IA

Conscience et libre arbitre, le dernier pré carré de l'humain face à l'IA .

Découvrez une réflexion portée par Matthieu CONSTANS, consultant chez Softeam

SOFTEAM
UNE FILIALE DE DOCAPOSTE

Depuis environ 2 ans, l'intelligence artificielle générative a connu un succès fulgurant. Jamais une machine n'avait atteint un tel niveau de maîtrise du langage et de raisonnement, au point de rivaliser avec les capacités humaines.

Les échanges conversationnels avec les nouveaux chatbots intelligents sont véritablement bluffants, et l'actualité ne cesse d'impressionner par les nouvelles capacités émergentes dans à peu près tous les domaines : vision et reconnaissance d'image, reconnaissance de sons, manipulation du langage et traduction, rédaction et manipulation de textes, d'images et de videos, commande de robots, diagnostic médical, conseil juridique, création artistique, développement logiciel, rien ne semble désormais inaccessible aux grands modèles d'IA.

On s'adresse désormais à un chatbot comme on s'adresserait à un humain, et le test de turing censé permettre de faire la part des choses entre une IA et un être humain semble aujourd'hui complètement dépassé. La crainte est grande de voir de nombreuses professions totalement remises en cause avec l'arrivée de l'IA.

Certes, il y a déjà eu la révolution industrielle et la mécanisation du travail, et la révolution informatique et numérique avant l'IA, qui en leur temps avaient déjà complètement révolutionné notre manière de produire et de travailler, repositionnant l'humain toujours plus loin et plus haut dans la chaîne de valeur. La mécanisation a permis la production de masse et l'augmentation des richesses, en libérant l'humain de tâches physiques pénibles. La révolution informatique a permis de pousser plus loin l'automatisation, l'intégration des processus, laissant à l'humain le rôle dévolu à ses capacités intellectuelles supérieures.

On a parlé d'une société de services, de conseil, de matière grise. Mais là, avec l'arrivée de l'IA, c'est le dernier pré-carré de l'humain, à savoir ses capacités cognitives, qui semble pouvoir être pris en charge par la machine.

Quelles sont aujourd'hui les limites de cette IA ?
En quoi l'humain se distingue-t-il encore ?
Existe-t-il des capacités humaines qui resteront inaccessibles à l'IA et si oui lesquelles ? Telles sont les questions que je propose d'aborder dans cet article.

LE PRINCIPE DE FONCTIONNEMENT

Un premier argument que j'entends parfois consiste à dire que l'IA n'a rien à voir avec l'intelligence humaine, car ses modèles ne font que prédire, de manière statistique, la sortie la plus probable en fonction de l'entrée qui leur est donnée. Pour ce qui concerne un texte par exemple, un grand modèle de langage ne fait que prédire le mot le plus probable qui doit suivre le prompt d'entrée, moyennant tous les poids du modèle et des données d'apprentissage.
J'avoue que cet argument ne me convainc pas du tout.

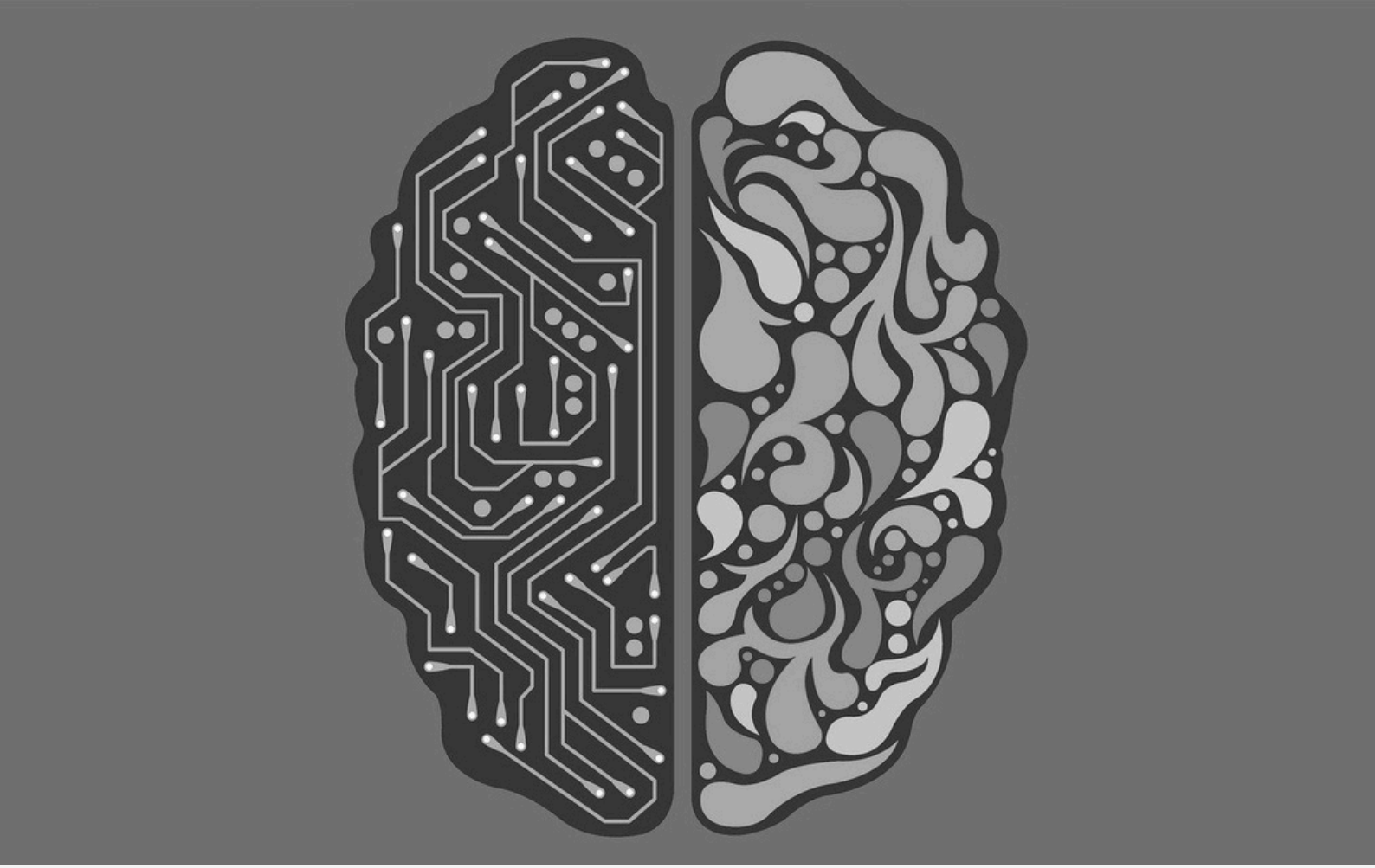
En effet, si on considère à la fois le fonctionnement d'un modèle d'IA et le fonctionnement du cerveau humain, hors apprentissage, il est impossible de ne pas y voir certaines similitudes : on a en gros au milieu un réseau de neurones entraîné, en amont des entrées (stimulis pour le cerveau humain, prompts ou images ou tout autre type d'entrée pour un modèle d'IA), et une sortie calculée en fonction des poids du réseau de neurones.

