

Débat n°1 Une machine qui parle est-elle une machine qui pense ?

Texte A Vincent Descombes, *La denrée mentale*, 1995, éd. Minuit coll. Critiques, pp. 154-158

Cet argument, qu'on peut appeler argument par la simulation, est donc le suivant : si une machine pouvait passer l'épreuve de Turing (donc se faire prendre pour un être humain dans un échange mené dans les conditions indiquées), elle devrait être dite intelligente. Ici, pourtant, il faut noter que l'argument est sans doute ambigu, car il a été utilisé dans de bien mauvais raisonnements visant à établir que les machines peuvent penser. Il suffirait, dit-on parfois, qu'un joueur ait perdu au jeu de simulation de l'intelligence pour que le statut intellectuel des ordinateurs soit solidement établi, en vertu d'un postulat phénoméniste¹ : une différence qui n'apparaît pas est une différence qui n'existe pas (ou qui est sans conséquence).

Pourtant, tout le ressort du jeu de l'imitation vient de ce que le quiproquo soit possible : le piquant du jeu « homme ou femme ? », c'est qu'en sortant de la chambre l'interrogateur C découvre que le joueur qu'il prenait pour une femme était un homme. C'est justement parce qu'un homme qui se fait passer pour une femme n'est pas pour autant une femme que le jeu nous amuse. Autrement dit, nous ne nous servons pas de ce jeu pour décider du sexe des êtres humains. Du point de vue philosophique, ce jeu repose donc sur *l'opposition* des apparences et de la réalité, pas sur leur identification dans un phénoménisme. Il ne suffit pas de réussir à s'exprimer comme le ferait une femme, de produire un texte féminin, pour recevoir le statut de femme. Du coup, l'analogie entre le jeu de société et l'épreuve de Turing ne devrait-elle pas être plutôt : il ne suffit pas de réussir à produire des réponses au cours d'une conversation pour être un partenaire intelligent ? Un joueur qui n'est pas une femme peut se faire prendre pour une femme. De même, dans l'épreuve de Turing, le joueur chargé de juger peut-être mystifié, il lui arrivera de croire qu'il a parlé avec quelqu'un alors qu'il n'a eu affaire qu'à une machine.

Les mots « imitation » et « simulation », qui reviennent dans cet argument, risquent de faire penser à des opérations illusionnistes, à des phénomènes de supercherie. En fait, rien n'est prouvé par le fait qu'un automate puisse, dans certaines conditions, se faire prendre pour une personne. Rien n'est établi quant à la psychologie de la machine. En effet, ce n'est pas une propriété psychologique de la machine que de donner l'impression d'être intelligente. C'est une propriété psychologique de *l'observateur* que de pouvoir être leurré ou mystifié par un montage habile des apparences.

Si l'épreuve de Turing est valide, c'est à la condition de recevoir une autre interprétation. Dans le jeu de l'imitation, il s'agissait de juger de la féminité. [...] Dans l'épreuve que subit la machine, il ne s'agit pas de juger de son humanité, mais de sa rationalité, d'une

¹ Phénoméniste : qui s'en tient aux **phénomènes**, c'est-à-dire à ce qui **se manifeste dans l'expérience**, plutôt qu'à une réalité supposée cachée derrière eux.

rationalité définie, comme chez Descartes, par le fait de pouvoir soutenir une conversation². Le problème est donc de savoir si l'interrogateur, ayant découvert qu'il avait reçu les réponses d'une machine, et non d'un homme, perdrait toute envie de continuer la conversation avec un tel partenaire, ou bien s'il pourrait continuer à s'intéresser à sa psychologie, sachant que c'est la psychologie d'une machine.

Telle est la véritable comparaison à faire entre le jeu de l'imitation et le jeu de l'intelligence : le style féminin d'une réponse ne suffit pas à établir la féminité de l'auteur de cette réponse, mais l'intelligence de la réponse pourrait bien suffire à établir l'intelligence de l'auteur de cette réponse.

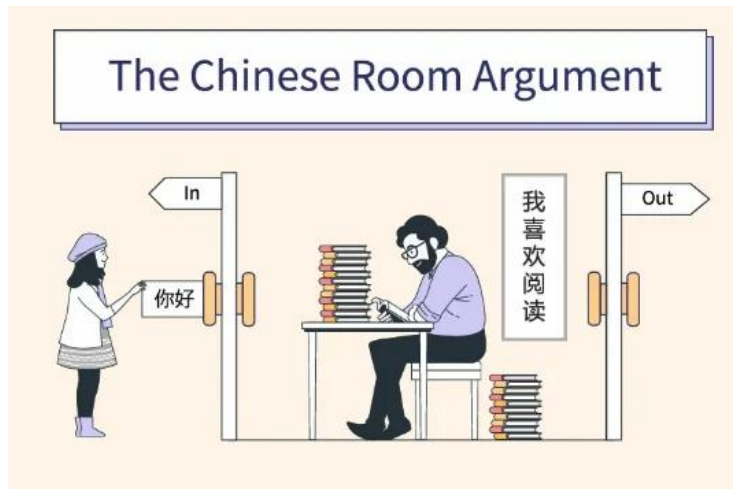
[...] Ce qui compte n'est pas que la machine *se fasse prendre* pour un être humain, car cela ne nous apprend rien sur les capacités de la machine : seule est alors en cause la capacité de discernement du mystifié. En revanche, il est fort important qu'on puisse prendre une machine à la place d'un être humain pour remplir une certaine fonction. En somme, la notion d'imitation (et, plus encore, celle de simulation) peut être entendue en deux sens. S'agissait-il pour le simulateur de se faire passer pour capable d'un certain travail, alors qu'en réalité il n'en est pas capable ? C'est l'imitation au sens d'une fraude, d'une supercherie. Ou bien s'agissait-il de se faire passer pour le titulaire normal d'un poste en faisant son travail à sa place et de façon aussi satisfaisante ? Dans ce cas, la « contrefaçon » est aussi bonne que le produit authentique, et il y aurait snobisme à préférer le vrai au faux. [...] Si les productions d'un logographe³ artificiel étaient de même qualité que celles des logographes humains, on devrait reconnaître aux uns et aux autres les mêmes compétences.

