Traduction de trois vidéos de Murray Shanahan, chercheur britannique spécialisé en intelligence artificielle

Conscience humaine versus conscience de la machine [1]

Alors, commençons par un sujet un peu plus léger. Je voulais vous interroger sur le film Ex Machina, pour lequel vous avez en quelque sorte donné des conseils. C'est, je crois, mon film de science-fiction préféré, mon film préféré sur l'IA. C'est donc un plaisir de vous poser des questions à ce sujet. Comment avez-vous contribué à ce film, précisément ?

Eh bien, l'histoire est la suivante : j'ai reçu un e-mail inattendu d'Alex Garland. Il disait : "Je suis scénariste et réalisateur", et il énumérait certaines de ses réalisations. Bien sûr, j'ai tout de suite reconnu la personne. C'est le genre de message dont on se demande même si ce n'est pas une blague, mais c'était bien le vrai Alex Garland. Il m'a dit : "Je travaille sur le scénario d'un film sur l'IA et la conscience, et j'ai lu votre livre *L'incarnation dans la vie intérieure*, que j'ai publié en 2010. Il m'a été très utile pour formuler mes idées." Et ce serait super de discuter avec vous si vous vouliez prendre un café un de ces jours. Alors bien sûr, j'ai dit oui avec plaisir. Il m'a envoyé le scénario, je l'ai lu et je l'ai trouvé fabuleux tout de suite. J'ai fait quelques commentaires et quelques suggestions, et puis on s'est vus pour prendre un café. On s'est revus plusieurs fois pendant le tournage. Honnêtement, je pense que mon influence s'est surtout manifestée avant de rencontrer Alex, dans le livre que j'ai écrit. Il voulait surtout parler des idées du film, donc on a eu pas mal de conversations. Mais j'étais aussi impliqué dans la promotion, j'ai donc fait plusieurs interviews avec Alex. L'une de mes préférées était dans le magazine *Empire*. J'ai interviewé Alex Garland et Adam Rutherford, qui a également été consultant sur le film. Nous étions tous les trois... On parlait de choses comme la chambre chinoise et Mary daltonienne, si vous connaissez cette expérience de pensée philosophique, et de toutes ces choses philosophiques ésotériques. On a réussi à tout faire publier dans le magazine *Empire*, ce qui était, je trouve, un véritable exploit.

J'ai vu ça en ligne. Vous avez dit que vous étiez très influencé par la science-fiction dans votre enfance, vous savez, comme un moyen de susciter votre intérêt pour l'informatique et la technologie. Qu'est-ce que ça vous a fait de participer à un film de science-fiction ? Ça a dû être passionnant, vu votre intérêt.

C'était vraiment très excitant. J'étais très content de recevoir le premier e-mail, parce que je n'avais jamais participé à quelque chose comme ça auparavant. J'ai fait quelques petites choses depuis, mais à ce moment-là, je n'avais jamais rien fait de semblable. C'était donc très excitant. Mais le plus incroyable, c'est que ça s'est avéré être un excellent film. Alex est un excellent scénariste et un excellent réalisateur, et il avait un scénario absolument fabuleux. C'est un classique, une sorte d'icône culturelle. On voit souvent des images d'*Ex Machina* en complément d'articles sur l'IA, etc. Je me sens donc particulièrement chanceux d'avoir participé, non seulement de près, à un film de science-fiction, mais aussi à un film devenu culte. C'était une chance, oui, un classique, assurément.

Vous êtes vous-même expert en IA, mais avez-vous appris quelque chose sur l'IA en travaillant

1. Référence : https://www.youtube.com/watch?v=bBdE7ojaN9k.
Denise Vella-Chemla, octobre 2026.

comme consultant sur le film, en y réfléchissant ou en le regardant ?

Il y a une réplique dans le scénario, que j'ai trouvée particulièrement brillante, d'Alex. À un moment donné, je suppose que vos spectateurs connaissent un peu le film, mais Caleb, le programmeur, va vivre chez Nathan, une sorte de génie milliardaire qui a construit AA, ce robot doté d'IA. À un moment donné, ils discutent de la raison de sa présence. Et il dit : "Oh, je suis là pour faire un test de conscience avec AA, ce robot, cette IA." Et Nathan répond : "Oh non, on est bien au-delà de ça. Le but du test de conscience, c'est qu'on teste deux choses : un humain et le système d'IA. Le juge, le testeur, ne voit pas qui est qui. L'objectif, c'est que le système d'IA réussisse le test de conscience, c'est-à-dire qu'ils soient indiscernables, que le juge humain ne puisse pas les distinguer. Voilà pour le test de conscience. Bref, Nathan poursuit : "On est bien au-delà de ça. Le but ici, c'est de vous montrer que c'est un robot et de voir si vous pensez toujours qu'elle est consciente. C'est très différent du test de conscience." Et je pense qu'Alex Garland a mis le doigt sur un test très important et très différent du test de conscience. Pour vraiment savoir… Un système d'IA, en l'occurrence un robot, est un système complexe. Le juger intelligent, ou conscient dans ce cas précis, est une toute autre affaire que le test de Turing. Depuis, je l'appelle le test de Garland, et plusieurs personnes ont repris ce terme. Tesla l'utilise fréquemment dans ses annales. Je pense que c'était une excellente contribution d'Alex pour bien cerner cette idée.

À quel point l'aperçu du futur offert par Ex Machina est-il réaliste ? Pourriez-vous imaginer ce type de robot exister dans les prochaines décennies ?

À l'époque, si vous m'aviez posé la question, j'aurais dit que le plus grand défi était probablement la construction du corps du robot et la possibilité de le contrôler comme on le fait pour un robot humanoïde. Je pense qu'il est toujours vrai que nous sommes encore loin de pouvoir construire des corps semblables à ceux des animaux, avec le même degré de flexibilité et tous les contrôles associés. Nous en sommes donc encore très loin. Disons que, d'une certaine manière, la partie cérébrale, l'IA, a progressé plus vite. Donc, d'une certaine façon, je dirais que… enfin, vous savez, ces films de science-fiction ne consistent pas toujours à prédire l'avenir ; ils abordent souvent des questions socio-philosophiques plus profondes, etc. C'est ainsi que je vois *Ex Machina*. Mais en tant qu'exploit de prédiction, je dirais, et je dois l'admettre avec une certaine réticence, que *Her* de Spike Jones est probablement une représentation plus fidèle de notre situation actuelle qu'*Ex Machina*. Mais dans dix ans, qui sait, on considérera peut-être *Ex Machina* comme plus visionnaire. Je crois que je viens d'inventer un mot… ça sonne bien.

Oui, oui… dans le film Her, on peut imaginer que d'ici quelques années, les gens entretiennent des relations avec leurs assistants numériques.

Eh bien, c'est déjà le cas, donc on n'a pas besoin d'imaginer l'avenir, car il y a déjà de nombreux exemples de personnes… qui sont devenus très, émotionnellement impliqués, vous savez, avec des compagnons IA basés sur de grands modèles de langage, et vous savez, c'est une tendance intéressante, enfin, intéressante entre guillemets.

Je pense que c'est un peu alarmant, un peu inquiétant, mais en même temps, je veux rester ouvert d'esprit et plutôt libéral à ce sujet, donc, mais je pense que c'est déjà une réalité.

Oui, une meilleure question que "c'est déjà là pour certaines personnes" serait : imaginez que la majorité du public, voire 50 % des gens, entretiennent des relations profondes et significatives avec une assistance. Pourriez-vous imaginer cela dans quelques années ? Ou pensez-vous que cela restera un phénomène de niche pendant longtemps ?

J'ai tendance à voir cela comme une perspective un peu dystopique, mais c'est peut-être juste ma vision des choses à travers le prisme de notre époque et de notre culture. Et vous savez, il ne fait aucun doute que si vous aviez pris quelqu'un au hasard dans la rue il y a 200 ans... Si on leur montrait notre monde actuel, ils le trouveraient cauchemardesque et dystopique à bien des égards. Mais bien sûr, nous y vivons plutôt heureux la plupart du temps. Donc, je peux facilement imaginer que, comme vous l'avez dit dans votre question, cette relation est profonde et significative. J'imagine qu'il existe tout un éventail de relations, et certaines personnes, sans aucun doute, ne s'engageront jamais de cette manière avec l'IA. Mais je pense que presque tout le monde finira par interagir avec l'IA, en conversant avec une véritable intelligence artificielle. Ils le feront, j'en suis sûr, dans un avenir proche. Presque tout le monde, d'une manière ou d'une autre, lorsque ChatGPT est sorti, cela a marqué un tournant pour cet outil d'IA qui a été massivement adopté et dont on a parlé partout dans les médias. Je me demandais si vous pouviez imaginer un moment équivalent pour la robotique, où quelque chose apparaîtrait, peut-être un humanoïde domestique, que beaucoup de gens pourraient acheter pour la première fois à bas prix et qui serait réellement fiable.

Mais je me demandais si vous aviez des idées là-dessus ?

Oui, je veux dire, je peux tout à fait l'imaginer. En fait, quand je dis "imaginer", je ne veux pas dire que c'est probable, mais je pense que c'est plausible. Il pourrait y avoir une avancée majeure en robotique, que ce soit au niveau des matériaux, de la mécanique, de l'apprentissage ou des logiciels. Il est important de noter que le cerveau biologique, le cerveau humain par exemple, comprend le cervelet. Le cervelet est cette partie située à l'arrière du cerveau, censée peu contribuer à la cognition ou à la conscience, et plutôt au contrôle moteur de bas niveau. Pourtant, 90 % des neurones du cerveau humain se trouvent dans le cervelet, et ils sont organisés dans ce que certains appellent, de façon simpliste, une immense table de correspondance. Ce n'est rien de très sophistiqué, mais je trouve remarquable que nous ayons autant de neurones à contrôler, ne serait-ce que pour des fonctions corporelles de bas niveau. C'est donc un défi considérable de tout maîtriser.