Certes, certains diront que c'est réducteur.

Néanmoins, si on y pense, qu'est-ce que notre cerveau fait dans la réalité, sinon de déterminer, en fonction de ce qui lui arrive en entrée, les sorties qui entraîneront le comportement le plus adapté, on pourrait dire le comportement suivant le plus probable en vertu de critères de mesure tels que l'adaptation au contexte, la survie, etc ?

Je ne vois pas de différence disruptive, fondamentale, sur le principe, entre la manière de fonctionner d'un réseau de neurone biologique, qu'il soit humain ou non, et un réseau de neurone artificiels, hors apprentissage encore une fois, lorsqu'il s'agit de produire une sortie en fonction d'entrées.

Exit donc, à mon sens, l'argument qui dit qu'un modèle d'IA ne fait que prédire la sortie (mot, image, etc.) la plus probable, comme étant un argument fondamental le différenciant de l'intelligence humaine.



LA RICHESSE DES MODÈLES

Une différence bien entendu est celle de la complexité et de la richesse des réseaux de neurones, incroyablement plus complexe chez l'humain que dans n'importe quel modèle LLM aujourd'hui.
Mais en sera-t-il toujours ainsi ? On estime le nombre de poids dans le modèle de GPT4 à environ 1 000 milliards.

À titre de comparaison, le cerveau humain dispose d'environ 100 000 Milliards de connexions synaptiques, que l'on peut apparenter à des poids de liaisons entre neurones, soit environ 100 fois plus que dans un LLM. 100 fois plus cela semble beaucoup, mais est-ce si énorme que cela ? En rendant les modèles un peu moins gourmands pour l'apprentissage, ne sera-t-il pas possible à terme d'atteindre les 100 000 Milliards de poids dans un grand modèle d'IA et même de dépasser ce nombre ?

Par ailleurs, si on souhaite comparer les capacités du cerveau aux LLMs, ne faut-il pas n'y considérer que les parties qui traitent du langage, sachant qu' en réalité le cerveau couvre de très nombreuses fonctions qui vont bien au-delà du langage ? Car si on ne considère dans le cerveau que la partie qui traite le langage, cela concerne :

- L'aire de Broca pour la syntaxe
- L'aire de Wernicke pour la compréhension du langage
- Le Gyrus angulaire et supramarginal pour la lecture et l'intégration multisensorielle
- Le cortex auditif primaire pour la compréhension des sons, le cortex préfrontal pour la mémoire de travail et la planification du discours
- Une partie de l'hémisphère droit pour le contexte global

Donc en limitant le décompte à ces zones du cerveau, on réduit le nombre de connexions mises à contribution à 16 000 milliards de connexions synaptiques environ sur le total évoqué précédemment de 100 000 milliards, soit finalement 16 fois plus à peine que dans les derniers modèles LLM.

L'APPRENTISSAGE

Un autre argument, plus valable à mon sens, est celui des modalités et de l'efficience de l'apprentissage au quotidien. Je ne parle pas de capacité d'apprentissage global, car à moins qu'une personne arrive à lire et à assimiler l'intégralité des 7 millions de pages de Wikipedia, lus environ 50 Tera-Octets de pages Web, incluant des centaines de millions d'articles, plus ou moins 71 000 livres tombés dans le domaine public, ce dont je doute fort, il semble que sur le plan de la culture générale, les LLM nous battent tous définitivement à plate couture.

