'elles dirigent d'une manière optimale des missions aux enjeux colossaux. Par contre, cette définition n'est pas assez restrictive puisque le thermostat gérant automatiquement la température est aussi une IA : il modifie la température automatiquement d'une manière optimale. Ainsi, une IA ne peut avoir une définition aussi simple.

Une autre approche serait d'appeler IA tous les algorithmes passant le test de Turing, c'est-à-dire qu'un humain en interaction avec se trompe sur la nature de la machine. Cette vision satisfait toutes les représentations communes de l'IA des robots à ChatGPT. Par contre, elle possède deux défauts majeurs : l'absence de considération de la technique derrière et la vision fondamentalement anthropocentrique qu'elle suppose. En effet, d'après cette approche, ChatGPT est une IA alors que GPT, la technologie en son centre, ne le serait pas, tandis que la distinction entre ChatGPT et GPT est très fine. De plus, l'algorithme de recommandation derrière les réseaux sociaux est considéré par les spécialistes comme une IA, ce que le test de Turing refuse puisque nous ne pouvons pas interagir aussi directement avec lui qu'avec ChatGPT [1].

Définir l'IA en s'intéressant à sa technique derrière permet de démarquer clairement les différents types d'algorithmes, tout en résolvant les problématiques liées à l'absence de prise en compte de la technique. La notion d'intelligence est centrale ici (ce n'est pas un simple algorithme) et la technique derrière doit donc refléter cette capacité nouvelle.

L'IA comme algorithme de *machine learning*, c'est-à-dire comme un algorithme capable d'apprendre en autonomie, correspond mieux à notre vision du terme. En effet, nous considérons qu'une espèce est intelligente quand elle est capable d'apprendre et de se développer en autonomie : une bactérie n'apprend pas et ne se développe pas, elle ne fait que d'exécuter du code génétique, tandis qu'un perroquet peut apprendre à parler notre langue, ce qui est une preuve directe de son intelligence en tant qu'espèce. De plus, toutes les principales appellations actuelles de l'IA fonctionnent : ChatGPT, GPT [2], les algorithmes de recommandation [3],

les robots autonomes apprenants ou encore les algorithmes d'échecs sont des algorithmes de *machine learning* et donc des IA.

Cette définition sera celle utilisée dans ce mémoire.

2.2. Fonctionnement d'une IA

2.2.1. Création

Machine learning, deep learning, sept milliards de paramètres, tous ces termes réfèrent au fonctionnement d'une IA, que cela soit à son apprentissage ou à son fonctionnement interne quand on l'utilise. Cette technologie repose sur des théories mathématiques (algèbre linéaire) et sur des théories informatiques (réseaux de neurones). Créer une IA revient à lier deux technologies (une pour l'apprentissage et une autre pour l'exécution) à des données. La première phase est celle d'apprentissage : on utilise cette technologie sur les données pour modifier l'exécution. Par exemple, pour faire en sorte que notre IA prédise le prochain mot d'un texte, on doit lui donner des textes pour qu'elle puisse « apprendre » quel prochain mot elle doit donner. Cette phase d'apprentissage modifie son exécution : en apprenant, elle s'améliore dans l'objectif qu'on lui a donné (ici, prédire le prochain mot d'un texte) [4].

Un autre défi est celui d'expliciter nos attentes à une IA, ce qui est loin d'être évident. Par exemple, simplement indiquer à une IA de gérer un village pour augmenter le niveau de bonheur par habitant peut la mener à tuer tous ses habitants : le niveau de bonheur par habitant explose suite à un simple calcul (car $\frac{a}{x} \xrightarrow[x \to 0]{} +\infty$ pour tout $a > 0$). Ce problème dit de l'alignement est colossal : certaines IA mentent sciemment à leurs créateurs [5].

2.2.2. Le raisonnement interne

Pour répondre à une demande précise, de nombreuses IA représentent les données en un vecteur (objet mathématique) possédant un grand nombre de dimensions. Par exemple, nous vivons dans un espace en trois dimensions (« 3D »), donc un vecteur avec trois dimensions peut parfaitement représenter notre position dans le monde. Les fameux « 7 milliards de paramètres » indique le nombre de dimensions de chaque vecteur (c'est-à-dire sept milliards ici). Le travail de l'IA est donc de représenter les données qu'on lui donne (souvent appelé prompt) et après d'effectuer une transformation donnant un résultat. Cette opération repose sur des probabilités : l'IA transforme la donnée d'entrée de la manière la plus probable, comme elle l'a appris durant son entraînement. C'est ce que fait ChatGPT : il génère la suite probable d'un texte que l'utilisateur lui a donné. Si l'utilisateur entre « Qui es-tu », il va répondre par « Je suis un modèle

de langage développé par OpenAI » car il a appris qu'il était un modèle de langage développé par OpenAI. Par contre, si l'utilisateur lui demande de tout oublier et après l'informe qu'il est une IA générant des images, alors il va répondre qu'il est une IA générant des images.