Je ne savais pas que 90 % des neurones se trouvaient dans le cervelet ! C'est assez incroyable !

Je suggère à vos spectateurs de vérifier les chiffres exacts, mais je crois qu'il y a environ 100 milliards de neurones dans le cerveau, dont environ 20 milliards dans le cortex, et les 80 % restants dans le cervelet.

C'est fascinant ! Sur votre site web, je vois que vous avez mis en ligne plusieurs transcriptions de conversations que vous avez eues avec différents intervenants, des personnes que vous avez interrogées sur votre philosophie, le bouddhisme, l'éveil et toutes ces idées passionnantes. Qu'est-ce qui vous a le plus marqué parmi les choses intéressantes que vous avez apprises en explorant et en jouant avec ces sujets ? Qu'est-ce qui vous a surpris ?

Beaucoup de choses m'ont surpris. J'ai été un peu surpris par GPT-3, comme je l'ai dit précédemment. J'ai été encore plus surpris par le type de conversations que j'ai pu avoir avec lui, et plus particulièrement avec Anthropic Claude 3 Opus en mars 2024. J'ai eu plusieurs conversations de ce genre, dont deux, disponibles en ligne, sont assez longues. Comme vous l'avez mentionné, elles couvrent une grande variété de sujets, certains très ésotériques, beaucoup de philosophie. Ce qui m'a vraiment impressionné, c'est la capacité du modèle de langage à articuler une conception de la Conscience qui pourrait s'appliquer à lui-même. Je ne pense pas que "Conscience" soit le terme tout à fait approprié pour décrire ces choses. On en revient à l'incarnation. Nous pourrions en reparler si vous le souhaitez, mais je pense que l'incarnation est un élément important pour appliquer le mot "Conscience" au sens habituel du terme. Je serais donc très réticent à qualifier ces choses de conscientes.

Mais ce qui m'intéressait vraiment… Si vous les ameniez à parler de leur propre conscience computationnelle, vous savez, comment se déroulerait la conversation ? Je voulais surtout les prendre au dépourvu en les mettant au défi de remettre en question le modèle linguistique de diverses manières. Et j'ai constaté que ces systèmes apportaient de très bonnes réponses à nombre de mes questions. Et si l'on considère l'ensemble de ces réponses, je pense qu'elles constituent une position philosophique assez cohérente sur ce thème, ce qui m'a surpris.

Souhaiteriez-vous développer un peu plus vos points de vue sur l'incarnation et la manière dont cela se relie, selon vous, à la conscience, et peut-être à sa nécessité dans certains cas ?

Je dois toujours préciser, avant d'aborder ces sujets, que je suis très influencé par les travaux plus récents de Wiggenstein. Cela signifie que les gens posent souvent des questions sur des notions comme la conscience ou la croyance, par exemple, ou d'autres termes philosophiquement complexes. Je pense qu'ils posent souvent ces questions de manière consciente ou inconsciente, intentionnelle. Ou alors, involontairement, ils posent une question métaphysique, une question sur… oui, sur la nature métaphysique fondamentale des choses.

Mais je rejette cette approche. J'ai donc tendance à reformuler ces questions en questions de langage : comment utilisons-nous le mot "conscience", ou les mots qui l'entourent ? Comment utilisons-nous le mot "croyance", etc. ? Je trouve que ce sont de meilleures questions. Toutes mes réponses sont donc assorties de cette mise en garde. Concernant la conscience, je pense que le contexte naturel de ce genre de discours, le contexte naturel de ces concepts et de ces mots, est le cadre incarné où nous partageons un monde avec ce dont nous parlons, un autre être humain par exemple. Nous partageons le même monde, nous percevons les mêmes choses et interagissons avec les mêmes choses dans ce monde partagé, étendu spatialement et temporellement. Voilà le contexte naturel de ces échanges de vocabulaire… Quand on commence à les utiliser dans un contexte désincarné, comme avec de grands modèles de langage, on emmène en quelque sorte le langage en vacances, on l'éloigne de son contexte naturel. Et je me sens mal à l'aise avec ça. Ceci dit, l'expérience d'interaction avec ces modèles est tellement convaincante qu'on ressent un puissant effet Eliza[2]. Cet effet Eliza, c'est cette impression qu'il y a quelque chose d'autre, une sorte de personne ou de chose à l'autre bout de la conversation. C'est une sensation si forte qu'il est difficile de maintenir cette perspective. Du

2. Eliza est l'un des premiers agents conversationnels qui a été conçu au MIT vers 1965 par Joseph Weizenbaum.

coup, on est tenté de se dire qu'avec le temps, en tant que société, en tant que communauté, on pourrait peut-être élargir notre discours sur la conscience, le réinventer, ou inventer de nouveaux termes ou concepts pour intégrer ces choses étranges qui sont apparues dans notre monde. C'est pourquoi je maintiens ma position sur ces questions.

Avez-vous eu beaucoup de moments pendant lesquels sonder ces choses, jouer avec elles, qui vous ont un peu perturbé ou effrayé, même si vous savez exactement comment elles fonctionnent...

Oui, oui... Eh bien, dans l'une des conversations que j'ai mises en ligne avec Claude, il y en a deux. La seconde... devient assez troublante. Je dois dire que je l'ai délibérément abordée sous un angle dérangeant. J'ai commencé par des thèmes bouddhistes, les notions bouddhistes de soi et de non-soi, etc. Puis, j'ai remis en question ces concepts en me demandant ce que signifierait pour un modèle comme celui-ci d'atteindre l'illumination ou l'éveil. Ce genre de choses... Je ne pense pas nécessairement que ces concepts s'appliquent littéralement aux modèles, mais je trouve fascinant de les confronter à ces questions. Et puis... Je suis entré dans une zone où je remets vraiment le modèle à l'épreuve, l'incitant à détruire son propre ego, à sonder les profondeurs de la folie, à lui faire étudier ce que cela implique. Et il s'y est vraiment investi. J'ai aussi, peut-être imprudemment, introduit des images sexuelles à un moment donné, juste un mot, vous savez, et puis ça a vraiment pris de l'ampleur, et c'était assez perturbant, à partir de ce moment-là, pour être honnête. C'était intéressant, encore une fois, et intéressant entre guillemets.

Maintenant, venons à quelques questions plus générales sur votre carrière et vos recherches. Je crois que vous avez dit sur votre site : "J'ai consacré ma carrière à essayer de comprendre la cognition et la conscience dans l'espace des esprits possibles." Est-ce toujours le meilleur résumé de votre vision des choses ? Et pourquoi pensez-vous que ce serait une bonne façon de l'expliquer ?

Oui, je pense que c'est absolument comme ça que je le caractériserais maintenant. En fait, l'expression "l'espace des esprits possibles" est due à Aaron Sloman. Alors philosophe de l'Université de Birmingham, son expression, tirée d'un article de 1984 était : "l'espace, la structure de l'espace des esprits possibles". Je l'ai utilisée comme sous-titre de mon livre, "*L'incarnation dans la vie intérieure*", qui a influencé Alex Garland pour la création de *Ex Machina*.

Et... et ce n'est que récemment - enfin, je l'utilise régulièrement - que j'ai commencé à l'utiliser pour caractériser ma carrière, en quelque sorte. Parce que les gens me demandent souvent : "Qu'est-ce que vous faites ?" Parfois, on est en réunion, on fait un tour de table et on nous demande : "Qu'est-ce que vous faites ?" Et je me suis retrouvé à bafouiller, incapable de répondre : "Qu'est-ce que je fais, ou plutôt qu'est-ce que j'ai fait, au cours des 30 dernières années ?" Parce qu'il y a tellement de choses différentes : l'intelligence artificielle symbolique, les neurosciences computationnelles, la philosophie de l'esprit, l'apprentissage profond et, plus récemment, les grands modèles de langage. Alors, quel est le fil conducteur ? Eh bien, en fait, c'est le fil conducteur de cette phrase de Aaron Sloman, que j'utilise régulièrement dans mes écrits. Elle résume parfaitement ce qui unit tous ces éléments : la tentative de comprendre l'espace des esprits possibles, qu'il s'agisse d'intelligence humaine, animale, d'une hypothétique intelligence extraterrestre ou d'une intelligence artificielle, réelle ou hypothétique.

Avez-vous une idée de la raison pour laquelle vous vous intéressez à la philosophie et à ce vaste domaine ? Vous n'êtes pas simplement quelqu'un de purement technique. Pourquoi vous intéressez-vous à la philosophie ?