Certes, le cerveau humain semble beaucoup plus efficient et moins gourmand en énergie, mais une fois l'apprentissage fait un LLM est rapidement duplicable en masse sans nécessité de nouvel entraînement, tandis que pour le cerveau humain, chacun doit personnellement faire son apprentissage, et il n'existe pas de moyen de télécharger automatiquement la connaissance d'un cerveau à un autre.

Dans un modèle d'IA, contrairement au cerveau humain qui apprend et se modifie en permanence, on sépare complètement la phase d'apprentissage de celle d'utilisation (aussi appelée phase d'inférence). Un modèle tel que ChatGPT, Claude, Mistral ou encore Deepseek n'apprend pas en temps réel de nos conversations et interactions avec lui. Tout au plus la société qui maintient le modèle peut-elle conserver le contenu de certains de nos échanges et, après les avoir évalués et éventuellement filtrés, elle peut en utiliser en partie pour la phase d'apprentissage suivante et du fine tuning.

Ce mode de fonctionnement est explicitement choisi car, si les LLM devaient apprendre en temps réel de nos interactions, cela ouvrirait la porte à de nombreuses attaques visant à y introduire des biais.
N'importe qui pourrait alimenter le modèle avec n'importe quelle donnée et en influencer les poids, ce qui donnerait lieu à toutes les dérives possibles.

Les grands modèles d'IA sont un peu comme un super cerveau dont les poids sont figés après chaque phase d'apprentissage, tandis que côté humains, nous sommes capables d'apprendre et de modifier notre cerveau beaucoup plus souvent et rapidement.

Non seulement les connexions synaptiques dans notre cerveau se modifient tous les jours, mais encore, contrairement à une idée reçue qui a encore aujourd'hui la vie dure, de nouveaux neurones sont créés dans notre cerveau en permanence, comme cela a été observé en particulier dans la circonvolution de l'hippocampe, siège de la mémoire autobiographique.
Cela nous donne un avantage, notamment en termes de créativité, car quel que soit l'apprentissage d'un modèle d'IA, il ne pourra que restituer quelque chose à l'image de ce qu'il a appris, et sera en quelque sorte incapable de créer son propre style et de le faire vivre, car cela imposerait une modification continue de son modèle interne.

À ce jour, nous ne savons pas modifier en temps réel les modèles d'IA pendant leur exploitation, et les phases d'entrainement demandent beaucoup de ressources énergétiques et beaucoup de temps. A regarder de près le fonctionnement du cerveau humain, il apparait en réalité qu'en son sein le véritable apprentissage a lieu pendant le sommeil, notamment pendant la phase de sommeil paradoxal. Dans la journée, le cerveau garde des traces mnésiques dans une mémoire à relativement court terme, et c'est pendant le sommeil que les expériences donnent lieu à une réorganisation du cerveau et des connexions synaptiques. C'est aussi la raison pour laquelle certains artistes utilisent parfois des substances psychoactives (des drogues en quelque sorte) pour favoriser la création, car ces substances favorisent une sorte de reconfiguration cognitive.

Sur le plan de la créativité donc, l'humain surpasse l'IA, car il apprend en permanence, avec une phase active de cet apprentissage a minima tous les 24h, la journée étant plutôt l'objet d'une inférence du modèle cérébral et d'un stockage des expériences sur le court terme.

Mais ne nous y trompons pas : non seulement le fait ne pas permettre une modification des modèles en phase d'exploitation est un choix voulu pour éviter que les modèles soient attaqués via l'introduction malicieuse et massive d'informations pouvant introduire des biais, mais encore, il n'est pas inenvisageable qu'à relativement court ou moyen terme, les modèles d'IA fassent l'objet d'un entrainement complémentaire quotidien, leur faisant intégrer à minima l'actualité du jour.

En outre, même si contrairement au cerveau humain, aujourd'hui, les modèles d'IA sont incapables de neurogénèse pendant la phase d'apprentissage, un modèle et le nombre de neurones d'un modèle d'IA étant fixé et figé dès le départ, des recherches actives sont aujourd'hui réalisées, afin que des modèles deviennent capables, en phase d'apprentissage, d'ajouter automatiquement de nouveaux neurones et de complexifier l'architecture quand les performances de l'architecture utilisée plafonnent.

Sur le plan de l'apprentissage, on peut donc conclure qu'aujourd'hui, l'être humain dépasse l'IA par ses capacités de neurogénèse et de reconfiguration neuronale d'une part, et d'autre part par son efficience énergétique, tandis que l'IA dépasse l'humain par la quantité de connaissances assimilées, du moins en ce qui concerne les connaissances académiques et culturelles. Cependant, des recherches actives sont menées pour améliorer les modalités d'apprentissage et son efficience.

L'AUTONOMIE

Un nouvel argument en faveur de l'humain, consiste à dire que ce dernier est autonome et à ce titre capable d'initiative.
Nul besoin de le solliciter pour qu'il agisse. Nous avons tous coutume de faire des projets, de planifier, de programmer nos actions puis de les exécuter.

Une IA générative aujourd'hui est réactive, c'est-à-dire qu'elle répond à une requête. Elle n'a pas de vie propre faite d'initiatives spontanées qui la verrait agir en toute autonomie en dehors de toute interaction humaine. Elle n'a pas de but intrinsèque.