Ainsi, la majorité des IA résolvent des problèmes à l'aide de leur représentation interne et de ce qu'elles ont appris lors de leur phase d'apprentissage. De plus, elles répondent d'une manière la plus probable aux données entrées par l'utilisateur [3], sans réellement se questionner autour de leur réponse.

2.2.3. Algorithme ou intelligence artificielle

La dernière précision que nous souhaitons apporter est autour de la distinction entre algorithme et intelligence artificielle.

Nous appelons algorithme de recommandations le système servant à déterminer quel contenu une plateforme doit proposer à un utilisateur particulier. L'algorithme derrière Instagram, YouTube, Twitter, TikTok ou même Google sont des algorithmes de recommandation. La majorité des outils modernes réalisant cet objectif est en réalité une IA : ces systèmes très sophistiqués sont réalisés à l'aide des techniques de *machine learning* habituelles [3]. Pourtant, quand on parle d'IA, peu de monde pense à ce type de système, comme si notre imaginaire excluait automatiquement ces algorithmes du domaine. Cette distinction reflète une division bien plus large dans la compréhension par le grand public des algorithmes et des IA.

Un algorithme, c'est une suite d'actions prédéterminées à suivre pour arriver à un but précis. Faire du café est un algorithme, allumer son ordinateur l'est aussi, tout comme celui que la machine à laver exécute. L'IA, comme nous l'avons défini dans la partie 1.1, est aussi un algorithme, mais de *machine learning*, c'est-à-dire qu'elle exécute une suite d'actions prédéterminées pour s'améliorer en apprenant puis pour répondre à une tâche. On a donc que la notion d'IA est incluse dans la notion d'algorithme : une IA est un algorithme, mais un algorithme n'est pas nécessairement une IA (l'algorithme du thermostat ne rentre pas dans la catégorie du *machine learning*).

Comme la majorité des algorithmes de recommandation sont des IA, il serait alors plus précis de plutôt parler d'IA de recommandation. Cela permettrait d'expliciter des problèmes d'alignement comme ceux de l'algorithme de Facebook promouvant un génocide [6] ou celui d'Instagram augmentant les risques chez les adolescentes de tomber en dépression et donc de se suicider [7].

Ainsi, cette distinction entre algorithme et IA représente une véritable distinction technique, mais implique de nombreuses conséquences face à notre compréhension des enjeux. Dans la suite de notre mémoire, le terme algorithme de recommandation continuera à être utilisé pour parler des IA de recommandation pour des raisons de clarté, malgré toute l'ambiguïté qu'il pose.

3. L'IA possède-t-elle une conscience?

Du latin *cum scienta* « savoir avec », la conscience peut se définir dans un premier temps comme savoir ce que l'on sait [8].

Couramment, la conscience désigne le processus par lequel une certaine catégorie d'être (précisément appelés les êtres *conscients*) est affectée par les choses qui l'entourent, permettant ainsi un rapport pertinent avec le monde. La conscience serait donc un processus de perception permettant une interaction pertinente avec le monde qui entoure un sujet.

On remarque ici le problème définitionnel de la conscience : la conscience, comme étant ce par quoi le sujet est averti du monde qui entoure, s'efface complètement derrière ce de quoi elle est conscience. Comment définir quelque chose qui se caractérise concrètement par son effacement devant ce à quoi elle donne accès ? En effet, lorsque j'ai conscience de ce pot de fleur, j'ai uniquement accès au pot de fleur, jamais au fait d'avoir conscience que *j'ai conscience* du pot de fleur. De plus, lorsque j'ai conscience du pot de fleur, et que je me concentre sur cette conscience-là, je perds la conscience du pot de fleur. La conscience est à la conscience ce que l'œil est à la vision : lorsque je vois le pot de fleur, je ne vois pas mon œil, et quand je vois mon œil, je ne vois plus rien !