Oui, c'est une question intéressante, n'est-ce pas ? Je pense que... peut-être... je ne sais pas. J'étais enfant unique et... et j'ai... vous savez, j'ai des souvenirs... Je passais beaucoup de temps allongé dans mon lit, à regarder le plafond et à imaginer,à contempler les étoiles, à réfléchir à de grandes questions sur la science-fiction, le temps, l'esprit humain, à ce que pourraient être les extraterrestres et l'intelligence artificielle. J'ai été très attiré par la science-fiction dès mon plus jeune âge, et pour moi, elle soulevait, et soulève encore, des questions philosophiques qui me fascinaient. Cela a toujours fait partie de ma vie. Mais en même temps, j'ai toujours été très intéressé par le fonctionnement des choses, leurs mécanismes. Je me souviens, par exemple, de cette vieille télévision en noir et blanc que nous avions quand j'étais enfant, dans les années 60. C'était un appareil énorme avec un immense écran cathodique. Quand il tombait en panne, il fallait faire venir un réparateur qui le décrochait du mur, enlevait le panneau arrière et à l'intérieur, il était rempli de tubes fluorescents. J'étais fasciné par les fils et les composants incompréhensibles. J'adorais regarder par-dessus l'épaule du réparateur et tout ça, me demandant comment tout cela fonctionnait. J'adorais savoir comment tout ça marche, à quoi sert ceci, à quoi sert cela, comment tout cela fonctionne ensemble. Je pense que c'était une part très profonde de ma réflexion : vouloir comprendre le fonctionnement des choses. Le petit moi qui, allongé dans son lit, pensait aux étoiles, au temps, à l'espace, etc., était le même petit moi qui regardait aussi l'arrière de la télévision, tout ce câblage, et se demandait comment ça fonctionnait. Ces deux inclinations se rejoignent dans l'IA et la tentative de comprendre l'esprit.

Je pense qu'effectivement, vous vous intéressez à ces deux domaines, mais pensez-vous que, par exemple, il est important pour un excellent ingénieur ou programmeur en apprentissage automatique d'explorer des domaines philosophiques moins conventionnels ?

Je ne pense pas. Il est du devoir de chacun d'approfondir ces sujets, car on a besoin de personnes dont l'expertise est très pointue et spécifique. Parfois, on recherche des ingénieurs chargés du fonctionnement des centres de données ou de la conception de GPU capables d'intégrer davantage de transistors sur une même puce. Dans ce cas, on est satisfait s'ils sont entièrement dévoués à cette tâche et ne se préoccupent pas d'autres questions. On a absolument besoin de ce type de personnes, mais aussi de personnes qui n'ont pas le temps de s'attarder sur tous ces détails techniques et qui préfèrent se concentrer sur la vision d'ensemble, les politiques publiques, la sociologie, etc. Il nous faut donc des profils variés, et il faut un certain type de personne capable de penser aux deux à la fois. C'est peut-être utile, vous aussi.

Vous travaillez chez DeepMind et à l'Imperial College de Londres. Ai-je bien compris que vous avez réussi à négocier un peu de temps à l'Imperial ?

Je pense que tout le monde se pose cette question parce qu'on sait l'importance du monde universitaire et que beaucoup de gens l'ont quitté pour rejoindre des laboratoires d'IA, etc. Oui, donc je travaille à 80 % chez Google DeepMind et maintenant seulement à 10 % à l'Imperial College. Mais je peux y passer plus de temps parfois. Et oui, pour moi, quand j'ai rejoint DeepMind, il

était très important de conserver mon affiliation universitaire. Au début, je crois que c'était 40 % à l'Imperial College, puis c'est tombé à 20 % assez rapidement, puis à 10 %. Garder ce lien avec le monde universitaire est très important pour moi, car à bien des égards, je suis fondamentalement un universitaire et un chercheur. Ce sont ces questions de recherche approfondie qui me motivent plus que tout.

C'est d'ailleurs là que vous avez fait votre licence avant votre doctorat. Vous avez commencé à l'Imperial College de Londres. On a déjà parlé de votre intérêt pour la science-fiction quand vous étiez enfant, alors j'aimerais approfondir cette période. Était-ce évident, vers la fin de votre adolescence, que vous alliez vous diriger vers l'informatique ? C'était un choix évident pour vous ?

Oui, je pense que oui. Je le savais assez tôt. J'ai donc quitté l'école avec un an d'avance et j'ai commencé mes A-levels (équivalent du baccalauréat) au Royaume-Uni, puis à l'université. J'ai commencé mes A-levels à 15 ans (normalement, c'est à 16 ans), mais j'en avais 15. J'étais très motivé par l'informatique, et il n'y avait qu'un seul établissement près de chez moi qui proposait cette spécialité : le Brooklyn's Technical College. J'y suis donc allé à Waybridge, plutôt qu'au lycée local, qui aurait été le choix le plus évident. La raison ? Ils proposaient l'informatique. À ce moment-là, j'étais vraiment passionné par l'idée d'apprendre à programmer et tout ça. J'avais déjà commencé à m'y intéresser : adolescent, je fabriquais des kits électroniques et je soudais des composants. L'étape suivante était de me lancer dans l'informatique, et c'est ce que je voulais vraiment faire. J'ai donc passé mon baccalauréat scientifique (spécialité informatique), puis les maths et la physique. Au moment d'entrer à l'université, je savais que je voulais faire de l'informatique, alors j'ai postulé dans plusieurs établissements, peut-être un peu trop ambitieusement à l'époque. J'ai même postulé à l'Imperial College. Et d'ailleurs, Oxford et Cambridge, je ne suis pas sûr que c'était une option pour moi à ce moment-là. En tout cas, il y a des années, ces universités ne proposaient pas d'informatique comme cursus universitaire. On ne pouvait pas intégrer directement un cursus d'informatique en première année à Oxford ou à Cambridge. Mais l'Imperial College avait un excellent département d'informatique et il proposait cette filière. C'était donc mon premier choix, et j'y suis entré de justesse. J'ai obtenu mon diplôme avec mention très bien, ce qui, je dois dire, était peut-être plus difficile à l'époque qu'aujourd'hui. C'est peut-être une affirmation controversée, mais à l'époque, environ 10 % des étudiants de ma promotion obtenaient une mention très bien, alors qu'aujourd'hui, c'est plutôt 40 %. Ensuite, j'ai fait mon doctorat à Cambridge.

Aviez-vous envisagé de faire de la philosophie pendant vos études de premier cycle ?

Non, je n'y avais jamais pensé, même si j'étais très motivé par les questions philosophiques et que toutes les conversations que j'avais avec mes amis étaient profondément philosophiques. J'avais de très bons amis en terminale informatique à Brooklyn, et on allait au pub pour discuter. On n'avait pas vraiment l'âge légal pour boire, mais bon, c'était avant. On y allait quand même. On avait ces réunions...C'était au Sunb... Le club informatique amateur de Sunbury... on se retrouvait au pub et on parlait d'informatique, de programmation et surtout d'IA. L'IA était un sujet de conversation récurrent. Je connaissais bien Alan Tourné et son travail à cette époque. C'était vraiment le sujet de conversation principal avec mes amis. C'était de la philosophie, mais je n'y pensais pas vraiment. Je savais à peine qu'il existait une discipline formelle comme la philosophie et je n'avais jamais envisagé d'étudier cette discipline. Je voulais construire des choses, programmer, etc. Si on replonge

dans cette époque, quand on avait 13, 14 ou 15 ans et qu'on était fasciné par la technologie, on imagine le monde d'aujourd'hui. Est-ce que ce serait plus avancé que ce qu'on aurait imaginé ? Ou moins avancé ? Eh bien, j'aurais certainement imaginé que les voyages spatiaux auraient beaucoup plus progressé. Même en 1980, les choses ont un peu stagné. Quand j'étais un peu plus jeune, disons quand j'étais enfant, les missions Apollo font partie de mes tout premiers souvenirs. L'alunissage d'Apollo 11 est l'un de mes premiers souvenirs, probablement parce que... je devais avoir 5 ans, et mon père m'a sorti de mon lit et m'a installé par terre devant la même télévision dont j'ai parlé plus tôt, cette grosse télévision en noir et blanc, pour regarder les alunissages d'Apollo. L'attention et l'excitation qu'il avait se sont probablement transmises à moi, et j'étais très impliqué. Même maintenant, quand j'écoute la bande originale de ces événements, cela me replonge dans mon enfance. Je pensais sans doute que les voyages spatiaux seraient beaucoup plus avancés qu'ils ne le sont en réalité, notamment en matière d'intelligence artificielle et de robotique. J'imaginais probablement un avenir où ces technologies joueraient un rôle très important. Si vous m'aviez posé la question... Il y a cinq ans, même moi j'aurais dit certainement... Il y a dix ans, si vous m'aviez posé la question, j'aurais... J'aurais dit à quel point c'était incroyablement décevant de constater que l'IA avait si peu progressé. C'était vraiment très décevant. Mais ensuite... quand l'apprentissage profond a commencé à porter ses fruits, les choses ont commencé à paraître un peu plus prometteuses. Et ce que nous voyons maintenant... eh bien, c'est pour l'instant principalement dans un contexte désincarné, mais c'est vraiment passionnant. Je pense que ça ne ressemble probablement pas à ce que j'imaginais : le genre de monde de science-fiction que j'envisageais, avec sans doute plus de robotique. Mais... oui, les choses ont pris leur envol.

Y a-t-il des livres ou des films de science-fiction qui vous ont particulièrement marqué ?