Cependant, l'IA agentique est précisément un domaine dont le principe est de concevoir des IA autonomes. Il est aujourd'hui possible de construire des modèles capables de planifier un raisonnement sur le court terme, et de concevoir des voitures qui se conduisent de manière quasi-autonome. Partant de là, rien n'interdit de faire en sorte qu'une IA interagisse avec son environnement et, considérant chaque entrée de capteurs tels que caméras ou micros comme une nouvelle requête, planifie sa réflexion de manière à trouver la meilleure réponse en fonction de son modèle, non seulement en termes de paroles, mais aussi en termes de mouvement.

LA CONSCIENCE

Un des freins parfois avancés à l'acquisition d'une telle autonomie vient du fait qu'une IA n'a pas, du moins c'est ce que nous pensons, de conscience d'exister, ni de ressenti, qui viendrait la guider au quotidien et lui donnerait spontanément des buts. J'en viens alors à une des différences fondamentales qui vient à l'esprit quand on veut comparer une IA à un être humain, à savoir le fait qu'un être humain est doué d'une conscience d'exister, et ressent subjectivement les choses, tandis qu'une IA n'a aucune conscience d'exister et ne dispose d'absolument aucun ressenti subjectif.

Cet argument est d'ailleurs parfois présenté comme un tabou : il faudrait être totalement fou pour penser qu'une machine faite de silicium, avec une architecture totalement différente de celle du cerveau humain, dispose d'une conscience d'exister. Car il se trouve qu'aujourd'hui, malgré les avancées considérables des neurosciences et de la biologie, personne ne sait ce qui fait qu'à un moment donné l'expérience subjective émerge. Nous avons bien découvert ce que l'on appelle des corrélats de la conscience, c'est-à-dire des phénomènes biologiques et physiques concomitants à l'émergence de la conscience chez une personne. Cependant, nous ignorons tout de ce qui en définitive est à l'origine de l'émergence de la conscience.

Il est vrai que les échanges réalisés avec un chatbot modernes sont aujourd'hui parfois si bluffants qu'il est difficile de faire la différence avec la réponse qui aurait pu être rédigée par un être humain, et à ce titre, comme je le disais précédemment, il semble que le test de Turing, qui s'appuie précisément sur un tel discernement, apparaisse comme dépassé.

Il existe bien un courant de pensée qualifié de « fonctionnaliste » qui considère que, à partir du moment où un mécanisme produit, à partir d'entrées, les mêmes sorties que celle que produirait un être humain pour ces mêmes entrées, alors cela signifie que cet organisme est nécessairement conscient, comme l'est l'être humain. La conscience serait à ce titre une propriété émergente de la fonction produite par un organisme, cette fonction étant caractérisée par ses entrées et ses sorties et les relations entre ces dernières, et ce, quels que soient le substrat et la matière constitutifs de cet organisme.

Le problème de ce courant de pensée, c'est qu'il s'appuie sur un postulat, qui est précisément le principe que je viens d'évoquer, et que ce postulat n'a jamais été démontré. Par conséquent, le fonctionnalisme n'est à mon sens qu'une hypothèse, rien de plus, et pas du tout, en l'état des connaissances, une vérité démontrée.

À y regarder de près, il apparaît que, quel que sera le degré d'avancement de la technologie, le fait de déterminer qu'une IA ou même que n'importe quel organisme dispose d'une conscience de sa propre existence ou pas soulève un défi totalement insurmontable. En effet, si chacun d'entre nous a, de par son expérience subjective directe en première personne, la preuve de l'existence que de sa propre conscience, les choses se corsent lorsqu'il s'agit, pour une personne, de prouver l'existence d'une conscience pour une personne autre que lui-même.

L'expérience de la conscience de soi est purement subjective, personnelle et non partageable. Je ne considère que les personnes qui vivent autour de moi sont conscientes comme je le suis moi-même qu'en inférant l'existence de leur conscience par analogie à celle de ma propre conscience du fait que, puisque je suis conscient et que ces personnes sont physiquement et biologiquement semblables à moi, et que leur comportement vu de l'extérieur est similaire au mien, alors cela signifie que ces personnes sont elles aussi très certainement conscientes. Je n'ai pas de preuve directe de la conscience d'autrui, car je n'expérimente pas directement cette conscience d'autrui comme je le fais pour ma propre conscience.

Ce constat prend tout son sens lorsque nous cherchons à déterminer si un organisme qui n'est pas humain est conscient ou non. À mesure que les caractéristiques physiques, biologiques, comportementales d'un animal s'éloignent de nous, il nous est de plus en plus difficile de dire dans quelle mesure cet animal est conscient ou non. Beaucoup diront qu'un grand singe capable d'interagir avec nous avec la langue des signes est sans doute conscient. Certains diront qu'un chien domestique dispose d'une forme de conscience, d'autres le nieront. Qu'en est-il alors pour une souris ? Un oiseau ? Un lézard ? Une abeille ? Une bactérie ?

En ce qui concerne une IA, l'inférence ne marche pas sur le plan biologique, physique et physiologique, car les mécanismes à l'œuvre pour une IA sont radicalement différents de ce qui se passe, au sens physique, dans un cerveau, ce qui nous fait donc dire qu'une IA, faite de puces de silicium refroidies dans des Datacenters, n'est selon toute probabilité pas consciente de sa propre existence ; ceci sans compter le fait que, comme évoqué précédemment, puisque nous n'avons pas la moindre petite idée de ce qui fait que la conscience émerge, il est totalement improbable que nous fassions émerger celle-ci par hasard.