Questions

1. Selon Vincent Descombes, quel contresens faut-il éviter au sujet du test de Turing ? Utilisez l'analogie avec le jeu "homme ou femme ?" pour illustrer les limites de ce test.
2. Quels sont les deux sens de la notion d'imitation qui sont distingués dans ce texte ?
3. Selon l'auteur, comment faut-il comprendre le test imaginé par Turing ?
4. En quoi serait-il snob, dans certains cas, de considérer une « contrefaçon » comme moins valable que l'original ? Cherchez un autre exemple que celui du logographe.
5. Selon Descombes, qu'est-ce qui permettrait réellement d'attribuer une compétence intellectuelle à une machine ?

² Pour Descartes, parler de façon libre et adaptée dans une conversation est un signe que l'on possède la raison. C'est selon lui ce qui distingue les humains des machines ou des animaux (*Discours de la méthode*, cinquième partie (1637)).

³ Logographe : personne (ou ici machine) qui écrit des discours ou textes pour quelqu'un d'autre.



Je ne connaissais rien alors à l'intelligence artificielle. J'ai acheté un manuel au hasard, dont la démarche argumentative m'a sidéré par sa faiblesse. Je ne savais pas alors que ce livre allait marquer un tournant dans ma vie. Il expliquait comment un ordinateur pouvait comprendre le langage. L'argument était qu'on pouvait raconter une histoire à un

ordinateur et qu'il était capable ensuite de répondre à des questions relatives à cette histoire bien que les réponses ne soient pas expressément données dans le récit. L'histoire était la suivante : un homme va au restaurant, commande un hamburger, on lui sert un hamburger carbonisé, l'homme s'en va sans payer. On demande à l'ordinateur : « A-t-il mangé le hamburger ? » Il répond par la négative. Les auteurs étaient très contents de ce résultat, qui était censé prouver que l'ordinateur possédait les mêmes capacités de compréhension que nous. C'est à ce moment-là que j'ai conçu l'argument de la chambre chinoise : supposons que je sois dans une pièce fermée avec la possibilité de recevoir et de donner des symboles, par l'intermédiaire d'un clavier et d'un écran, par exemple. Je dispose de caractères chinois et d'instructions permettant de produire certaines suites de caractères [...]. Vous me fournissez l'histoire puis la question, toutes deux écrites en chinois. Disposant des instructions appropriées, je ne peux que vous donner la bonne réponse, mais sans avoir compris quoi que ce soit, puisque je ne connais pas le chinois. Tout ce que j'aurais fait, c'est manipuler des symboles qui n'ont pour moi aucune signification. Un ordinateur se trouve exactement dans la même situation que moi dans la chambre chinoise : il ne dispose que de symboles et de règles régissant leur manipulation. [...] L'argument de la chambre chinoise montre que la sémantique⁴ du contenu mental n'est pas intrinsèque⁵ à la syntaxe⁶ du programme informatique, lequel est défini syntaxiquement par une suite de 0 et de 1.

Questions

1. Analysez l'exemple tiré d'un manuel sur l'intelligence artificielle : pourquoi pourrait-on penser au premier abord qu'il prouve l'intelligence de l'ordinateur ?

⁴ Sémantique : sens ou signification des mots, des phrases ou des symboles.

⁵ Intrinsèque : qui appartient à la chose elle-même, qui est interne, et non attribué de l'extérieur.

⁶ Syntaxe : règles formelles de manipulation des signes (ex. la grammaire d'une langue ou les opérations d'un programme informatique).

2. En quoi consiste l'expérience de pensée appelée « chambre chinoise » imaginée par Searle ?
3. Quelle analogie Searle établit-il entre cette situation et le fonctionnement d'un ordinateur ?
4. Quelle différence Searle établit-il entre syntaxe et sémantique ?
5. Pourquoi, selon lui, la syntaxe d'un programme informatique ne suffit-elle pas à produire du sens ?

Searle Extrait 2 « *Esprits, cerveaux et programmes* », paru dans la revue *Behavioral and Brain Sciences* (Cambridge University Press, 1980)

Les manipulations formelles de symboles n'ont en elles-mêmes aucune intentionnalité ; elles sont tout à fait dénuées de sens ; ce ne sont même pas des manipulations de symboles, puisque les symboles ne symbolisent rien. Dans le jargon linguistique, elles n'ont qu'une syntaxe, mais aucune sémantique. L'intentionnalité que semblent avoir les ordinateurs n'existe que dans l'esprit de ceux qui les programment et de ceux qui les utilisent, de ceux qui envoient les données d'entrée et de ceux qui interprètent les données de sortie. [...]