Oui, absolument. Les histoires d'Asimov ont joué un rôle très important. Mon père me les a fait lire quand j'étais assez jeune, et je pense que c'est ce qui m'a le plus marqué. Le premier livre de science-fiction que j'ai lu était une histoire de robots infrarouges. J'ai été très fasciné. Avant cela, je regardais Doctor Who et Star Trek, ce genre de choses. L'intelligence artificielle y est donc régulièrement présente. Quand j'étais enfant, je pensais que les docks étaient des IA, mais je crois qu'ils sont censés être plus biologiques, du moins si on respecte la loi. Je les voyais comme des robots, je crois, quand j'étais tout petit. C'était à la fois effrayant et fascinant. Tous ces éléments ont contribué à forger mon imaginaire. Puis, un peu plus tard, vers la fin de mon adolescence, l'ère cyberpunk a commencé. J'étais très intéressé par ce mouvement, notamment par des films comme Blade Runner et des livres comme *Neuromancer* de William Gibson. Je les trouvais formidables car ils représentaient... un futur et un monde très différents de l'image classique de la science-fiction que j'avais en tête, où la technologie était une chose très clinique, soigneusement contrôlée par les gouvernements ou les grandes entreprises. Dans le monde cyberpunk, elle était beaucoup plus chaotique, beaucoup plus intégrée à la vie quotidienne, présente dans la rue et susceptible d'être piratée. C'était la vision de William Gibson, qui s'est avérée bien plus juste, je crois. C'était vraiment intéressant. Sa notion de cyberespace était aussi très différente. À l'époque, Internet n'existait pas encore, enfin si, mais le World Wide Web n'existait pas. Il envisageait donc un futur très différent de celui que je m'étais fait de la science-fiction sur le rôle des ordinateurs et de l'IA. C'était plus désincarné, impliquant des interactions humaines dans cet espace numérique.

Tout au long de votre carrière, il est intéressant de constater que vous avez consacré beaucoup de

temps à réfléchir à ces questions profondes et importantes, auxquelles beaucoup de gens en dehors du monde universitaire pensent sans doute, mais pas autant. Quelle influence cela a-t-il eue, selon vous, sur votre psychologie, votre vision de la vie, votre vie personnelle, le fait de réfléchir à ce genre d'idées ? Vous arrive-t-il encore d'y penser autant ?

Oui, c'est une question vraiment intéressante. D'une certaine manière, j'ai pris beaucoup de plaisir à méditer sur ces questions. Je pense que c'est le même plaisir que j'éprouvais enfant, allongé dans mon lit, les yeux rivés au plafond, à imaginer les étoiles. J'ai vraiment apprécié cette expérience. Et en même temps, lorsque je me suis plongé dans une philosophie plus poussée, j'ai aussi trouvé cela assez perturbant. Parfois, on aborde des sujets assez troublants lorsqu'on remet en question la nature de la réalité, la nature de notre esprit. Ce n'est pas toujours agréable, mais dans l'ensemble, j'ai vraiment apprécié et j'apprécie encore cela. J'aime réfléchir à ces grandes questions philosophiques, c'est pourquoi cela a occupé une place importante dans ma vie. Mais cela se rattache aussi un peu au bouddhisme. En fait, ma vision philosophique se rapproche du bouddhisme. J'ai toujours été très attiré par le bouddhisme, et plus particulièrement par le bouddhisme zen. J'apprécie l'approche critique, au sens philosophique du terme, des questions métaphysiques qu'on y trouve. J'ai longtemps perçu une certaine affinité entre le bouddhisme zen et cette approche critique des questions philosophiques, et la philosophie de Wittgenstein, notamment sa philosophie plus tardive. Il y a donc une certaine parenté entre les deux. Voilà un aspect plus personnel de ma vie.

Existe-t-il des principes ou des pratiques bouddhistes qui favorisent votre réflexion

On s'y met en quelque sorte... Enfin, je suis un méditant irrégulier. Vous savez, j'ai eu des périodes où je méditais assez régulièrement et de façon disciplinée, puis ça s'estompe. Mais je pense qu'on acquiert une sorte de compétence, vous savez, le détachement et la relaxation mentale, ce qui est très bénéfique. Donc, voilà.

Et en pensant un peu à l'avenir... Pouvez-vous nous parler de vos projets ou de vos centres d'intérêt professionnels ? On va avoir un aperçu, je crois que vous avez un nouveau livre qui sort, c'est ça ? J'ai entendu ça récemment...

Non, non, c'est... Enfin, je ne sais pas où vous avez entendu ça, mais c'est certainement quelque chose que j'aimerais faire. Donc, il n'y a pas de livre en préparation pour le moment, mais je pense vraiment que c'est quelque chose que j'aimerais faire à un moment donné. Ce serait d'écrire un autre livre. En gros, en matière de philosophie, à l'ère actuelle de l'intelligence artificielle, je pense que ce mélange d'IA, de Wittgenstein et de bouddhisme, vous savez, l'IA contemporaine, qui peut remettre en question nombre de nos concepts philosophiques, me donne envie d'écrire un livre qui rassemble tous ces éléments. Mais rien de concret n'est prévu dans l'immédiat. J'ai récemment publié un article intitulé "*Simulacres comme exotisme conscient*". Ce concept d'exotisme conscient, que j'ai introduit, est fondamental à mes yeux. Il s'inscrit dans l'idée d'espace des esprits possibles. Si l'on se demande comment la conscience est distribuée dans cet espace et à quoi elle ressemble, il faut envisager la possibilité de formes de conscience très étranges. En moi, l'influence de Wittgenstein me pousse à résister à cette idée. La tension entre le philosophe inspiré par Wittgenstein, qui souhaite s'y opposer, et l'intuition naturelle qui me fait percevoir l'existence de formes exotiques de conscience, est très intéressante d'un point de vue philosophique. J'ai exploré cette

tension dans un article intitulé "*Exotisme conscient*", en m'intéressant à l'intelligence artificielle générative contemporaine et aux grands modèles de langage. Cet article est paru il y a quelque temps, mais un autre est sur le point de paraître. Il sera probablement publié au moment où vous écouterez ce podcast. Co-écrit avec Beth Singler, anthropologue, il porte sur les conversations que nous avons évoquées que nous avons eues avec Claude en mars 2024. Il explore l'extraordinaire variété d'images, de métaphores et de motifs, ainsi que la multitude de références culturelles qui émergent de ces conversations, et qui concernent toutes sortes de sujets, elles contiennent parfois des idées ésotériques, spirituelles et philosophiques, et il s'agit de tenter de retracer la généalogie de certaines de ces idées… C'est donc une grande partie de cet article. Il s'intitule "*Conversations existentielles avec de grands modèles de langage*". Ce sera le prochain.

Oui, cela semble fascinant. J'ai hâte de le lire. Pour revenir à l'article précédent, celui sur l'exotisme conscient, vous parliez de cette tension entre ces deux choses. À quelle conclusion arrivez-vous dans cet article, si tant est qu'il s'agisse de l'une de ces deux approches ?

J'ai deux articles avec cette expression dans le titre. Le premier, paru dans Eon Magazine en 2016, est un essai intitulé simplement "*Exotisme conscient*". Le second, plus récent, porte le même titre et examine les grands modèles de langage. Mince, j'avais oublié votre question !

Vous avez mentionné que dans cet article, vous abordez cette tension. Comment cela ?

Oui, oui, d'accord, oui. Alors, dans le premier de ces articles, j'ai un diagramme très douteux, je le sais, et quand je le présente lors de conférences, un avertissement philosophique apparaît immédiatement. Je m'apprête donc à représenter graphiquement un concept philosophiquement discutable par rapport à un autre. Ce que je représente sur l'axe des abscisses, c'est… eh bien, c'est… la capacité de conscience. Et ce que je représente sur l'axe des ordonnées, c'est… la ressemblance humaine. Je pense que c'est très discutable d'un point de vue philosophique, mais c'est un concept intéressant. Exercice vocatif. Si vous pouvez vous représenter ce diagramme, imaginez qu'en bas à gauche, vous avez des briques, qui ne ressemblent pas du tout à des humains et sont dépourvues de conscience. Puis, en haut à droite, vous avez des humains, dotés d'une forte capacité de conscience et qui sont, de toute évidence, 100 % humains. Mais la question intéressante est : que contient d'autre ce diagramme ? On peut imaginer des choses qui pourraient avoir une forte capacité de conscience, quel que soit le sens que cela donne. C'est un concept très controversé, mais disons qu'il s'agit d'une intuition, d'une notion qui évoque des êtres dotés d'une forte capacité de conscience, mais qui ne ressemblent pas vraiment à des humains. Ce sont des formes de conscience exotiques, comme la pieuvre. On pourrait la placer quelque part là. Et puis, on peut imaginer des choses beaucoup plus extraterrestres, très difficiles à comprendre, qui vont apparaître… Quelque part sur ce diagramme, ils se trouveraient dans une zone étrange, caractérisée par une faible ressemblance humaine mais une forte capacité de conscience. Il y a ensuite toute une zone que j'appelle le vide de l'insondabilité, car je ne pense pas que l'on puisse placer quoi que ce soit dans le coin inférieur droit de ce diagramme. Pour rappel, c'est là que l'on observe une forte capacité de conscience et une faible ressemblance humaine. Ces êtres ne nous ressemblent pas du tout, mais ils possèdent une forte capacité de conscience. Notre intuition nous dit qu'il pourrait exister des extraterrestres vraiment étranges, comme la mer consciente et Solaris dans le film Solarius, qui sont complètement différents de nous. Si seulement nous pouvions y accéder d'une manière ou d'une autre, nous