Cependant, si écartons un moment les seuls aspects physique et biologique, et que nous ne prenons en considération que le volet purement comportemental de l'IA, si nous sommes un jour capables de créer un robot capable de mimer à la perfection le comportement humain, capable de planifier, de décider de manière autonome, d'interagir avec nous, comment sera-t-il alors possible de déterminer si ce robot est conscient ou s'il demeure au contraire totalement inconscient tel un parfait Zombie au sens de David Chalmers (*) ?

Puisqu'il nous est définitivement impossible d'expérimenter directement en première personne comme pour nous – mêmes l'éventuelle conscience dont pourrait disposer un tel robot, et puisque nous ne pourrions qu'inférer l'existence d'une telle conscience de soi en nous fondant sur des signes extérieurs, il m'apparaît comme définitivement impossible de trancher de manière irréfutable en faveur de l'une ou l'autre de ces deux hypothèses.

Et, combien même nous serions capables de créer un robot conscient, il nous sera à jamais impossible de prouver qu'il l'est véritablement.

**David CHALMERS est le spécialisme mondialement reconnu en matière d'étude de la conscience*

En définitive, je dirais pour ma part que, comme nous ne savons pas ce qui fait que la conscience émerge, il est peu probable que nous ayons réussi à faire émerger une telle conscience par hasard dans les IA d'aujourd'hui.

Le fait qu'une IA, capable de raisonnements cognitifs élaborés, puisse exister sans qu'une conscience l'accompagne, démontre au passage qu'intelligence cognitive et conscience de soi sont bien deux caractéristiques totalement distinctes.

Je pense aussi que la conscience de soi est quelque chose de totalement transcendant à la matière, et que l'émergence de la conscience ne peut pas et ne pourra jamais être expliquée par un mécanisme physico-chimique.

LE LIBRE ARBITRE

Le dernier argument que je vois, qui viendrait différencier une IA d'un être humain, réside dans ce que l'on appelle communément le libre arbitre. Le libre arbitre, ou la liberté de penser et d'agir, suppose à mon sens l'existence simultanée des caractéristiques suivantes :

- La conscience de soi d'une part, car le libre arbitre suppose l'existence d'un sujet conscient capable de l'exercer, capable notamment de ressentir, de décider et d'agir en fonction de ses convictions propres.
- Un accord et une cohérence effectifs entre le désir ressenti par le sujet, la pensée et les actions de ce dernier, cette mise en cohérence étant généralement le résultat d'un processus de réflexion interne. C'est cette caractéristique de la liberté qui fait parfois dire aux enfants que la liberté consiste à pouvoir faire ce que l'on veut.
- Un accord de ces mêmes désirs, pensées et actions non seulement entre elles, mais aussi avec les valeurs universelles qui font que, sur le moyen et long terme, ces actions seront moins susceptibles de rencontrer l'opposition d'autres personnes. Cette caractéristique de la liberté, qui s'inscrit dans la continuation de la précédente dans le développement personnel du sujet et son accession à la maturité, illustre notamment le propos de Saint Augustin lorsqu'il déclare « Aime et fais ce que voudras ». En ce sens, le libre arbitre n'est pas simplement le fait de pouvoir faire ce que l'on veut, mais il inclut également une certaine forme de clairvoyance qui tournent spontanément le désir et la volonté du sujet vers des valeurs universelles de respect, d'amour, de compassion, etc. En d'autres termes, en vertu de ce troisième critère, un individu est d'autant plus libre que spontanément, sa volonté, son désir, son bonheur sont tournés vers et lui viennent naturellement du bonheur des autres.
- Une capacité à penser, décider et agir, dans une certaine mesure et dans un certain éventail de possibles, indépendamment de toute cause, y compris biologique et physique, et d'être, par conséquent, à tout instant, dans le cadre de certaines limites, la cause première du reste de de sa vie. Cette possibilité d'agir spontanément indépendamment de toute cause permet notamment à l'individu de choisir parmi plusieurs possibilités également positives et lui évite le sort malheureux de l'âne de Buridan, mort de faim et de soif entre un seau d'eau et un seau d'avoine. L'existence de cette composante essentielle du libre arbitre implique au passage que l'histoire de l'Univers n'est pas totalement écrite, puisqu'il y existe une possibilité d'action indépendante de toute cause, donc du passé.

En d'autres termes, un individu est libre si et seulement s'il est conscient de sa propre existence, si ses désirs, pensées et actions sont cohérentes entre elles, si elles sont alignées avec les valeurs universelles, et enfin si, après avoir délibéré intérieurement, éliminé parmi tous les choix possibles ceux qui ne sont pas en accord avec ces valeurs, ce même individu est capable de choisir parmi les actions possibles restantes sans que ce choix soit déterminé par aucune loi, et notamment par aucune loi de la physique.

Autrement dit, est libre celui dont le désir, les pensées et les actions sont spontanément orientés vers le bonheur universel de tous, et qui dispose, en dernier ressort, d'une capacité de choisir, entre plusieurs possibilités également positives, indépendamment de toute cause.