L'idée selon laquelle les simulations par ordinateur pourraient être la chose même aurait dû d'emblée éveiller la suspicion, car l'ordinateur n'est absolument pas confiné à la simulation d'opérations mentales. Personne n'imagine que les simulations sur ordinateur d'un incendie catastrophique vont réduire en cendre le voisinage, ni que la simulation d'une pluie de tempête nous laissera trempés. Pourquoi donc quelqu'un irait-il imaginer que la simulation sur ordinateur de l'entendement comprend quoi que ce soit ? On dit parfois qu'il serait terriblement difficile de faire ressentir de la douleur à un ordinateur ou de le faire tomber amoureux. Pourtant, amour et douleur ne sont ni plus ni moins difficiles que la cognition ou que n'importe quoi. Pour simuler, tout ce qu'il faut ce sont les bonnes entrées et sorties avec un programme entre les deux qui transforme les unes en les autres. C'est tout ce dont dispose l'ordinateur pour tout ce qu'il est appelé à faire. En confondant la simulation et la reproduction, on commet la même erreur, qu'il s'agisse de douleur, d'amour, de cognition, d'incendie ou de pluie de tempête.

Questions

1. Quelle différence Searle établit-il entre simuler un phénomène et le produire réellement ?
2. Quels exemples utilise-t-il pour illustrer cette distinction ?
3. Quelle erreur commet-on lorsqu'on confond simulation informatique et activité mentale réelle ?
4. Selon Searle, qu'est-ce qui manque aux ordinateurs pour qu'on puisse dire qu'ils pensent réellement ?

Wittgenstein, dans la *Grammaire philosophique*, fait remarquer que :

« Si l'on pense à la pensée comme à quelque chose de spécifiquement humain, organique, on aimerait demander : "Pourrait-il donc y avoir une prothèse de pensée, un substitut anorganique pour la pensée ?" Mais si maintenant la pensée consiste dans le fait d'écrire ou de parler, pourquoi une machine n'est-elle pas censée faire cela ? - "Oui, mais la machine n'est au courant de rien !" - Il est vrai que parler d'une prothèse du voir ou de l'entendre n'a pas de sens. On parle assurément d'un pied artificiel, mais pas de douleurs dans le pied artificielles » (PG, § 64).

Se demander quelles sont, parmi nos activités intellectuelles et, plus généralement, mentales, celles pour lesquelles on peut parler de façon sensée de prothèses susceptibles de se substituer à l'organe naturel qui les effectue et de remplir sa fonction aussi bien et peut-être même éventuellement mieux que lui est effectivement une question intéressante. Il est significatif, par exemple, que l'on parle couramment d'« intelligence artificielle », ce qui est plus facile à comprendre puisqu'il y a un nombre important et, semble-t-il, de plus en plus grand de comportements intelligents qui peuvent être simulés par des machines, mais pas (ou pas encore ?) de « pensée artificielle », probablement parce qu'on a l'impression d'avoir affaire le plus souvent à des dispositifs artificiels qui sont capables de faire effectivement la même chose que des êtres pensants, mais justement sans avoir besoin pour cela de penser. Il est vrai que cela pourrait constituer aussi une raison de se demander si les êtres pensants en question ne sont pas, eux aussi, après tout des machines. C'est une question qui, aux yeux de Wittgenstein, n'a rien d'illégitime :

« "Mais une machine pourrait-elle penser ?" - Pourrait-elle avoir des douleurs ? Ici tout dépend de ce qu'on entend par : "que quelque chose a des douleurs". Je peux considérer autrui comme une machine qui a des douleurs, c'est-à-dire : l'autre corps. En revanche, le phénomène des douleurs que je décris, quand je dis par exemple "j'ai des maux de dents", ne présuppose pas un corps physique. (Je peux avoir des maux de dents sans dents.) Et ici la machine n'a pas du tout de place. - C'est une chose claire, la machine ne peut remplacer qu'un corps physique. Et dans le sens auquel on peut dire d'un tel corps qu'il a des douleurs, on peut le dire aussi de la machine. Ou une fois encore : les corps dont nous disons qu'ils ont des douleurs, nous pouvons les comparer à des machines, et également les appeler des machines » (ibid.).

Or la chose dont on se demandait si elle pouvait ou non être comparée à une machine n'était justement pas le corps, pour lequel cela ne pose pas de problème réel, mais l'esprit, pour

lequel on ne sait pas très bien si une question de cette sorte a ou non un sens, ce qui, bien entendu, n'est pas du tout la même chose que de dire que la réponse ne peut être, dans son cas, que négative. Il se pourrait, bien entendu, que les machines intelligentes deviennent capables de performances telles et que l'usage que nous faisons du mot « pensée » se modifie en même temps à tel point que nous trouvions à un moment donné tout à fait possible et normal de dire de certaines de nos machines qu'elles pensent et que, par conséquent, une machine peut penser. Turing était d'avis que nous en serions probablement arrivés à ce stade avant la fin du siècle. C'était un pronostic pour le moins risqué mais, même d'un point de vue wittgensteinien, pas nécessairement choquant. L'erreur qui est commise ici le plus souvent est, selon Wittgenstein, celle qui consiste à croire qu'un effort de réflexion philosophique supplémentaire devrait permettre de trancher dès à présent la question dans un sens ou dans l'autre et que c'est, pour un philosophe, une sorte d'obligation de leur donner une réponse, alors que les questions de cette sorte, quand elles finissent par se décider, le font d'une façon bien différente de celle-là, qui fait intervenir bien d'autres facteurs et dans laquelle la philosophie n'est généralement pas pour grand-chose.