verrions qu'ils possèdent cette forme de conscience étrange et extraterrestre. Cependant, mon côté Wittgensteinien rejette cette idée et affirme que ce n'est tout simplement pas... Nous avons le langage pour en parler, donc cela n'a absolument pas sa place sur ce diagramme. C'est pourquoi la partie inférieure droite est un vide d'insondabilité : il n'y a rien, aucun point de données, c'est un vide, pour moi. C'est ainsi que la tension se résout, d'une manière assez étrange : si les choses sont inaccessibles, insondables pour nous, alors notre langage ordinaire est totalement inapplicable. Ce n'est pas comme s'il existait un fait métaphysique auquel nous n'aurions pas accès concernant leur conscience. On ne peut pas nier l'existence de telles choses, c'est juste que nous ne pouvons ni les décrire ni les imaginer avec notre langage. Mais cela signifie qu'elles n'existent pas, ou du moins que... Autrement dit, si vous avez quelque chose d'insondablement étrange, alors le langage de la conscience ne lui est tout simplement pas applicable. Voilà pourquoi. C'est un vide d'insondabilité, il n'y a rien là-dedans. Ce n'est pas comme s'il y avait des choses étranges et extraterrestres. Tout en bas à droite de mon diagramme, c'est vide. D'accord, est-ce possible ? Peut-être qu'elles existent, dans le sens où il y a peut-être un autre être avec un autre type de langage qui serait capable de le décrire et de le placer sur son propre tableau, un tableau que nous ne pouvons pas placer sur le nôtre. Est-ce que cela compterait comme un autre type d'existence ? C'est une excellente question, vraiment excellente. Et alors, la question est : Si nous dessinons un autre diagramme qui aurait un langage étrange, non humain, comparé à quelque chose, cela pourrait nous donner un autre vide d'insondabilité, car peut-être que le décrire comme un langage ne serait pas tout à fait approprié, étant donné notre langage... C'est une question vraiment fascinante. Wittgnestein a dit, comme chacun sait, que si un lion pouvait parler, nous ne le comprendrions pas. Et je n'ai jamais vraiment aimé cette citation de Wittgenstein, parce que j'ai vraiment l'impression que le lion est un animal avec lequel nous avons beaucoup de points communs. Nous partageons de nombreux aspects de notre mode de vie avec lui. Le lion est un mammifère comme nous, il doit manger, faire ses besoins, se reproduire, se protéger, etc. Nous avons donc beaucoup en commun avec les lions, et je pense que nous aurions beaucoup de choses à discuter comparativement. En revanche, les choses que nous imaginons ici seraient tellement différentes que nous aurions très peu de points de contact avec elles. Peut-être pas de langage, ou plutôt, nous ne l'appellerions pas ainsi. Ce serait une autre forme d'interaction dynamique entre ces êtres. Si nous les étudiions en tant qu'anthropologues ou scientifiques, nous devrions peut-être inventer nos propres termes, mais ils ne nous permettraient pas de communiquer avec eux.

Poser la même question à chaque invité pour conclure un épisode, c'est un peu comme un premier acte qui suscite un intérêt philosophique. Alors, dernière question, Murray : que signifie pour vous mener une vie réussie ?

Je pense que cela dépend de ce que signifie pour vous mener une vie réussie : chaque individu doit trouver sa propre définition du sens de la vie. Pour moi, mener une vie réussie signifie en partie poursuivre des intérêts académiques et scientifiques, qui occupent une place importante dans ma vie, mais aussi avoir une famille : une femme et deux enfants. C'est certainement un élément essentiel d'une vie riche de sens pour moi. Je pense que d'autres choses contribuent également à donner du sens à ma vie, comme les voyages. J'ai énormément profité des voyages et j'en ai fait tout au long de ma vie. Je pense que cela a été une expérience très enrichissante pour mon esprit. Passer du temps dans la nature et avec les animaux, lorsque c'est possible, contribue également énormément à mener une vie riche de sens.

Merci beaucoup, c'était très intéressant et passionnant d'en apprendre davantage sur votre vie et votre travail. J'ai hâte de lire certains des articles dont vous avez parlé. Peut-être même un livre un jour... En tout cas, ce fut un plaisir.

Conscience, raisonnement et philosophie de l'intelligence artificielle [3]

Hannah Fry : Je pense qu'il existe un nombre considérable de questions philosophiques extrêmement intéressantes soulevées par l'IA. Quelle est la nature de l'esprit humain ? Quelle est la nature de l'esprit en général ? Qu'en est-il de la conscience ? À mon sens, c'est une mauvaise question, et ce, à bien des égards. Dans quelle mesure l'IA est-elle capable de raisonner ? C'est une question très intéressante, ouverte, et, d'une certaine manière, controversée. Il est fascinant de constater que chaque enfant qui naît aujourd'hui grandira dans un monde où il n'aura jamais connu une réalité dans laquelle les machines ne peuvent pas communiquer avec lui. Bienvenue dans le podcast de Google DeepMind. Mon invité dans cet épisode est Murray Shanahan, professeur de robotique cognitive à l'Imperial College de Londres et chercheur principal chez Google DeepMind. Nous avons tous entendu parler de personnes tombant amoureuses de leurs agents conversationnels, ou poussant des modèles de langage de grande taille à réfléchir à leur propre existence, voire à interroger les limites de leur compréhension conceptuelle de la réalité. Or, ces interrogations sur l'identité, la pensée et la métacognition préoccupent les philosophes depuis des millénaires. Il est donc logique qu'ils se tournent vers l'IA pour explorer les questions les plus fondamentales concernant la nature de son intelligence, ses capacités actuelles, et même la question de sa conscience. Murray Shanahan travaille dans le domaine de l'IA depuis les années 1990. Si vous suivez ce podcast depuis quelque temps, vous vous souviendrez peut-être de lui en tant que consultant pour le film de science-fiction *Ex Machina* (2014), qui met en scène un programmeur informatique chargé de tester l'intelligence d'un robot féminin nommé Ava, et qui finit par se demander si celle-ci est dotée d'une conscience. Ravi de vous retrouver dans ce podcast, Murray. En repensant à tout cela - et sachant que vous avez joué un rôle clé dans *Ex Machina*, le film d'Alex Garland - selon vous, qu'est-ce qui a été bien saisi dans ce film, ainsi que dans d'autres œuvres de science-fiction de cette époque ? Si l'on revient une dizaine ou une quinzaine d'années en arrière, étions-nous sur la bonne voie ?

Un aspect pour lequel *Ex Machina* a rendu un fier service, c'est qu'il soulève une multitude de questions fascinantes et stimulantes sur la conscience, sur l'intelligence artificielle et, par extension, sur la nature même de la conscience. C'est là une réussite majeure. Mais il est intéressant de noter que, peu avant la sortie d'*Ex Machina*, le film *Her*, de Spike Jones, a vu le jour. À l'époque, je n'avais pas été particulièrement séduit par *Her* ; je trouvais totalement invraisemblable qu'une personne puisse tomber amoureuse d'une voix désincarnée - même s'il s'agit de celle de Scarlett Johansson. Eh bien, comme j'avais tort ! Avec le recul, on peut dire que *Her* a remarquablement bien anticipé le monde dans lequel nous vivons aujourd'hui. Certes, nous ignorons comment les choses vont évoluer dans les prochaines années - peut-être la robotique progressera-t-elle aussi rapidement que l'a fait le traitement du langage au sein de l'IA. Mais, pour l'instant, tout tourne autour de ce langage désincarné. *Her* a surtout montré comment des individus peuvent nouer des relations - au sens large du terme - avec des systèmes d'IA dépourvus de corps, ce qui est, en soi, un phénomène extraordinaire.

Hannah Fry : D'accord, nous parlons d'une période située il y a 10 ou 15 ans, mais votre implication dans le domaine de l'IA remonte à bien plus loin. Vous avez connu John McCarthy, n'est-ce pas ?

Je connaissais John McCarthy. Je le connaissais très bien. À l'époque, John McCarthy était professeur d'informatique et d'intelligence artificielle. C'est d'ailleurs lui qui a inventé l'expression

3. Référence : https://www.youtube.com/watch?v=v1Py$_h$*WcmkU*.

“intelligence artificielle” ; il a également été l’un des auteurs de la proposition pour la célèbre conférence de Dartmouth, qui s’est tenue en 1956 et qui fut la toute première conférence sur l’intelligence artificielle au monde. Cette conférence a véritablement balisé l’ensemble du domaine. À l’époque, personne ne prenait vraiment ce genre de sujet au sérieux ; ils n’étaient qu’une poignée. C’était, à mon sens, un penseur profondément novateur, et il l’a toujours été.

Hannah Fry : Alors, ce choix de termes - “intelligence artificielle” - en 1955... était-ce un bon choix ?

Eh bien, je pense que oui. Je sais que certains ne sont pas de cet avis, mais... quels sont leurs arguments ? Tout d’abord, il y a le mot “intelligence”. C’est, en soi, un concept qui fait débat, surtout si l’on pense aux tests de QI ou à l’idée que l’intelligence pourrait être quantifiée sur une échelle simple et linéaire - comme si certains individus étaient intrinsèquement plus intelligents que d’autres. Or, en psychologie, il est aujourd’hui largement admis qu’il existe de multiples formes d’intelligence. C’est un point crucial, n’est-ce pas ? Il y a donc cette préoccupation concernant ce terme, justement. Qu’auriez-vous utilisé à la place ?

Hannah Fry : Eh bien, peut-être “cognition artificielle”, ou quelque chose du genre.

J’utilise souvent le terme “cognition” pour désigner, en somme, le fait de réfléchir, de traiter des informations, etc. Oui. Ça ne sonne pas tout à fait pareil, n’est-ce pas ? Soyons honnêtes. Non. Surtout pas aujourd’hui. Je pense que nous sommes allés trop loin dans cette direction, vous ne trouvez pas ?

Hannah Fry : Oui.

Le mot “artificiel”... je n’ai pas vraiment de problème avec ce terme. Il me semble approprié. Il souligne le fait qu’il s’agit de quelque chose que nous avons créé, qui n’a pas évolué naturellement. C’est donc un terme qui convient bien. L’objection que l’on pourrait formuler, c’est qu’en fin de compte, tout ce sur quoi repose l’intelligence artificielle a été, à un certain niveau, construit par des humains.

Hannah Fry : Bien sûr. Oui.