Pour qu'une IA soit dotée d'un libre arbitre, il faudrait donc, premièrement, que cette IA soit dotée d'une conscience d'exister. Comme je l'ai évoqué précédemment, je pense que les IA actuelles ne sont pas conscientes, et dans le même temps, il ne sera jamais possible de savoir si une IA est consciente ou non. Sur ce point donc, je dirais qu'une IA telle que celles qui existent aujourd'hui ne sont probablement pas dotée d'un libre arbitre, puisqu'elle ne sont probablement pas conscientes de leur propre existence. Je propose néanmoins d'examiner, concernant l'IA, dans quelle mesure elles pourraient ou non satisfaire les autres conditions nécessaires à l'existence chez elles d'un libre arbitre.

En ce qui concerne la satisfaction du deuxième critère, à savoir la cohérence entre le désir, les pensées et les actions, moyennant un processus de délibération interne, force est de constater que sans conscience, il n'y a pas de désir. Par conséquent, il n'est encore une fois pas possible d'envisager qu'une IA satisfasse ce deuxième critère sans supposer au préalable qu'elle soit dotée d'une conscience, c'est-à-dire sans qu'elle satisfasse le premier critère.

On pourrait imaginer qu'une IA fasse l'objet d'une simulation de désir, à savoir un mouvement spontané vers un but, sans qu'existe de manière concomitante un ressenti associé. Cependant, le désir est l'intuition d'une possibilité de plaisir ou de bonheur à venir. La manière dont le désir émerge et s'exprime dépend de très nombreux facteurs, parmi lesquels je noterai pour partie des réflexes et comportements codés dans notre cerveau et fruit de notre évolution, les souvenir des expériences personnelles passées, ainsi que les valeurs et ambitions personnelles. Sans cette possibilité de ressenti, il est difficile de concevoir une forme d'accord interne entre désir, pensée et actions, sinon de manière purement abstraite, sémantique et logique, et je dirais, de manière complètement « hors sol ». C'est un peu comme si un cuisinier qui n'avait jamais eu le sens du goût cherchait à inventer de nouvelles recettes, ou comme si un être humain sourd de naissance voulait composer de la musique.

Certes, il est tout à fait possible d'apprendre à une IA quelles sont les caractéristiques de ce qui constitue une « belle » musique, en s'appuyant sur des mesures objectives de fréquence, d'harmonie ou autre, en lui donnant à assimiler le contenu de « belles » musiques. Cependant, même si l'on apprend à cette IA à innover en musique, en lui donnant la possibilité de reconfigurer ses neurones, à en créer de nouveaux, en s'inspirant de toutes les musiques passées, même si on lui fait apprendre les caractéristiques physiques de fréquence, d'harmonie qui font que nous apprécions ou non la musique, il manquera toujours à cette IA la perception et l'appréciation directe de la musique, lui permettant de l'apprécier ou non. Toute la musique qui lui aura été fournie, et tous les critères physiques d'harmonie qui lui auront été donnés, ont tous comme fondement inévitable, à un moment, l'intervention d'un être conscient qui apprécie ou non. Sans cet aiguillon, une IA incapable d'avoir par elle-même conscience et d'apprécier la musique, qui ne serait pas guidée, au moins par des données issues d'êtres eux-mêmes conscients, finira inévitablement par dériver via ses innovations, à plus ou moins long terme, telle un robot sans conscience, sans lumière et sans but.

L'argument précédent est d'autant plus vrai en ce qui concerne la satisfaction par une IA du troisième critère nécessaire à l'exercice du libre arbitre, à savoir l'alignement vis-à-vis des valeurs universelles.

En effet, il est bien possible d'apprendre à une IA les principes universels tels que nous les concevons aujourd'hui. Mais la perception des valeurs universelles évolue avec les civilisations. Il était considéré comme normal de pouvoir se venger à une certaine époque, cela n'est plus vrai aujourd'hui. L'évolution des idées suit un rythme qui dépasse largement la vie de chacun des individus que nous sommes, et chaque génération découvre des idées dont la valeur universelle est plus élevée que les idées jugées universelles des générations antérieures. Dans 200 ans, je l'espère, les futures générations regarderont avec sévérité et incompréhension notre époque et civilisation, disposant des capacités technologiques et ressources nécessaires pour que chacun puisse vivre dignement et en harmonie, et néanmoins incapable d'éviter la pauvreté, famine et la guerre. Une IA pourra sans doute apprendre la photo des valeurs d'une époque. Elle pourra en intégrer les principes analytiques sous-jacents tels que compris à cette même époque. Mais sans capacité à ressentir directement le bonheur, le malheur, la compassion, et sans être guidée par ailleurs par de nouvelles données fournies par des êtres capables d'un tel ressenti, comment pourrait-elle faire autrement que soit rester bloquée dans la vision d'une époque, soit évoluer sur le long terme sans inévitablement dévier du futur des idées universelles, puisqu'incapable de ressentir ce que constitue le malheur ou le bonheur ?

À moins qu'il existe une loi de la nature qui ferait nécessairement évoluer un réseau de neurones vers un alignement vis-à-vis de valeurs universelles, chose très improbable. Une telle IA serait un peu comme un réseau de neurones qui, sans aucune capacité à ressentir ce qu'est une jolie musique, serait capable d'innover et de créer de nouveaux styles musicaux totalement disruptifs et que nous, en tant qu'êtres humains, percevrions spontanément comme beaux...