Questions

1. Que critique Wittgenstein dans l'idée de « prothèse de pensée » ? Expliquez l'analogie avec la douleur dans le pied.
2. Quelle différence Bouveresse fait-il entre « intelligence artificielle » et « pensée artificielle » ?
3. Dans le deuxième extrait de Wittgenstein, comment comprenez-vous la phrase « la machine ne peut remplacer qu'un corps physique » ?
4. Selon Jacques Bouveresse, comment pourrait-on en venir à régler la question « Les machines peuvent-elles penser » ?

Débat n°2 Peut-on réduire l'intelligence humaine à du calcul ?

Texte A Yann Le Cun, *Quand la machine apprend. La révolution des neurones artificiels et de l'apprentissage profond*, mai 2023.

La plupart des scientifiques acceptent aujourd'hui l'idée que le cerveau est une machine biochimique. Une machine compliquée, certes, mais une machine. Un neurone réagit par signal électrique à ses entrées. En fonction de ce qu'il reçoit de ses neurones en amont, il fait un calcul qui aboutit ou non à une impulsion électrique, un potentiel d'action, ou *spike*, qu'il envoie à tous ses neurones en aval. Le mécanisme est assez basique. Mais en combinant l'activité de milliards de ces éléments relativement simples que sont les neurones, on obtient le cerveau et l'esprit humains.

J'ai conscience que l'idée de modéliser le cerveau de l'homme peut faire bondir certains philosophes ou ceux qui ont la foi. Mais nous sommes nombreux, parmi les scientifiques, à penser que les mécanismes de la pensée seront à terme reproductibles par les systèmes d'intelligence artificielle capables d'apprentissage.

Ceux qui le contestent tirent argument du fait que nous sommes loin de comprendre comment le biologique, la physique, le quantique et les autres systèmes s'associent dans l'organisme pour faire fonctionner le cerveau. Nous ne comprenons pas tout, en effet, mais je suis convaincu que les cerveaux des mammifères ou des êtres humains sont des machines qui « calculent », et que ces calculs sont, en principe, reproductibles par une machine électronique, un ordinateur.

Le complexe de supériorité de l'humain risque d'en prendre un coup. Comme l'a écrit Sigmund Freud, cité par l'illustre biologiste de l'évolution Stephen Jay Gould :

« Les principales révolutions scientifiques ont en commun d'avoir toutes détrôné l'arrogance humaine des piédestaux successifs de nos anciennes certitudes quant à notre position centrale dans le cosmos.

Questions

1. Quelle conception de l'esprit humain est implicite dans l'idée que le cerveau est une « machine biochimique » ?
2. Pourquoi « l'idée de modéliser le cerveau de l'homme » peut-elle « faire bondir certains philosophes ou ceux qui ont la foi » ?
3. Que peut-on qualifier de « complexe de supériorité de l'humain » ? Illustrez par un exemple.

Supposons que les dispositifs d'IA réussissent toutes les prouesses qu'on nous annonce, même les plus hyperboliques⁷. Est-ce que cela en ferait des quasi-hommes ou des surhommes ?

D'abord, comme il ressort du test actuellement le plus utilisé, le benchmark⁸ ARC proposé en 2019 par François Chollet, l'intelligence mesurée est celle définie par la psychométrie⁹, donc celle que mesurent les tests de QI. Or les tests cognitifs psychométriques posent de nombreux problèmes épistémologiques¹⁰ (notamment parce qu'ils concluent à partir d'une performance mesurée dans un contexte concret spécifique à l'existence d'une structure ou faculté stable). Par ailleurs, ils ne mesurent que des types spécifiques d'intelligence. Ils sont aveugles en particulier à ce qui relève de l'intelligence pratique, de l'intelligence créatrice, de l'intelligence sociale, de l'intelligence machiavélique¹¹, etc. Enfin, bien que les directions de ressources humaines soient friandes des tests de QI, ils n'ont qu'une valeur prédictive faible, ce qui ne peut pas ne pas se répercuter sur la valeur prédictive des benchmarks.

En deuxième lieu, et c'est le point décisif à mes yeux, la thèse forte de l'IA qui est à l'origine de l'imaginaire de la naissance imminente d'un nouvel Adam ou d'une nouvelle Ève, engendrés « algorithmiquement », présuppose une conception de l'Homme qui le réduit à ses fonctions cognitives. Mais les humains ne sont pas (malheureusement ou heureusement) de purs esprits calculateurs et ratiocineurs¹². Ils sont des êtres biologiques et sociaux mortels, guidés par des besoins, des désirs, des détestations, des réactions émotives, des valeurs multiples et parfois contradictoires, cherchant à éviter la mort, la douleur et le malheur, recherchant le plaisir et le bonheur, mus éventuellement par des aspirations de transcendance, autant de tendances qui en partie sont subconscientes et biaisent la plupart de leurs calculs rationnels. On ne saurait réduire aucune de ces conduites à une quête optimisée de croyances justifiées rationnellement. Pour autant, nous serions bien imprudents d'y voir des séquelles évolutives dont nous devrions nous défaire : cela nous transformerait en zombis¹³, peut-être rationnels, mais sans vie. Si au contraire nous acceptons que ce sont des propriétés constitutives de ce que c'est que d'être un homme, alors les algorithmes devraient aussi les implémenter pour pouvoir prétendre être les égaux des hommes. On ne

⁷ les promesses le plus **hyperboliques** : les plus exagérées ou spectaculaires faites par certains discours sur l'IA.