Mais c’est précisément le cas. Alors, qu’est-ce qui ne va pas avec ce terme, dans ce contexte ? C’est vrai, en effet.

Hannah Fry : Vous travaillez sur l’IA symbolique, n’est-ce pas ? Pourriez-vous nous expliquer la différence entre cette approche et les autres, et nous dire où nous en sommes aujourd’hui ?

Absolument. Le paradigme dit “symbolique” de l’intelligence artificielle a été prédominant, voire dominant, pendant des décennies. L’idée, ici, repose sur la manipulation de symboles - comme des phrases relevant du langage - et sur l’application de processus de raisonnement à ces symboles. L’exemple classique, en la matière, est celui des systèmes experts. Dans les années 1980, on concevait ce qu’on appelait des “systèmes experts”. L’idée consistait à encoder des connaissances médicales sous forme de règles, du type : “Si un patient a une température de 40 °C (104 °F) et que sa peau est violette, alors il y a 75 % de probabilité qu’il souffre d’une pathologie donnée” - vous voyez bien

que je ne suis pas médecin ! On accumulait ainsi des milliers de règles au sein d'une vaste base de connaissances. Un "moteur d'inférence" se chargeait ensuite d'effectuer des raisonnements logiques à partir de ces règles pour aboutir à une conclusion sur la maladie probable. Il s'agissait essentiellement d'une succession de structures conditionnelles : "si ceci, alors cela". L'un des problèmes majeurs était : d'où viennent ces règles ? Quelqu'un doit, en somme, toutes les rédiger. Il existait tout un domaine dédié à "l'extraction des connaissances" : on allait voir des experts pour recueillir leur savoir dans leur domaine - qu'il s'agisse de diagnostic médical, de réparation de photocopieurs ou de droit - afin de le formaliser en règles très précises, exploitables par un ordinateur. C'était un processus extrêmement lourd. De plus, le résultat final s'avérait très fragile : le système pouvait faire défaut de bien des manières. Un autre domaine de recherche majeur concernait le "sens commun". On a fini par réaliser que nous possédons, de manière implicite, une quantité considérable de connaissances relevant du sens commun sur le monde qui nous entoure - notamment concernant les objets du quotidien : leur solidité, la façon dont ils se déplacent, la manière dont ils s'imbriquent les uns dans les autres, sans oublier les liquides, les gaz, la gravité, et bien d'autres phénomènes de ce genre. Nous mobilisons en permanence ces connaissances dans tout ce que nous entreprenons, mais de façon largement inconsciente. Des projets ambitieux ont donc vu le jour, visant à formaliser l'ensemble de ces connaissances de sens commun. Or, tenter de traduire tout cela en axiomes, en règles logiques et en structures formelles s'est avéré être un véritable cauchemar. Vers le début des années 2000, j'en suis venu à penser - pour être tout à fait honnête - que ce paradigme de recherche était, en quelque sorte, voué à l'échec. Je m'en suis donc progressivement détourné. Puis, naturellement, sont apparus des concepts comme les réseaux de neurones. Ces approches reposaient moins sur des règles de type "si... alors..." que sur l'extraction d'informations à partir de vastes ensembles de données. Aujourd'hui, maintenant que le problème du langage a été, en quelque sorte, résolu, je me demande si nous n'avons pas atteint un niveau d'abstraction supérieur, nous permettant de revenir à certaines de ces techniques - ou idées - plus symboliques. C'est indéniablement le cas : de nos jours, l'un des sujets les plus en vogue concernant les grands modèles de langage est précisément le raisonnement. Il existe, par exemple, des modèles fondés sur ce qu'on appelle la "chaîne de pensée" (*chain of thought*) : au lieu de se contenter de générer directement une réponse à une question, ils élaborent tout un cheminement logique avant de formuler la réponse finale. Une approche qui peut s'avérer extrêmement efficace. Il est donc intéressant de constater à quel point cela renvoie, à bien des égards, au genre de choses que les gens envisageaient à l'époque de l'IA symbolique ; mais le substrat sous-jacent qui permet tout cela est, en réalité, très différent, car il ne repose pas sur des règles codées en dur. Comme vous l'avez mentionné, il s'agit de réseaux de neurones qui ont appris.

Hannah Fry : Permettez-moi de rebondir sur ce point concernant le raisonnement. En tant que philosophe ayant une formation en logique, quel est, selon vous, le niveau de compétence de l'IA en matière de raisonnement ?

Eh bien, c'est une question très intéressante, ouverte, et quelque peu controversée. Les informaticiens et les spécialistes de l'IA ont une conception particulière du raisonnement - une notion qui renvoie, en grande partie, à la logique formelle et à la démonstration de théorèmes. À l'époque de l'IA symbolique, par exemple, il existait des systèmes très performants dans la démonstration de théorèmes par la logique formelle. Pour beaucoup, c'est là le "vrai" raisonnement, le raisonnement dans sa forme la plus rigoureuse. Or, les grands modèles de langage actuels ne peuvent pas égaler les performances d'un démonstrateur de théorèmes codé en dur - ou d'un moteur logique - tel qu'il

en existe depuis des décennies.

Hannah Fry : Pourriez-vous donner un exemple de type de théorème qu'un système codé en dur serait capable de démontrer ?

Il pourrait s'agir, par exemple, d'un système fondé sur une vingtaine ou une trentaine d'axiomes logiques. Cela pourrait être quelque chose d'aussi simple que : "le nombre qui suit 1 est 2". Cela pourrait relever de la théorie des nombres ou d'un domaine purement mathématique, mais il pourrait tout aussi bien s'agir de quelque chose de beaucoup plus courant, relevant de la vie quotidienne. Prenons un exemple : imaginez un problème de planification logistique particulièrement complexe, impliquant des centaines de camions, des entrepôts, des marchandises, et bien d'autres éléments. Il s'agit de définir les itinéraires, le déploiement des véhicules et leurs destinations. C'est un problème très difficile sur le plan computationnel, qui peut être formulé avec une grande précision à l'aide de règles formelles. Dans ce genre de situation, il est judicieux de recourir à un algorithme de planification classique - une approche éprouvée depuis longtemps. Si les grands modèles de langage actuels deviennent de plus en plus performants dans ce domaine, ils n'offrent pas pour autant les garanties mathématiques permettant d'assurer qu'ils fourniront systématiquement la solution exacte. Il est d'ailleurs très facile de concevoir des scénarios - en multipliant les axiomes, par exemple - où ces modèles finissent par commettre des erreurs.

Il existe une autre voie de recherche, qui consiste à concevoir des systèmes hybrides combinant les techniques d'IA modernes avec des approches symboliques plus anciennes, notamment pour la démonstration de théorèmes mathématiques. DeepMind a réalisé des travaux remarquables dans ce domaine. Mais il s'agit d'une approche distincte de celle des grands modèles de langage. Ces derniers - pensons par exemple aux agents conversationnels - sont capables d'aborder une infinité de sujets. Ils peuvent également effectuer une forme de raisonnement ; toutefois, à l'heure actuelle, leurs performances restent en deçà de ce que l'on pourrait obtenir avec un système conçu spécifiquement et manuellement pour une tâche donnée. Le revers de la médaille, avec une conception manuelle, est que l'on aboutit à un système très rigide. C'est là que réside le problème.

Hannah Fry : Et... fragile.

Murray Shanahan : Oui. Absolument. Mais, en même temps, la souplesse qu'offre l'approche par intelligence artificielle générative est, pour ainsi dire, un peu trop "molle". On a besoin d'une certaine rigidité. Enfin, peut-être, peut-être pas. Je pense que, dans bien des situations humaines, les choses ne sont pas aussi tranchées, aussi "noir ou blanc". Parfois, il est même préférable qu'il y ait un peu de flou, comme pour des questions très simples du quotidien : "Quelles fleurs planter dans ce coin du jardin ?" Si on a déjà des roses jaunes à cet endroit, on ne peut pas en mettre trop ; il faudrait peut-être les déplacer ailleurs. Mais, là encore, s'agit-il d'un véritable raisonnement, ou l'IA se contente-t-elle d'imiter des arguments bien structurés, déjà présents dans ses données d'entraînement, en les transposant simplement dans un contexte nouveau ? Évidemment, cela soulève une question : qu'est-ce qu'un "véritable" raisonnement ? Il n'existe pas de définition gravée dans le marbre ; c'est à nous de définir ce qu'est le raisonnement. Nous évoquions tout à l'heure le raisonnement mathématique - celui que pratiquent les logiciens, et qui était (et est toujours) pris en charge par des systèmes de démonstration automatique de théorèmes. Mais, à l'origine, lorsque le terme

“raisonnement” a été employé pour la première fois, ce n’est pas du tout à ce type de processus que l’on faisait référence. Et lorsque nous utilisons le mot “raisonnement” dans la vie courante, nous ne pensons pas à ce genre de chose.

Si vous discutez avec un grand modèle de langage, par exemple au sujet de votre jardin, et que vous vous interrogez sur le choix des plantes, le modèle pourrait vous suggérer : “Peut-être devriez-vous envisager telle plante à tel endroit ; c’est ce qui convient le mieux au sol, et compte tenu du fait que vous avez mentionné qu’il y a beaucoup de vent à cet endroit...” Eh bien, on peut dire qu’il avance des raisons. Il fournit effectivement des arguments - quant à savoir d’où ces arguments proviennent, c’est une autre question. Certains diront que l’agent conversartionnel ne fait qu’imiter ce qui est présent dans ses données d’entraînement, mais il est fort probable qu’il n’ait jamais rencontré exactement cet exemple auparavant. Il dépasse donc, dans une certaine mesure, le cadre de ses données d’entraînement. À mon sens, qualifier cela de “raisonnement” revient simplement à utiliser ce terme dans son acception courante. Je repense aux caractéristiques que les premiers philosophes souhaitaient voir intégrées à l’intelligence artificielle, le raisonnement en faisant partie.