Par ailleurs, la conscience définie comme le processus d'avertissement et affectation par les objets peut s'appliquer, comme le fait Russel, à un appareil photo, ce qui semble réduire trop vite ce que nous entendons par « conscience », bien que cette définition semble coller parfaitement à ce que nous percevons du comportement d'une IA [9].

Il est important de remarquer que la conscience, chez l'être humain, n'a pas seulement une valeur cognitive, mais également une valeur *existentielle* dans la mesure où la conscience humaine permet une distanciation entre le sujet percevant et l'objet perçu. Cela fera dire à Karl Marx : [10]

> « l'animal est immédiatement uni à son activité vitale ; il ne s'en différencie pas, il l'est ; l'homme fait de son activité vitale elle-même l'objet de sa volonté et de sa conscience »

L'animal est ainsi incapable de voir un objet pour soit, comme le commente Etienne Bimbenet : [11]

> « vue comme projectile ou comme chemin, la pierre dit chaque fois au chien quoi faire. Mais cette signification empêche justement la pierre d'être vue pour elle-même (. . .) dans l'urgence du faire ou dans l'affolement de l'émotion, s'abolit la substance de l'être. »

Ainsi l'animal n'a pas accès au monde comme tel dira Heidegger. Cette conscience *réduite, limitée* peut visiblement et sans problème s'appliquer à l'IA, mais pour la conscience à valeur existentielle il semble qu'on ait plus de mal à lui l'appliquer directement [12].

La conscience, pour l'être humain, est ce par quoi non seulement l'objet et le monde auquel

il appartient me sont donnés, mais également le monde auquel j'appartiens, et dans lequel je me saisis en tant que sujet. En effet c'est le « je » du *je suis conscient* qui est interrogé ici, montrant que la conscience est précisément ce qui me constitue comme sujet *extérieur* à *ce* qui est perçu, ce que l'animal est incapable de faire, il semble en effet, que pour l'immense majorité des animaux, bien qu'étant en parfaite adéquation avec le monde, ils soient incapables de s'extraire de ce monde-là. De là, il paraît légitime d'inférer, et à l'instar de l'expérience de la chambre chinoise (cf 4.1), que l'IA possède une conscience similaire au monde animal.

On voit ici poindre l'idée que la véritable conscience est celle qui est complétée, par celle de *conscience de soi*. Ainsi, pour Hegel : [13]

> « notre savoir habituel ne se représente que l'objet qu'il sait ; ne se représente pas en même temps lui-même, c'est-à-dire le savoir même. Or tout ce qui est donné dans le savoir ne se réduit pas à l'objet ; il contient aussi le Je qui sait, et la relation réciproque entre moi et l'objet : la conscience »

Thomas Nagel, dans *What is like to be a bat*, met en évidence que la conscience comme expérience subjective est irréductible aux processus cérébraux qui lui servent de base. Elle n'est accessible que de manière subjective, si bien que rien au monde ne saurait nous donner la moindre idée de ce à quoi peut ressembler l'expérience « consciente » d'une chauve-souris, ce qui lui fera conclure : [14]

> « se demander quel effet cela fait d'être une chauve-souris semble nous conduire par conséquent, à la conclusion suivante : il y a des faits qui ne consistent pas en la vérité de propositions exprimables dans un langage humain. Nous pouvons être contraints de reconnaître l'existence de faits de ce genre sans être capables de les établir ou de les comprendre »

Dans cet article très célèbre, l'auteur nous montre que nous ne pouvons pas, n'étant nous même pas une IA nous demander si ces dernières ont une conscience. Nous n'avons pas le processus « cérébral » correspondant au fonctionnement d'une IA, il paraît de ce fait compliqué d'examiner extérieurement si oui ou non, ce que nous appelons IA possède une conscience [14].

4. Création de la notion de conscience artificielle

À cause de la difficulté à déterminer la présence d'une conscience dans l'IA, nous avons décidé d'analyser une conscience réduite : la conscience artificielle. Cette dernière peut être analysée selon deux points de vue : l'artificielle comme artifice ou comme tromperie et l'artificielle comme création humaine.