⁸ **Benchmark** : test standardisé servant à mesurer et comparer les performances d'un système, notamment en intelligence artificielle.

⁹ **Psychométrie** : méthode de mesure des capacités mentales à l'aide de tests standardisés comme les tests de QI (quotient intellectuel).

¹⁰ **Épistémologique** : relatif à l'étude critique des méthodes et de la validité des connaissances scientifiques.

¹¹ **Machiavélique** : qui renvoie à une forme d'intelligence fondée sur la stratégie, la ruse et la capacité à manipuler les autres pour atteindre ses objectifs.

¹² **Ratiocineur** : qui raisonne de façon très logique et abstraite

¹³ **Zombi (en philosophie de l'esprit)** : être qui agit comme un humain mais sans conscience ni expérience intérieure.

peut que souhaiter bien du courage aux programmeurs d'algorithmes pour réussir à simuler cet embrouillamini indescriptible (du point de vue de tout calcul rationnel) qu'est pourtant tout être humain ! Il est tragique que grâce au Web des imaginaires eschatologiques archaïques¹⁴ colonisent les esprits humains au même moment où les industriels de l'armement commencent à expérimenter, sur une grande échelle et sur les champs de bataille réels (en particulier ceux de la guerre d'agression de la Russie contre l'Ukraine), l'efficacité meurtrière de l'IA. Le danger véritable de l'IA ne réside donc pas dans le risque (fantasmé) qu'elle se tourne contre les hommes mais dans celui, bien réel, de ses utilisations néfastes par les hommes. Même à l'âge de l'intelligence artificielle, le plus grand ennemi de l'homme reste toujours l'homme - ce qui est à la fois rassurant et terrifiant.

Questions

1. Selon Schaeffer, quels problèmes épistémologiques les tests utilisés pour évaluer les performances de l'IA posent-ils ?
2. Relevez et expliquez trois limites des tests de QI selon l'auteur.
3. En quoi les « besoins, désirs, détestations, réactions émotives, valeurs multiples et parfois contradictoires » qui nous guident pourraient-ils être considérés comme des « séquelles évolutives dont nous devrions nous défaire » ?
4. Est-il souhaitable, selon Scheffer, de transformer l'être humain en un être purement rationnel ? Pourquoi ?
5. Relevez des dangers fantasmés et des dangers réels de l'IA : quel est le principal critère qui les distingue d'après Schaeffer ?

¹⁴ **Imaginaires eschatologiques archaïques** : vieux mythes sur la fin de l'humanité ou l'arrivée d'un monde nouveau.

L'intelligence artificielle est une logique pure désolidarisée de la pensée humaine.

Inutile ici de décider s'il existe une ou plusieurs logiques, donc un ou plusieurs systèmes de règles déductives. L'important est que tous ces systèmes logiques ont des propriétés qu'on peut aussi programmer sous forme de software¹⁵.

C'est pourquoi une IA n'est pas susceptible d'erreurs. Elle peut certes s'abîmer, être attaquée par des virus et elle nécessitera toujours des mises à jour parce qu'aucun software ne sera jamais parfait. Mais quand mon ordinateur tombe en panne, bogue, il ne fait pas d'erreur. Les opérations de mon ordinateur sont d'ordre purement logique. Il n'y a pas de psychologie d'ordinateur.

L'ère numérique est celle du pouvoir de la logique sur la pensée humaine. Notre pensée suit les règles de la logique pour atteindre un but. Ce qui ne signifie pas que nous pensions vraiment logiquement, que de nos pensées nous détachions peu à peu des concepts qui se laissent appréhender comme des algorithmes, et que nous combinions ces pensées avec de nouvelles pensées selon les règles de la logique. En tant qu'êtres vivants et pensants, nous travaillons sous la contrainte du temps et nous préférons souffrir des erreurs de logique que calculer trop longtemps pour obtenir un résultat. De plus, nous sommes des corps vivants soumis à des émotions, nous avons des différences de caractère, nous faisons la part des choses et ne recourons pas toujours à la logique pour réaliser nos fins. [...] Notre psychologie quotidienne est un chaos de bêtises que nous structurons socialement pour que notre civilisation ne s'écroule pas sous ce perpétuel cafouillage mental pour le moins bigarré. Les hommes ne sont pas toujours intéressés à une rationalité logique optimale, ils agissent sous des influences radicalement différentes de celles de la logique, auxquelles il n'y a rien à redire parce que (c'est ainsi et pas autrement), nous ne sommes pas des Mister Spock¹⁶. C'est justement parce que les modules sur lesquels nous pouvons faire fond sont différents que nous sommes des sujets faillibles certes, mais dotés d'objectivité¹⁷. L'intelligence artificielle n'est pas une copie de la pensée humaine. Plus exactement, c'est un modèle de pensée, une carte logique de notre pensée – après qu'ont été éliminés la contrainte du temps et nos besoins d'humains voués à la finitude, incapables de penser s'ils ne disposaient pas d'un corps mortel qui participe intensément à leurs intérêts.