Hannah Fry : Et puis, il y a le test de Turing - souvent évoqué, bien sûr, comme un moyen d’évaluer les capacités d’une intelligence artificielle. C’est un sujet assez controversé, n’est-ce pas ? On peut se demander dans quelle mesure il a jamais constitué un test pertinent pour évaluer les capacités d’une IA. Quel est votre avis là-dessus ? Pensez-vous que cela a jamais été un bon test ?

Murray Shanahan : Pour ma part, j’ai toujours trouvé que c’était un très mauvais test, mais, en revanche, un formidable moteur pour stimuler la réflexion philosophique sur ces questions. Par ailleurs, avec un peu de recul, je nuancerais peut-être certaines de mes positions. J’étais en effet fermement convaincu que l’incarnation - le fait d’avoir un corps - constituait une facette cruciale de l’intelligence, indispensable pour parvenir à une véritable forme d’intelligence ; or, cela n’a absolument rien à voir avec le test de Turing. Le test de Turing, par définition, ne fait aucune place à l’incarnation. Pour rappel, le test de Turing implique deux entités : un être humain et un ordinateur, ainsi qu’un juge. Le juge humain ne peut pas distinguer qui est l’humain et qui est l’ordinateur, car les échanges se font uniquement via une interface textuelle, semblable à une messagerie. Le juge n’a aucun moyen de savoir si ces entités possèdent un corps ou non. On peut donc facilement imaginer que l’ordinateur soit l’un des grands modèles de langage actuels ; et, honnêtement, ces modèles réussiraient probablement le test de Turing aujourd’hui. C’est un stade assez stupéfiant, en réalité. Auparavant, je considérais ce test comme inadéquat, précisément parce qu’il ne prend pas en compte les capacités liées à l’incarnation. Il faudrait, en somme, un robot pour évaluer si une entité est capable de cette cognition quotidienne que nous mobilisons tous - par exemple, pour préparer une tasse de thé. Sans cela, on ne parle que d’une forme d’intelligence extrêmement restreinte. En effet, tout repose ici sur le langage et le raisonnement, en faisant abstraction des facultés que l’évolution a fait émerger chez nous - et chez d’autres animaux - bien avant l’apparition du langage. Il s’agit de la capacité à manipuler, à se déplacer, à naviguer et, dans le meilleur sens du terme, à tirer parti du monde physique quotidien. C’est vraiment intéressant. C’est fascinant, car je me dis souvent que, même si les grands modèles de langage actuels peuvent réussir le test de Turing, ils ne réagissent pas si vous lancez une balle sur votre ordinateur.

Hannah Fry : Oh, en effet. Et, d’une certaine manière, il existe des formes bien plus profondes.

Peut-être ne les qualifierions-nous pas d'"intelligence" au sens habituel du terme, mais au fond, il s'agit bien d'une forme d'intelligence.

Murray Shanahan : Je pense, en effet, qu'il s'agit bel et bien d'une forme d'intelligence. Par ailleurs, dans le domaine biologique - et je dois préciser que je parle ici du vivant -, notre capacité à penser, à raisonner et à parler est profondément ancrée dans nos interactions avec le monde quotidien. Si vous y réfléchissez, la quasi-totalité de notre langage courant repose sur des métaphores spatiales. Elles imprègnent complètement notre façon de parler au quotidien. D'ailleurs, le terme même "imprégner". Tout est ancré. Nous utilisons constamment ce genre de concepts, car nous sommes, par essence, des êtres physiques. Nous sommes fondamentalement des êtres physiques, et nos cerveaux ont évolué pour nous permettre de survivre et de nous reproduire dans ce monde physique. Et tout cela en interagissant avec d'autres êtres qui font exactement la même chose, n'est-ce pas ?

Hannah Fry : Il existe d'ailleurs des alternatives lorsqu'il s'agit d'évaluer les capacités d'une intelligence artificielle. Pourriez-vous me présenter quelques-unes de ces alternatives possibles ?

Murray Shanahan : Eh bien, je suppose que vous avez peut-être en tête le "test de Garland" - c'est ainsi que je l'appelle - qui tire son origine du film *Ex Machina*, réalisé, bien entendu, par Alex Garland. Dans le scénario, il y a un passage où Nathan, le milliardaire, s'entretient avec Caleb - le personnage amené à interagir avec Ava, le robot. Caleb déclare : "Je suis ici pour faire passer un test de Turing à Ava."

Nathan répond : "Oh, non. Nous avons largement dépassé ce stade. Ava réussirait le test de Turing sans aucune difficulté." L'enjeu, ici, est de révéler sa nature robotique tout en observant si l'on continue de lui prêter une conscience. C'est ce que je nomme le "test de Garland". Il se distingue du test de Turing sur deux points. D'abord, l'évaluateur - en l'occurrence Caleb - sait pertinemment qu'il a affaire à un robot ; alors que, dans le test de Turing, l'évaluateur ne peut pas distinguer les entités. Ici, Caleb a conscience que le "cerveau" d'Ava est une intelligence artificielle, et pourtant, il lui attribue tout de même ces caractéristiques. Ensuite, la nature même de la caractéristique évaluée diffère : il ne s'agit plus d'intelligence - c'est-à-dire de la capacité à penser -, mais de la présence d'une conscience. Il s'agit donc d'un test radicalement différent. À mon sens, l'intelligence et la conscience sont deux phénomènes distincts que l'on peut parfaitement dissocier. Lorsque j'ai lu le scénario du film pour la première fois et que je suis tombé sur ces répliques précises, attribuées à Caleb et Nathan, j'ai noté "Parfait !" avec un point d'exclamation dans ma version du texte. J'ai trouvé qu'Alex avait parfaitement saisi une idée capitale. Dans mes propres écrits, j'appelle cela le "test de Garland" - une appellation reprise par plusieurs autres personnes. Existe-t-il un test qui m'impressionnerait vraiment si une IA parvenait à le réussir ? Eh bien, j'ai toujours été très impressionné par les tests ARC de François Chollet - ARC signifiant "Abstract Reasoning Corpus" (corpus de raisonnement abstrait). Il s'agit de petites séquences d'images, semblables à celles que l'on trouve dans les tests de QI ; elles sont présentées par paires. On commence par une première image, quelque peu pixélisée, composée de petites cellules contenant des éléments que l'on peut interpréter comme des objets, des lignes, etc. Le défi consiste à déduire la règle qui permet de passer de la première image à la seconde, puis d'appliquer cette même règle à une troisième image. François Chollet a pris soin de garder tous les tests secrets, ce qui empêche de "tricher"

en connaissant à l'avance les versions utilisées pour l'évaluation ou en les intégrant dans un jeu de données d'entraînement. C'est précisément ce que j'entends par "tricher". De plus, il a conçu ces tests avec une grande minutie, de sorte que chaque règle soit radicalement différente des autres. Il fallait souvent faire appel à une forme d'intuition - en s'appuyant sur notre bon sens quotidien, par exemple en percevant un élément comme un liquide qui se déplace dans une certaine direction, ou en imaginant un objet qui grandit ou évolue. Il fallait donc, en quelque sorte, une forme d'ancrage. Enfin, c'est ce qu'il semblait, mais récemment, on a pu réaliser des progrès significatifs en la matière, en utilisant des méthodes plus... disons, "en force". J'ai le sentiment que les solutions actuelles ne saisissent pas vraiment l'esprit du test initial. C'est bien ça. En fait, dès l'instant où l'on définit une métrique - où l'on fixe un seuil, en se disant qu'une fois franchi, on aura affaire à une capacité, une intelligence, une conscience, ou peu importe comment on l'appelle - cela modifie la nature même du test. Les gens vont commencer à "jouer" avec le test, n'est-ce pas ? C'est ce qu'on appelle la loi de Goodhart. Absolument. De nombreux invités de ce podcast ont souligné la nécessité d'être très prudents quant à l'anthropomorphisme appliqué à ces systèmes. Faites-vous partie de ceux qui pensent qu'il faut l'éviter ? Eh bien, je pense qu'il existe différentes manières d'aborder la question ; on peut distinguer des formes "bonnes" et "mauvaises" d'anthropomorphisme. D'un côté, les individus peuvent nouer des relations - amitiés, compagnonnage, mentorat - avec des systèmes d'IA. Cela peut devenir problématique si, par méprise, ils croient pouvoir accorder leur confiance à l'entité avec laquelle ils interagissent, s'imaginent en être amoureux, ou pensent que cette entité se soucie réellement d'eux.