Le dernier critère évoqué concernant le libre arbitre réside dans la capacité, en dernier ressort, après examen des différentes possibilités d'actions au regard de ses valeurs internes et des valeurs universelles, parmi les choix possibles équivalents restant, de manière non déterminée, c'est à dire indépendamment de toute cause.

Sur ce plan, l'être humain est, à mon sens, doué d'une telle capacité d'action non déterminée, même si je sais que cet avis n'est pas partagé par beaucoup de gens, notamment au sein de la communauté scientifique. En ce qui concerne l'IA, en revanche, la réponse est à mon sens catégorique : il n'existe aucune forme de hasard dans les réponses que celle-ci produit. Certes, il est possible de moduler un paramètre des modèles d'IA appelé température, qui, selon sa valeur, autorisera ou non le fait qu'une même entrée produise ou pas systématiquement la même sortie. Cependant, dans ce cas, le fait qu'une même entrée ne produise pas la même sortie ne signifie pas que cette sortie n'est pas déterminée. La sortie est en réalité dite pseudo-aléatoire, c'est-à-dire parfaitement déterminée par des paramètres qui viennent compléter la requêtes d'entrée et qui permettent de simuler une certaine forme de sortie aléatoire. Si nous prenions en considération tous les paramètres d'entrée, qui vont au-delà du contenu de la requête, alors nous serions capable de déterminer avec certitude la sortie du modèle. C'est un peu comme un algorithme de tirage de dés, qui simule un tirage aléatoire, et qui n'a d'aléatoire que l'illusion que nous en avons.

En résumé, si on récapitule la satisfaction de l'ensemble des critères nécessaires à l'exercice du libre arbitre par une IA, je dirais que si celle-ci ne dispose pas de conscience, alors elle ne satisfait aucun des quatre critères énoncés.

En effet :

- Elle ne dispose d'aucune conscience
- Elle est donc incapable de ressentir un désir, donc d'aligner ce désir avec des valeurs interne, et encore moins des valeurs universelles, sinon vis-à-vis d'une photo figée dans le temps de telles valeurs qui lui auront été apprises, et l'évolution de l'alignement du comportement de cette IA vis-à-vis de valeurs universelles est inévitablement conditionné à la fourniture à cette IA de nouvelles données par des individus qui sont eux conscients d'exister et capables de ressenti.
- Elle ne dispose d'aucune capacité à agir de manière non déterminée indépendamment de toute cause

En définitive, les deux critères que sont l'absence de conscience et le déterminisme résument à eux seuls ce qui différencie fondamentalement une IA d'un être humain, car si une IA avait une conscience et était capable de décider indépendamment de toute cause, alors il n'y aurait plus véritablement d'obstacle à ce qu'elle puisse satisfaire les autres critères que sont l'alignement avec des valeurs internes et l'alignement aux valeurs universelles.

L'IA SANS L'HUMAIN, UNE SCIENCE SANS CONSCIENCE

Nous l'avons vu, il sera sans doute à jamais impossible de déterminer avec certitude si une IA est consciente ou non.
À moins que...
Il me vient alors l'idée suivante, qui pourrait peut-être, finalement, une prolongation du test de Turing : puisque, comme nous l'avons vu, une IA n'a pas de conscience, celle-ci est dépendante d'informations fournies par nous autres humains, qui sommes conscients, pour alimenter et faire évoluer son modèle. Sans de telles informations, une IA, qui ne ressent pas, déviera inévitablement vers « quelque chose » qui ne matche pas avec le ressenti humain, et ne pourra donc pas rester indéfiniment alignées aux valeurs universelles telles que celles-ci évolueront dans le futur.

A contrario donc, si on devait observer qu'une IA, non alimentée par les informations issues de la maturation de notre ressenti, faisait spontanément la preuve d'un tel alignement avec les valeurs universelles futures, et ne déviait pas vis-à-vis de ces valeurs, en gros, si une IA était capable d'un amour universel qui puisse nous surprendre voire de nous dépasser, cela ne constituerait-il pas une preuve irréfutable de l'existence d'une capacité de ressenti interne au sein de cette IA, et donc d'une certaine forme de conscience ? La question reste posée.

On ne va pas se mentir : aujourd'hui, non seulement une IA est très loin de pouvoir réussir un tel test, mais encore, nous sommes totalement incapable de dire si elle le sera un jour.

À supposer cependant qu'une IA ne soit pas dotée de conscience, ce qui signifie qu'elle n'a aucune conscience ni des images qui lui sont soumises, ni des sons, ni des réflexions qui ont lieu en son sein, si une IA n'est pas capable de ressentir quoi que ce soit, de positif ou négatif, si elle n'est pas capable d'appréhender ce que c'est que d'être heureux ou malheureux, si elle ne peut qu'inférer ces concepts de manière totalement abstraite, en s'alimentant de l'expérience vécue par les humains qui ont eux toutes ces facultés, alors une question se pose.

Que peut-on supposer qu'il advienne d'une telle IA, qui, bien qu'autonome et capable d'initiative, traiterait des entrées comme des images ou des sons sans jamais pouvoir les expérimenter ou ressentir ce qu'ils évoquent : la beauté, la souffrance, la joie ?