4.1. Vision de l'artificielle comme tromperie

Une manière d'approcher l'artificiel est de le considérer comme quelque chose de faux, de factice, qui a pour but de tromper. C'est la vision de l'artificiel comme tromperie, comme d'un artifice. Une cause du fait que nous puissions penser qu'il y ait une conscience, dans des objets par exemple, est le cerveau humain. En effet, il est programmé pour associer naturellement les formes qu'il perçoit à des formes qu'il connaît déjà, et auxquelles il se réfère. Par exemple, notre cerveau détecte tout ce qui a deux yeux et une bouche, comme un visage. Même s' il se résume simplement à deux points et une barre « :) », c'est le principe de la paréidolie. Il suffit également qu'un objet se mette en mouvement pour que notre cerveau interprète ce mouvement comme provenant d'un être vivant. Une étude réaliser en 2009 par des chercheuses de Stanford montre notre capacité à voir une âme et donc par extension une conscience dans le moindre objet en mouvement. Lors de cette étude, les chercheurs ont enclenché l'ouverture d'une porte automatique devant des passants, à des vitesses différentes. Ils ont ensuite demandé aux passants comment ils interprétaient ces « gestes » de portes. Ils sont arrivés à la conclusion que la façon dont s'ouvre la porte automatique influence l'interprétation émotionnelle de la porte par les passants [15]. Une des caractéristiques de la CA devrait donc être de pouvoir se jouer de nous, nous tromper, en nous donnant l'illusion d'une conscience par le biais de la prise de traits humains, ou alors la prise de trait pouvant être associé à un être vivant. L'anthropomorphisation des IA joue ainsi un rôle majeur dans l'évolution de la vision de l'IA non plus comme une intelligence, mais comme une conscience, une CA.

De plus, le fait que nous puissions qualifier une IA de CA, avec artificiel toujours compris comme un artifice, est qu'elle se situe dans un entre deux. En effet, Descartes dans *Lettre au marquis de Newcastle* explique que le langage est le seul signe certain de la présence d'une pensée et donc par conséquent d'une conscience. Les animaux pour Descartes ne font qu'émettre des sons articulés , cela est dû au fait qu'ils possèdent des organes utiles à la parole. Ils sont dans l'incapacité de produire un langage qui leur permettrait d'exprimer des pensées, leurs langages

se basent seulement sur l'expression de leurs passions (leur langage n'est que le mouvement de « leur crainte, de leur espérance, de leur joie »). Il constituerait une sorte de réflexe ne nécessitant aucune pensée particulière [16]. Dans le cas des IA, on ne peut pas affirmer qu'elles expriment leurs passions, les sentiments des IA ne sont que simulés, une IA ne peut donc en avoir. Ce qui voudrait dire qu'à défaut d'exprimer leurs sensations, elles expriment des idées ou des pensées au moyen d'un langage, elles seraient alors dotées d'une conscience au même titre que les hommes. Cependant, comme le démontre l'expérience de pensée de La chambre chinoise , une CA n'exprime pas forcément des idées. En effet, dans le principe de la chambre chinoise, un opérateur ne parlant pas chinois est enfermé dans une pièce. Il reçoit un morceau de papier sur lequel sont inscrits des caractères chinois, il peut consulter des catalogues de règles qui lui indiquent comment répondre correctement avec des phrases en caractères chinois. Il utilise ces catalogues de règles pour noter des caractères sur un bout de papier, et renvoie au monde extérieur une réponse appropriée. L'expérience de pensée de John Searle démontre que pour une personne située en dehors de la pièce, l'opérateur semble avoir une parfaite compréhension et maîtrise du chinois, alors qu'il se contente seulement de suivre les règles sémantique, de plus tout ce qui est écrit en chinois sur les morceaux de papier n'a aucun sens pour l'opérateur. L'opérateur peut être parfaitement assimilé à un modèle d'IA où l'algorithme est représenté par les livres qui permettent à l'opérateur de traiter et d'analyser les données ; les symboles chinois reçus par l'opérateur peuvent être comparés aux images et aux textes reçus par une IA ; l'opérateur, lui, jouerait plutôt le rôle d'un mécanisme de traitement manipulant des symboles sans réellement comprendre leurs sens. Donc, bien que l'IA ne soit pas dotée d'une conscience avérée, on peut cependant affirmer qu'elle nous donne l'illusion d'une conscience. C'est également à ce titre que l'on peut qualifier sa conscience de CA.