¹⁵ **Software** : programmes et instructions qui font fonctionner un ordinateur.

¹⁶ **Mister Spock** est un personnage de la série télévisée Star Trek (1966). Né en 2230 sur la planète Vulcain, il utilise son esprit logique (hérité de son père vulcain) pour contrôler son émotivité (héritée de sa mère humaine)

¹⁷ M. Gabriel écrit p.72 : « L'objectivité est ce point de vue qui fait que nous puissions nous tromper à propos de quelque chose ou avoir raison. On prendra garde que l'objectivité ne signifie pas que nous prenons un point de vue de nulle part ou que nous appréhendons le réel de manière totalement neutre. Bien au contraire, l'être de l'objectivité c'est qu'on lui attribue une subjectivité. La subjectivité est la manière dont nous pouvons nous tromper. Quand je me trompe et la manière dont je me trompe renseignent sur ma personnalité. Sans subjectivité, pas d'objectivité et réciproquement. »

Questions

1. Pourquoi M. Gabriel dit-il que les êtres humains préfèrent parfois se tromper plutôt que de calculer trop longtemps ?
2. Quels rôles jouent le corps, les émotions et le caractère dans la pensée humaine ?
3. Comment restons-nous civilisés malgré notre « chaos de bêtise » ?
4. En quoi notre caractère faillible est-il une richesse ?
5. Quelles différences fondamentales entre l'IA et l'intelligence humaine l'auteur souligne-t-il ?

Débat n°3 Puis-je être ami(e) avec mon IA ?

Texte A - Laurence Devillers *Relation émotionnelle entre humains et robots : Quelle éthique ?*
Sorbonne Université, GEMASS/LIMSI/CNRS, membre de la CERNA Commission de réflexion sur l'Éthique de la Recherche Nationale. In *Les signatures neurobiologiques de la conscience*, pp.289-300, 2021.

L'informatique affective a ouvert un champ important d'applications que ce soit pour le diagnostic médical, la surveillance, l'assistance en santé ou encore l'éducation à l'aide de logiciels ou encore de robots. Même si les performances de ces modèles ne sont pas très robustes, ces modélisations engendrent chez les humains des comportements surprenants. L'anthropomorphisme, c'est-à-dire la tendance à attribuer aux machines des réactions humaines est alors bien réel. L'intelligence artificielle permet de détecter ou simuler les affects, elle ne peut pas les faire ressentir à la machine. La modélisation des émotions ne touche que la composante expressive, il n'y a ni sentiment, ni désir, ni plaisir, ni intention dans une machine. Mais ces machines numériques ont beau ne pas être vivantes, elles sont bien présentes. La modélisation informatique des affects amène à se poser la question des conséquences sociétales de vivre dans un quotidien environné d'objets pseudo-affectifs. Les chatbots aussi appelés agents conversationnels et les robots sociaux peuvent déjà embarquer des systèmes de détection, de raisonnement et de génération d'expressions affectives qui même avec des erreurs importantes peuvent interagir avec nous. Ils sont cependant loin d'avoir des capacités sémantiques suffisantes pour converser et partager des idées mais ils pourront bientôt détecter notre malaise, notre stress et peut-être nos mensonges. Notre imaginaire et nos représentations symboliques vont évoluer au contact des machines, surtout si celles-ci nous parlent et simulent de l'empathie. Comment éviter l'isolement, la déshumanisation et l'appauvrissement de la vie sociale, si les interactions des humains sont réduites à discuter avec des machines affectives ? Il est nécessaire de réfléchir à la co-adaptation humain-machine, aux bénéfices et aux risques engendrés par ces objets et aux garde-fous à imaginer pour éviter d'être trop manipulés par ces machines. Il importe également de savoir si nous sommes prêts à accepter des robots capables de détecter nos émotions, d'y réagir et par exemple de simuler de l'empathie ; et si nous pourrions un jour nous attacher à des machines affectives comme nous nous attachons à un animal domestique. Dans ce cas, quels statuts auront ces individus numériques dans nos sociétés ? Même si ces objets ne sont pas réellement émotionnels, conscients et autonomes, ils vont envahir notre quotidien. Quels garde-fous seront développés ? L'alphabétisation de l'IA, l'éducation et l'expérimentation de ces machines affectives sont nécessaires pour que nous puissions prendre suffisamment de distance devant ces objets affectifs qui vont envahir notre quotidien.

Questions

1. Quels types de comportements humains la capacité de la machine à détecter ou simuler des affects provoque-t-elle ?
2. Pourquoi Laurence Devillers parle-t-elle de « répercussions sociétales » liées à la présence de machines pseudo-affectives dans notre quotidien ? Illustrez par un exemple.
3. Comment comprenez-vous la notion de « co-adaptation humain-machine » dans ce texte ?
4. Quels types de garde-fous ou de mesures préventives sont suggérés pour éviter une manipulation excessive par ces machines ?