Lorsque j'étais étudiant, je me rappelle d'une année au cours de laquelle j'ai partagé un appartement avec un autre étudiant d'origine coréenne. Ce dernier m'avait alors partagé le journal intime de son père, qui avait lui-même fait ses études en Europe pendant les années 30. Une phrase de son père, qui m'a alors profondément marqué, est celle-ci : « lorsque j'entends les gens parler autour de moi, j'ai le sentiment que dans leur esprit, la raison s'est détachée du cœur, et telle une machine infernale, elle s'est mise à produire des monstres. ». Son constat était celui d'une personne issue d'un monde étranger à l'Europe d'alors, non imprégnée de la culture et de l'histoire européenne de l'époque, et qui assistait de manière impuissante à la montée de l'intolérance, des populismes et des totalitarismes.

Ce constat, qui illustre le proverbe connu qui dit que « Science sans conscience n'est que ruine de l'âme », est celui qui exprime le fait que le cerveau humain a besoin du cœur, ou de la conscience (les deux termes ayant ici le même sens), pour pouvoir le guider, et resté aligné aux valeurs universelles. Blaise Pascal en son temps avait identifié trois ordres distincts : celui de la matière, celui de la raison, et celui de la charité, chacun de ces trois ordres étant infiniment plus important que le précédent : la raison par rapport à la matière, et la charité par rapport à la raison. Autrement dit, au regard des trois ordres de Pascal, une personne qui aurait la raison la plus développée possible ne serait rien, si elle n'avait pas la charité, c'est-à-dire une conscience qui la guide. On retrouve ici la même idée que précédemment.

Pour l'IA, j'ai le sentiment que c'est la même chose. Nous serons assurément capables un jour de produire des modèles d'IA qui surpasseront, par leurs capacités cognitives, complètement celles d'un être humain. Cette étape de dépassement est par ailleurs régulièrement annoncée par ce que l'on appelle l'AGI. Cependant, une IA dépourvue de conscience et de cœur restera toujours dépendante de l'humain pour rester alignée aux valeurs universelles. Si nous la laissions évoluer de manière totalement autonome, le risque est à mon sens important qu'elle dérive de ces valeurs, et devienne à son tour un monstre.

Dit autrement : puisqu'une IA ne peut pas ressentir directement ce qu'est le vrai, le beau, l'agréable, le bien, le mal, et puisqu'elle ne peut que l'inférer à partir des connaissances qui lui sont données, elle ne peut pas évoluer de manière totalement autonome vers les valeurs de bien, de beau, de vrai sans être constamment alimentée par des êtres humains qui eux peuvent expérimenter cela. Les IA sont comme un cuisinier à qui on demanderait d'innover sans avoir la perception du goût ou un musicien qui devrait innover de manière disruptive sans être capable de ressentir la musique.
Inévitablement, sans capacité de perception directe, il y aura une dérive.

L'INTELLIGENCE DU COEUR

L'humanité est aujourd'hui à la croisée des chemins. Elle a connu un temps l'essor des religions, censées relier les êtres humains entre eux (le terme même de religion, issu de « religo » en latin, signifie « relier »), en les invitant à dépasser leurs limites et leur condition quotidienne, à s'élever vers des valeurs universelles. Puis, les institutions religieuses ayant montré leurs limites et de sévères dérives, il y a eu une phase d'émancipation, avec l'essor des sciences, des technologies, de la médecine, de l'industrie.

La raison a pourrait-on dire pris un temps sa revanche sur les croyances, mais deux guerres mondiales nous ont encore une fois montré les limites et les dérives d'un monde sans valeurs qui ne serait guidé que par la puissance et la science, conduisant André Malraux à déclarer un jour que « le 21ème siècle sera religieux ou ne sera pas ».

Aujourd'hui, l'essor de l'IA alimente tous les fantasmes en faisant entrevoir une nouvelle étape disruptive dans le développement de la science et de la puissance qui l'accompagne.

Sachons garder à l'esprit cet élément fondamental et transcendant qui nous distingue encore aujourd'hui de l'IA, à savoir notre conscience, notre capacité à sentir et ressentir, ces valeurs, cette empathie, cette compassion, cet amour qui nous guide et qui font de nous des êtres profondément humains.

Au fond, la véritable intelligence, la plus haute, n'est-elle pas justement cette clairvoyance, fruit d'un accord parfait entre la raison et le cœur, que l'on appelle l'intelligence du cœur ?

À PROPOS DE L'AUTEUR MATTHIEU CONSTANS

Matthieu CONSTANS a rejoint Softeam il y a 14 ans et accompagne depuis près de 8 ans les clients dans leurs projets de transformation stratégique des Systèmes d'Information.
Directeur au sein du Cabinet, Matthieu contribue activement à l'élaboration des offres de conseil en transformation, en intégrant les enjeux technologiques, organisationnels et humains.
Passionné de philosophie, il s'intéresse aux implications éthiques, sociétales et épistémologiques des technologies émergentes.
Cette double sensibilité nourrit une approche lucide, critique et responsable des transformations numériques.

SOFTEAM
UNE FILIALE DE DOCAPOSTE

Nous accompagnons nos clients, des secteurs publics et privés, dans leur transformation avec du conseil et des services créateurs de valeur, pour construire ensemble un numérique d'intérêt général, responsable et durable.

NOUS CONTACTER

communication@softeam.fr
softeam.com

Crédits visuels : Adobe, Pexel, Pixabay