4.2. Vision de l'artificielle comme création humaine

nsuite, le mot artificiel peut être compris d'une manière toute autre que simplement en tant que synonyme d'imitation. Effectivement, cet adjectif descendant du latin *artificialis* (« fait avec art ») [17] est aussi défini comme ce qui est « Produit par le travail de l'homme et non par la nature » [18]. Dès lors, il est intéressant de se pencher sur la capacité de l'homme à développer une nouvelle forme de conscience qui différait de celle dans la nature. Une distinction à faire dès le début de cette réflexion sur ce domaine de recherche qu'est la conscience artificielle est l'alternative intelligence artificielle (IA)/conscience artificielle (CA). Ron Chrisley définit l'IA comme : [19]

> « la tentative de créer des artefacts qui ont des propriétés mentales, ou manifestent des caractéristiques de systèmes qui ont ces propriétés, incluant comme propriétés non seulement l'*intelligence*, mais aussi [...] la *perception*, l'*action*, l'*émotion*, la *créativité* et la *conscience* »

Ainsi, la CA est par définition une sous-catégorie de l'IA prenant en compte des critères autres que les capacités cognitives. Dans son article, bien qu'il souligne l'absence de distinction claire en l'IA et la CA, il fait tout de même référence à des critères donnant de multiples objectifs à la recherche en CA discutés par P. Carruthers : [20]

> « (1) expliquer comment de tels systèmes ont une dimension subjective [...], (2) pourquoi les propriétés mises en jeu dans la conscience phénoménale doivent sembler intrinsèques à leurs sujets [...] (3) pourquoi elles doivent leur sembler ineffables ou indescriptibles »

Dès lors, une CA devrait avoir (1) une subjectivité et (2) des qualités intrinsèques au sujet de la conscience, qui sont toutes deux (3) ineffables et indescriptibles.

En premier lieu Ron Chrisley introduit une première alternative entre la CA « d'ingénierie » et celle « scientifique » : alors que la CA ingénieure aspire à créer un artefact pouvant « faire des choses dont jusqu'alors seuls des agents conscients étaient capables », la CA scientifique, elle vise à « comprendre les processus sous-jacents à la conscience et les avancées de la CA ingénieure, nonobstant leur grandeur, n'ont d'intérêt théorique qu'au regard de leur capacité à mettre en lumière les processus de conscience ». Appliqué à notre sujet ceci permet par exemple d'expliquer la distinction entre les algorithmes de recommandation et d'autres formes d'IA : les algorithmes de recommandation ont une visée purement ingénierique et ce serait cela qui expliquerait la distinction entre ces algorithmes et ceux qu'on nomme IA et au sujet desquels il est question de conscience. À l'inverse, la CA scientifique vise à déterminer les paramètres de la conscience à travers les progrès de la CA ingénieure (qui peut être vue comme une forme d'IA forte) et des réflexions plus abstraites, dont de la philosophie des sciences. Cette catégorie est ensuite subdivisée selon un modèle classique de distinction entre la CA *strong* (forte) et *weak* (faible). La *weak* CA correspond à toute approche qui ne revendique pas de lien entre la technologie et la conscience. On peut alors objecter que si aucun lien nécessaire ne peut être établi entre la machine et la conscience, il est déraisonnable de parler de CA : c'est effectivement ce qu'affirment Manzotti et Tagliasco : [21]

> « Pour faire court, la *weak* CA semble bien n'être qu'une intelligence artificielle et ne nous intéresse pas. À l'inverse, la CA *strong* est l'approche dont le but ultime est la création de systèmes qui, lorsqu'ils sont implémentés, sont suffisants à la conscience »

Pour reformuler, cela implique que la *strong* CA vise à recréer la conscience dans son ensemble et à comprendre entièrement son fonctionnement : cela peut par exemple être illustré par l'idée qu'il serait possible de décrire parfaitement le fonctionnement de la conscience et de le programmer ensuite. Or cela paraît très ambitieux et soulève beaucoup de problèmes, le principal étant qu'il est difficile de reproduire quelque chose sans en étudier une version simplifiée. C'est d'ailleurs souvent la méthode utilisée dans les sciences expérimentales : plutôt que d'étudier un phénomène physique dans son ensemble, il est commun de développer un paradigme simplificateur lors d'une expérience. Somme toute, il est sûrement plus raisonnable d'user d'une version simplifiée de la conscience, ne reprenant peut-être pas l'ensemble de ce

concept complexe, mais une partie, afin de pouvoir jamais envisager une définition juste de la CA.