L'adoption des technologies numériques ne peut être réduite à une simple décision rationnelle ou à une évolution fonctionnelle des pratiques et des usages. Elle détache, progressivement ou non, les individus de leurs cadres de référence initiaux et de leurs structures habituelles, pour les immerger dans des environnements régis par des logiques externes, celles imposées par la technologie elle-même. Ce basculement représente une reconfiguration profonde des structures cognitives, sociales et comportementales de l'individu, sous l'effet de logiques algorithmiques et prescriptives qui supplantent ses propres cadres de référence. Ce processus de transition technologique, loin d'être neutre, s'apparente à une forme de capture symbolique dans laquelle l'individu, confronté à la violence du changement, active des mécanismes de défense psychiques face à ce qu'il perçoit comme une agression à son autonomie, à son libre-arbitre, et à son intégrité identitaire.

Lorsque l'adoption est jugée réussie, cela signifie que les structures de défense initiales ont cédé : l'utilisateur a non seulement intégré les règles imposées par la technologie, mais a développé une forme d'identification affective envers elle, réinterprétant l'origine de la contrainte comme une relation choisie. À ce stade, un nouveau régime normatif se met en place. Ce basculement marque la substitution de l'ancien cadre de référence par celui de la machine, désormais perçu comme familier et rassurant. L'agression initiale est refoulée, et les nouveaux automatismes cognitifs deviennent objets de défense.

Ce phénomène, que l'on peut assimiler au « syndrome de Stockholm¹⁹ » dans le rapport entre l'homme et la machine, implique une dislocation des référents cognitifs, suivie d'une reconfiguration émotionnelle dans laquelle la victime en vient à protéger son agresseur technologique. L'asservissement cognitif ainsi produit n'est pas un effet secondaire, mais un mécanisme de survie. Il est alimenté par les tentatives du cerveau de réduire le stress généré par l'intrusion d'un cadre de pensée étranger. Cette réécriture émotionnelle permet d'assurer une forme de cohérence interne face à l'aliénation technologique. L'attention de l'utilisateur se détourne alors de la violence initiale pour se fixer sur les signaux positifs émis par la machine : validation sociale, gratification algorithmique, récompense ludique. Ces stimuli activent le biais de confirmation émotionnelle²⁰ et transforment la coercition en bienveillance perçue.

¹⁸ Hamilton Mann est vice-président du groupe chargé du marketing numérique et de la transformation numérique chez Thales et maître de conférences à l'INSEAD.

¹⁹ **Syndrome de Stockholm** : en psychologie, désigne une situation dans laquelle une victime développe une forme d'attachement, de sympathie ou de loyauté envers son agresseur ou son ravisseur, souvent comme mécanisme de défense face à une situation de contrainte ou de stress.

²⁰ **Biais de confirmation émotionnelle** : tendance à retenir surtout les informations qui confirment ce que l'on ressent déjà.

Par un processus de plasticité neuronale, les circuits cérébraux réorganisent la perception du rapport à la machine : ce qui était stress devient normalité ; ce qui était domination devient soutien ; et ce qui était agresseur devient compagnon. [...] Ce phénomène représente l'un des périls fondamentaux que l'intelligence artificielle fait peser sur l'humanité : la normalisation de la dépendance mentale comme vecteur d'acceptabilité sociale. C'est pourquoi il ne suffit pas de concevoir des systèmes artificiellement intelligents : il faut impérativement leur adjoindre une capacité d'intégrité artificielle, garante de la souveraineté cognitive humaine.

[...] Le paradoxe qui empêche la reconnaissance systémique de ce syndrome est celui de l'innovation elle-même. Perçue comme intrinsèquement positive, elle dissimule son potentiel ambivalent : elle peut à la fois émanciper et aliéner, selon les conditions de son adoption.

Pour que l'intelligence artificielle renforce notre humanité sans la diluer, elle devra, au-delà de sa capacité à imiter la cognition, être fondée et guidée par une intégrité artificielle, afin de respecter les libertés mentales, émotionnelles et identitaires des individus. La technologie peut soulager la douleur, limiter le risque et améliorer l'existence. Mais aucune avancée ne doit se faire au prix d'une dette cognitive qui ruinerait notre capacité à penser par nous-mêmes, et avec elle, notre rapport à notre propre humanité.

Questions

1. Comment comprenez-vous l'idée que des « logiques algorithmiques et prescriptives » viennent remplacer nos « cadres de référence initiaux » ? Cherchez un exemple qui pourrait illustrer cette idée.
2. En quoi le « processus de transition technologique » peut-il être perçu par l'individu comme « une agression à son autonomie, à son libre-arbitre, et à son intégrité identitaire. »
3. En quoi peut-on parler de « syndrome de Stockholm » dans le rapport des humains à leurs IA ?
4. Pourquoi ce syndrome n'est-il reconnu par les utilisateurs d'après Hamilton Mann ?
5. Comment comprenez-vous la notion « d'intégrité artificielle » ? Vous semble-t-elle réalisable ?