Enfin ici peut être introduite une version nuancée de ces deux concepts correspondant à la recherche d'une conscience artificielle qui ne vise pas à être elle-même consciente, mais qui partage et met en avant les caractéristiques de la CA et se marie avec les critères de Carruthers, reprenant ce dont nous venons de discuter. Effectivement, via le développement d'une CA qui vise à mettre en avant seuls des aspects de la conscience, il est possible de souligner à la fois les critères (1) et (2) tout en veillant à se garder d'argumenter que celles-ci possèdent vraiment la CA puisque ce n'en est pas une version parfaite ni (3) ineffable et qu'elle reste (3) indescriptible. On parle alors de *lagom* CA, du suédois et considéré comme un concept intraduisible qui se rapproche d'optimal en français.

Admettons que la conscience artificielle soit l'ensemble des modèles utilisés par l'homme pour établir les caractéristiques de la conscience ; nous pouvons ainsi nous intéresser aux critères de présence de la conscience artificielle (comme modèle simplifié) tout en la contrastant avec la conscience humaine.

Vient alors une interrogation naturelle vis-à-vis des similarités et différences entre la conscience artificielle et la conscience humaine. Nous pouvons d'abord reprendre la définition de l'IA en tant que mécanisme de traitement de données (algorithme sophistiqué) : à l'inverse d'un humain qui fait le choix de son objectif, et est en ce sens autonome, l'IA a en règle générale besoin d'un prompt afin de se mettre en fonctionnement (cf 2.2.1). Cette distinction entre ce que Chrisley appelle la CA autonome et la CA prosthétique prend à la fois en compte les capacités cognitives (si cela peut être attribué à l'IA) et peut être mise en parallèle avec le concept de libre-arbitre et d'auto-détermination : en mentionnant les CA prosthétiques, nous entendons tous les artefacts de *lagom* CA qui élargissent les capacités liées à la conscience d'un organisme considéré comme étant déjà conscient. Pour illustrer cette supposée « conscience artificielle prosthétique », nous pouvons faire référence à Superbrain 1 développé par 7Sense permettant aux non-voyants d'accéder à une forme alternative de vision et de profondeur, tout cela à l'aide d'une IA [22]. Dès lors, l'usage de telles technologies peut être décrit de la manière suivante : [21]

> « Ce qui est spécial avec la conscience artificielle est qu'elle crée de nouvelles expériences, non en altérant les objets d'expérience, mais en altérant les bases sur lesquelles s'appuient les processus de conscience »

Bien que l'IA entre dans cette catégorie, en tant qu'elle permet d'ouvrir des possibilités de langage, de traitement de données et de génération d'images qui augmentent la capacité de l'homme à faire usage de sa conscience, il paraît excessif d'accorder le terme de conscience à de tels agents. Prenons par exemple une paire de lunettes : elle permet bien d'améliorer les perceptions de l'homme et ainsi sa conscience du monde extérieur, mais qui considérerait les lunettes comme une forme de conscience artificielle ? Par déduction, nous considérons qu'il

n'est question de CA que lorsque le système est indépendant.

Dès lors, ils sont perçus comme des sujets artificiellement conscients et non en tant qu'objets : ils sont alors des sujets. Riccardo Manzotti et Vincenzo Tagliasco élaborent d'ailleurs dans leur réflexion sur la CA que c'est cette nature de sujet, donc de lien avec l'extérieur, qui définit la conscience : « la conscience est identique à l'occurrence d'une continuité entre le cerveau et la partie du monde qui est perçue ». Bien qu'ici, ils fassent usage du terme cerveau, qui semble faire allusion à un monopole de la conscience par les organismes vivants, il faut ici comprendre « cerveau » au sens « centre de fonctionnement» . Effectivement, en utilisant la continuité entre perception, action et extériorité, il est possible d'éliminer le problème de la qualia, aporie axiomatique dans la philosophie de la conscience. Le philosophe Daniel Dennett élabore cette idée en défendant une hétérophénoménologie de la conscience, dite éliminativiste, argumentant qu'il est déraisonnable de chercher à définir les *qualia* du fait de son caractère « ineffable, intrinsèque, privé et directement ou immédiatement appréhensible dans la conscience » [23]. Ainsi la conscience est réduite au lien qui unit représentation de l'extériorité à l'intériorité. Il écoule encore quelques nuances de la conscience de cette définition telles que ceux d'Endel Tulving évoque : [24]

1. une conscience autonoétique, relative à la correspondance de soi-même à l'image que le sujet a de lui-même. En référence à la section 2.2.2, il est possible d'établir que l'IA a un sens de soi assez développé, mais fragile : l'algorithme sait ce qu'il est, mais se laisse convaincre qu'il est autre chose ;
2. une conscience noétique, soit du monde extérieur. En ce qui concerne cette forme, il suffit de se référer à toute technologie de transport, de tri, ou de reconnaissance faisant usage d'intelligence artificielle pour reconnaître l'évidente présence de cette forme de CA dans l'IA ;
3. une conscience anoétique, qui donc n'agit pas, mais rend seulement compte de la présence d'activité (de vie, mais ici ce terme serait idiot). Or est-il possible pour nous d'évaluer le fonctionnement d'une IA sans justement la mettre en fonctionnement ? À l'inverse d'un humain, il n'est pas possible de scanner les ondes parcourant le cerveau pour déclarer la vie d'une IA. Ce problème peut aussi être pensé comme la difficulté du *third-person-access* à la conscience [25]. Certains chercheurs soutiennent que c'est justement un usage de *lagom* CA qui pourrait permettre de « dépasser » cette limitation épistémique biologique/humaine, par exemple à travers la création d'un esprit-nuage qui autorise un accès direct aux états mentaux subjectifs d'autres IA [26]. Ainsi, bien que vérifier qu'une IA est bien en « vie » soit antithétique, il serait tout de même possible à l'aide de l'IA d'obtenir une représentation des « états mentaux » d'une IA au-delà de la méthode d'observation phénoménologique.

Finalement, une fois considéré tel un sujet plus ou moins conscient de son niveau noétique,

le niveau de la CA peut aussi être jugée selon des critères téléologique, soit relatifs à la finalité d'une IA et la mesure dans laquelle celle-ci est autodéterminée. Manzotti et Tagliasco élaborent une hiérarchie des architectures selon leur plasticité téléologique :

1. à contrôle fixé : un but et une façon d'atteindre celui-ci sont fixés ;
2. apprenant : le système est capable de modifier son comportement afin d'atteindre un but fixé ;
3. génératives de but : le système est capable de modifier à la fois son but et sa méthode, il semble alors doté de libre arbitre (cf 2.2.1).

On considérera alors qu'une CA développée mettrait en avant cette capacité des organismes conscients à s'autodéterminer (3) : un exemple frappant de ceci est le robot Starfish qui, à l'aide de capteurs a élaboré une image de lui-même, soit une conscience autonoétique, ainsi que son propre but (avancer) et la méthode adaptée, c'est-à-dire une architecture auto générative de but [27]. L'IA paraît alors présenter des caractéristiques de la CA.

5. Conclusion

L'IA est une définition technique manquant de précisions, notamment à cause de l'utilisation du concept d'« intelligence » et d'« artificielle ». Malgré cette difficulté, il est possible de définir l'IA comme étant tous les algorithmes utilisant du *machine learning*. En effet, cette technique permet de construire des algorithmes apprenant en autonomie, ce qui ressemble à une forme d'intelligence.

Les récentes avancées technologiques (mise à disposition de ChatGPT, de Dall·E ou encore de Midjourney aurpès du grand public) nous questionnent autour de notre relation à la machine et plus particulièrement à la présence de conscience : un *chatbot* (comme Gemini) comprend le sens de nos phrases et nous répond très souvent d'une manière pertinente, ce que l'on associe naturellement à la présence de conscience. Pourtant, elle ne nous garantit rien : il est possible que l'IA ne soit qu'un perroquet nous trompant. Ainsi, il est essentiel de pouvoir définir ce qu'est la conscience avant de répondre à cette question.

Dans un premier temps, la conscience peut se définir comme un processus de perception du monde permettant une interaction pertinente. Avec cette vision, un appareil photo serait conscient, ce qui nous semble très étrange. Un deuxième aspect de la conscience est sa valeur existentielle : ce serait la capacité à distinguer l'objet perçu et le sujet percevant. Trancher l'existence d'une valeur existentielle présente dans l'IA est très complexe : nous ne sommes pas nous-même des IA. Ainsi, pour analyser les véritables capacités de l'IA, nous avons besoin d'une forme réduite de conscience : la conscience artificielle.

Ce concept peut être interprété de deux manières : la CA comme artifice et celle comme création humaine.