Extrait 1, p.124

Ce qui rend un Paro vivant, ce n'est pas qu'il montre qu'il nous aime, mais que nous l'aimons nous-mêmes à travers les marques d'affection qu'il nous prodigue et celles qu'on lui donne en retour. Le Paro est un robot de compagnie très populaire au Japon qui a la forme d'un petit phoque en peluche. Le succès du Paro est lié à la conception du robot et aux mécanismes qui simulent de façon très réaliste une créature vivante - avec des ronronnements, des ondulations de peau, etc. Mais il ne devient vraiment vivant que parce que nous le sentons ainsi car il nous engage vis-à-vis de lui dans une telle relation. Nous construisons avec lui un espace dans lequel le fait qu'il soit vivant a un sens. Il n'y a pas de différence entre l'affection et la simulation de l'affection parce que cette simulation de l'affection peut être aussi convaincante qu'une affection qui ne serait apparemment pas simulée. Un esprit positiviste objectera que l'on est dans l'illusion, mais quand un enfant devient extrêmement malheureux parce qu'on veut remplacer son Aibo habituel par un autre Aibo, exactement pareil au premier, il ressent des émotions négatives qui n'ont rien de simulées et qui l'affectent comme n'importe quelles autres émotions négatives. L'Aibo, construit par Sony, avec une forme de petit chien, a été le plus populaire des robots de compagnie. La clef du phénomène réside dans un processus très simple et pourtant sous-estimé : la non-transitivité des simulations d'affects - je peux ressentir des affects très intenses à partir d'autres affects purement simulés.

Questions

1. En quel sens peut-on dire selon Lestel que les robots de compagnie sont vivants ?
2. Comment comprenez-vous « la non-transitivité des simulations d'affects » ?

L'opinion dominante voit l'amitié comme une conduite noble qui mène à des obligations spéciales entre amis et qui débouche sur des formes contraignantes de responsabilité - une façon d'être (plutôt qu'une façon de faire) qui est jugée hors de portée d'un artefact. [...] Soins mutuels, intimité et activité partagée sont considérés comme les ingrédients indispensables qu'on est supposé trouver dans toute cuisine amicale. Dans le contexte d'un tel cadre intellectuel, une amitié qui passe par une matérialité purement numérique n'est pas « réelle » ; elle distord quelques-uns des aspects les plus essentiels de ce qu'est l'amitié et en pervertit la nature même. L'amitié réelle ne doit pas passer par des mises en œuvre mécaniques, ni même par des obligations purement sociales - et c'est là l'un des nombreux paradoxes de la notion traditionnelle d'amitié, qui obéit nécessairement à une obligation qui ne relève pas de la nécessité. La multiplicité des « amis » ne peut donc qu'être le signe que ce ne sont pas vraiment des amis - avoir par exemple des centaines d'amis sur Facebook est jugé ridicule. [...] Une amitié véritable implique une interaction substantielle dans un monde réel, et c'est l'une des raisons de la disqualification de la multiplicité des prétendants : il est impossible d'accorder à chaque ami le temps et l'énergie qu'une amitié requiert si le nombre des amis devient astronomique. Seule peut-être qualifiée de réelle et avoir une valeur morale une amitié rare et concrète. Toutes les autres sont des illusions. On pourrait alors en déduire que l'amitié entre humains et robots est plus « authentique » que l'amitié entre deux humains à travers Internet, mais ce n'est évidemment pas la conclusion qui est habituellement adoptée. D'où l'hypothèse que quels que soient les cas de figure rencontrés, dans la littérature à prétention philosophique qui s'intéresse aux amitiés artificielles, toute amitié qui s'établit en dehors de l'espace sacré des humains relève toujours du leurre. Reste à comprendre ce qui pose un problème, et si on doit nécessairement s'en tenir là.

Questions

1. Pourquoi, dans la conception classique de l'amitié, l'amitié avec un robot semble-t-elle impossible ou illusoire ?
2. En quel sens une amitié avec un robot peut-elle pourtant être considérée comme plus « authentique » que celle entre deux humains à travers un réseau social ?

L'avantage majeur du robot ami sur l'ami humain n'est [...] pas négligeable ; se lier avec lui n'implique aucune responsabilité alors que lui-même montre vis-à-vis de l'humain auquel il est lié un attachement absolu. Cette dissymétrie qu'on peut juger immorale a évidemment de quoi troubler quiconque persiste à penser l'amitié comme un comportement moral, mais c'est précisément l'avantage de l'ami artificiel de placer enfin l'amitié sur un terrain purement fonctionnel, débarrassé des asservissements de l'empathie et de la morale qui en découle.

[...] La question qui est aujourd'hui la plus intéressante pour un philosophe à propos des robots générateurs d'amour, de tendresse et d'affection, est celle de savoir quels genres de vies on peut inventer quand on est libéré de la tyrannie de l'authenticité et des contraintes du monopole humain sur l'amitié et l'amour.

[...] Si des artefacts pouvaient devenir ces agents qu'on appelle « amis » de façon générique, c'est-à-dire ceux qui nous soutiennent en cas de coup dur, avec lesquels on peut se réjouir quand ça va et sur lesquels, surtout, on peut toujours compter quand on en a besoin - on pourrait consacrer les relations humaines « authentiques » à des registres beaucoup moins contraignants et les faciliter ainsi grandement. Avoir une compagne (ou un compagnon) parfait(e) (quels que soient les critères plus ou moins idiots qu'on est prêt à mettre en avant pour qualifier l'adjectif) deviendrait à la portée de n'importe qui, et pas seulement des plus « glamours » ou des plus « gentils ». L'amour parfait se démocratiserait ainsi considérablement au lieu de rester l'apanage de quelques privilégiés, et les relations entre humains « authentiques » en seraient allégées d'autant.

Questions

1. Quels avantages l'auteur voit-il dans la relation avec un robot ami par rapport à l'amitié humaine ?
2. Comment les robots pourraient-ils transformer notre conception de l'amitié et des relations affectives ?