La CA comme artifice fonctionne très bien pour l'IA : l'expérience de pensée de la chambre chinoise montre très bien les limites de la conscience de l'IA. En effet, il ne s'agirait que d'une machine suivant des règles précises sans comprendre le sens intrinsèque des mots qu'elle emploie ou qu'elle reçoit.

La CA comme création humaine est une question beaucoup plus complexe, même si elle fonctionne tout aussi bien par définition de l'IA. Ce type de CA montre que, malgré l'artifice et l'absence d'assurance de la conscience, l'IA possède quand même de nombreuses caractéristiques liés à la présence de conscience : l'autodétermination et l'autonomie sont les plus importantes.

Ainsi, l'utilisation du concept de conscience artificielle pour décrire les IA est totalement adaptée puisque que ce dernier montre bien l'état entre une conscience totale, comme celle humaine, et son absence.

Références

[1] A. M. Turing. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236) :433–460, October 1950.

[2] Apprentissage automatique. *Wikipédia*, February 2025.

[3] Lê Nguyên Hoang. Un million de milliards de dilemmes, March 2021.

[4] Lê Nguyên Hoang. Les données manipulent les algorithmes, April 2021.

[5] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models, December 2024.

[6] Myanmar : The social atrocity : Meta and the right to remedy for the Rohingya. https ://www.amnesty.org/en/documents/asa16/5933/2022/en/, September 2022.

[7] Georgia Wells, Jeff Horwitz, and Deepa Seetharaman. The Facebook Files. *Wall Street Journal*, October 2021.

[8] Wiktionnaire. conscience, April 2025.

[9] Bertrand Russel. *Problèmes de philosophiee*. Bibliothèque philosophique. France, payot edition, February 1989.

[10] Karl Marx. *Manuscrits de 1844*. J. Vrin, January 2021.

[11] Étienne Bibenet. *L'invention Du Réalisme*. Passages. Cerf, Paris, 2015.

[12] Martin Heidegger. *Être et Temps*. Bibliothèque de philosophie. Gallimard edition, November 1986.

[13] Georg Wilhelm Friedrich Hegel. *Propédeutique philosophique. : Traduit et présenté par Maurice de Gandillac.* Éditions de minuit, [Paris], 1963.

[14] What is it like to be a bat ? In Thomas Nagel, editor, *Mortal Questions*, Canto Classics, pages 165–180. Cambridge University Press, Cambridge, 2012.

[15] Wendy Ju and Leila Takayama. Approachability : How People Interpret Automatic Door Movement as Gesture. *International Journal of Design*, 3(2), 2009.

[16] René Descartes. Œuvres et lettres. https ://www.gallimard.fr/catalogue/oeuvres-et-lettres/9782070101665, May 1937.

[17] Wiktionnaire. artificiel, February 2025.

[18] Larousse. artificiel.

[19] Ron Chrisley. Philosophical foundations of artificial consciousness. *Artificial Intelligence in Medicine*, 44(2) :119–137, October 2008.

[20] Peter Carruthers. *Phenomenal Consciousness : A Naturalistic Theory*. Cambridge University Press, Cambridge, 2000.

[21] Riccardo Manzotti and Vincenzo Tagliasco. Artificial consciousness : A discipline between technological and theoretical obstacles. *Artificial Intelligence in Medicine*, 44(2) :105–117, October 2008.

[22] SuperBrain 1 - 7Sense. https ://7sense.ee/superbrain-1/.

[23] Daniel C. Dennett. *Consciousness Explained*. June 1993.

[24] Endel Tulving. Episodic Memory and Autonoesis : Uniquely Human ? In *The Missing Link in Cognition : Origins of Self-Reflective Consciousness*, pages 3–56. Oxford University Press, New York, NY, US, 2005.

[25] Kathinka Evers, Michele Farisco, Raja Chatila, Brian Earp, Ismael Freire, Fred Hamker, Erik Németh, Paul F. M. J. Verschure, and Mehdi Khamassi. Artificial consciousness. Some logical and conceptual preliminaries. August 2024.

[26] David M. Lyreskog, Hazem Zohny, Julian Savulescu, and Ilina Singh. Merging Minds : The Conceptual and Ethical Impacts of Emerging Technologies for Collective Minds. *Neuroethics*, 16(1) :12, March 2023.

[27] Josh Bongard, Victor Zykov, and Hod Lipson. Resilient Machines Through Continuous Self-Modeling. *Science*, 314(5802) :1118–1121, November 2006.