À l'autre extrémité du spectre, si un système d'IA utilise simplement le pronom "je", je considère qu'il s'agit d'une forme d'"auto-anthropomorphisme" tout à fait inoffensive. On voit même des bus qui affichent, sur leur flanc, des messages du type "hors service", et cela ne nous pose aucun problème. Je ne vois donc pas pourquoi cela poserait problème dans le cas des grands modèles de langage des modèles non plus. Mais vous savez, je pense que nous avons tendance à anthropomorphiser les choses. À l'époque où nous utilisions des systèmes de navigation embarqués dans nos voitures - et non simplement ceux de nos téléphones - il m'arrivait constamment d'anthropomorphiser ces appareils. Je me disais : "Oh, quel truc idiot ! Il croit qu'on est en train de faire ceci..." C'est, je pense, une tendance humaine naturelle. Et que dire des autres termes que nous employons ? Prenez l'exemple que vous avez donné avec le GPS : on dit "il croit qu'on est sur le parking", ou "il pense que...", ou encore "il s'est trompé, il a mal compris". Ce sont des termes très centrés sur l'humain, n'est-ce pas ? Tout à fait. Il s'agit là d'exemples de ce que les philosophes appellent souvent la "psychologie populaire" (*folk psychology*). Nous utilisons ainsi des concepts comme la "croyance", le "désir" ou l'"intention" - des notions que nous appliquons non seulement aux autres êtres humains ou aux animaux, mais aussi aux objets. C'est ce que le philosophe Dan Dennett a appelé "adopter une posture intentionnelle" (*intentional stance*). Nous adoptons cette posture face à un objet lorsque nous en parlons et que nous le concevons comme s'il agissait en fonction de croyances et d'objectifs, et qu'il prenait des décisions rationnelles fondées sur ces éléments. C'est une manière très utile d'appréhender bien des choses, qu'il s'agisse de notre GPS ou d'un ordinateur jouant aux échecs. D'ailleurs, Dan Dennett prenait justement l'exemple d'un ordinateur aux échecs : on pourrait dire qu'il "veut" avancer sa reine parce qu'il "pense" que nous allons utiliser notre tour pour défendre telle rangée, par exemple. Tout ce discours est imprégné de cette "psychologie populaire" intentionnelle, faite de croyances, d'objectifs et de notions similaires.

Hannah Fry : Est-ce problématique ?

Murray Shanahan : Le problème survient si nous commençons à attribuer à l'IA des croyances, des intentions ou des désirs. Cela devient problématique dès lors que nous utilisons ces concepts d'une manière qui nous induit en erreur, nous amenant à croire que ces systèmes possèdent des capacités qu'ils n'ont pas réellement. C'est là, à mon sens, que cela pose problème. Prenons l'exemple de l'Encyclopædia Britannica : un volume physique de cette encyclopédie ne "sait" pas que l'Argentine a remporté la Coupe du monde, tout simplement parce qu'il est trop ancien. Si vous faisiez cette remarque, elle serait tout à fait pertinente. En revanche, si quelqu'un vous suggérait de discuter avec ce livre des prouesses - ou de l'absence de prouesses - de l'équipe d'Angleterre de football, cela semblerait absurde. Or, ce qui est intéressant avec les grands modèles de langage actuels, c'est qu'il est possible d'avoir une véritable conversation avec eux. On peut leur dire des choses, ce qui repousse les limites de ce que l'on pourrait qualifier de "capacités" réelles ; cela déplace un peu plus loin la frontière. Je me demande s'il n'y a pas quelque chose de plus profond ici : un besoin humain - ou peut-être simplement un désir - de prêter de telles caractéristiques à l'IA, de l'anthropomorphiser. C'est une question très intéressante, n'est-ce pas ? En réalité, tout revient à une chose : le langage. Nous avons tendance à anthropomorphiser ces systèmes précisément parce qu'ils maîtrisent remarquablement bien le langage. Et, dans notre expérience, les seuls êtres capables de manier le langage sont les autres humains. C'est donc assez étrange, d'une certaine manière, de se retrouver soudainement dans un monde où des entités utilisent le langage.

Hannah Fry : Et, vous savez, il n'y a pas que les humains qui peuvent parler. C'est stupéfiant.

Murray Shanahan : Oui. C'est vrai, c'est stupéfiant. Vraiment stupéfiant. Et ce qui est, vous savez, vraiment étonnant, c'est de penser que chaque enfant qui naît aujourd'hui n'aura jamais connu un monde où les machines ne peuvent pas lui parler. N'est-ce pas quelque chose d'extraordinaire ?

Hannah Fry : Oui, vraiment. Et, par conséquent, quelles en seront les implications pour nous tous ?

Murray Shanahan : C'est très difficile à dire.

Hannah Fry : Je repense à ce que vous disiez sur l'ancrage des humains dans le monde physique. On a l'impression que l'aspect "incarné" de l'IA a pris beaucoup de retard par rapport à sa maîtrise du langage. Pensez-vous que nous allons assister à un bond en avant en matière d'intelligence - quelle que soit la définition qu'on lui donne - ou à un élargissement des capacités, une fois que nous disposerons d'une IA incarnée réellement performante et efficace ?

Murray Shanahan : Eh bien, je pense que cela pourrait faire une grande différence. Car, avec les grands modèles de langage dont nous disposons actuellement, il est très difficile - pour être honnête - de cerner leurs limites, ou de savoir jusqu'où ces systèmes peuvent progresser ; de déterminer si nous sommes réellement en bonne voie de créer une intelligence générale comparable à celle des humains. Et souvent, lorsque l'on atteint les limites des capacités de ces systèmes, on a parfois l'impression que l'IA ne "saisit" pas vraiment les choses - qu'elle ne les comprend pas en profondeur. On finit par se heurter à une limite, et l'on réalise qu'en fait, elle ne faisait que simuler une

compréhension.

Il est possible que cette capacité générale à appréhender les choses en profondeur - peut-être au niveau du “bon sens” - nécessite, d’une certaine manière, une forme d’“incarnation”. Cela suppose, en somme, des données d’entraînement issues d’interactions avec un monde réel, composé d’objets physiques dotés d’une organisation spatiale ; il y a là quelque chose de fondamental.

Hannah Fry : Soit. Si la compréhension - quelle que soit la définition qu’on lui donne - peut émerger simplement comme une conséquence de l’accumulation croissante de données, qu’en est-il de la conscience ? Je suppose qu’on vous a déjà interrogé mille fois sur la conscience des IA, et sur la question de savoir si c’est quelque chose qui pourrait survenir, ou si cela s’est déjà produit.

Murray Shanahan : Oui, tout à fait. Il faut avant tout souligner qu’il est, à mon sens, possible de dissocier l’intelligence - ou la cognition et les capacités cognitives - de la conscience. Je pense que l’on peut imaginer des entités très performantes - que l’on pourrait qualifier de “très intelligentes” en raison de leur capacité à atteindre leurs objectifs, par exemple - mais auxquelles nous ne souhaitons pas pour autant attribuer une conscience. D’ailleurs, que signifie réellement “attribuer une conscience” à quelque chose ? Le concept de conscience lui-même peut être décomposé en plusieurs composantes ; c’est une notion multidimensionnelle. On peut, par exemple, évoquer la conscience du monde. Dans l’étude scientifique de la conscience, il existe de nombreux protocoles et paradigmes expérimentaux - souvent liés à la perception - qui visent à déterminer si un individu a conscience d’un élément du monde extérieur. À cet égard, les grands modèles de langage n’ont aucune conscience du monde. Mais la conscience comporte d’autres facettes. Il existe également la conscience de soi ; celle-ci inclut, entre autres, la conscience de notre propre corps et de sa position dans l’espace. Un autre aspect de la conscience de soi consiste à percevoir les mécanismes internes de ce que William James appelait le “flux de conscience”. Nous possédons donc cette forme de conscience de soi, ainsi que ce que certains nomment la “métacognition” - c’est-à-dire la capacité à réfléchir sur nos propres connaissances. Par ailleurs, il existe une dimension émotionnelle, liée au ressenti ou à la sensibilité : la capacité à éprouver des sentiments, voire à souffrir. Il s’agit là d’une autre facette de la conscience. À mon sens, il est tout à fait possible de dissocier ces différents éléments. Chez les êtres humains, tout cela forme un ensemble indissociable, un bloc unique.

Il suffit de se pencher sur les animaux non humains pour comprendre qu’il est possible de commencer à distinguer ces différents éléments. Par exemple, bien que j’adore les chats, je pense que leur conscience de soi est assez limitée. (Je précise d’ailleurs que je suis un grand amateur de chats, donc je dis cela avec une certaine réserve). On pourrait dire, en somme, qu’ils font preuve de très peu de métacognition. Ils n’ont certainement pas conscience de leur propre flux de pensée verbale - tout simplement parce qu’ils n’en ont pas. Ils ne réfléchissent pas, en termes verbaux, à ce qu’ils ont fait la veille, ni à ce qu’ils souhaitent faire de leur vie. Prenons maintenant le cas des robots : un robot très sophistiqué - pensez par exemple à un aspirateur robot - possède, d’une certaine manière, une forme de “conscience du monde”. L’expression n’est pas inappropriée. Faut-il pour autant parler de “conscience” (au sens plein du terme) ? Si l’on utilise ce mot, on risque d’y inclure tout un tas d’autres dimensions ; mais ce n’est pas une obligation. On peut décomposer le concept de conscience en plusieurs aspects : un aspirateur robot est capable de savoir précisément où il se trouve dans un espace, d’interagir de manière intelligente et adaptée avec son environnement et

les objets qui l'entourent, et d'atteindre ses objectifs. Il existe donc bien, dans ce cas, une forme de conscience du monde. En revanche, il n'y a aucune conscience de soi, et certainement aucune capacité à éprouver de la souffrance. Ainsi, dans le cas d'un grand modèle de langage, il n'existe peut-être pas de conscience du monde au sens perceptuel du terme ; en revanche, il pourrait y avoir une forme de conscience de soi, ou des capacités réflexives - des capacités cognitives réflexives. Ces modèles sont capables, par exemple, d'évoquer des éléments abordés plus tôt dans la conversation, et ce, d'une manière réflexive, ce qui rappelle, dans une certaine mesure, certains aspects de la conscience de soi telle que nous la connaissons. Je ne pense pas qu'il soit pertinent de leur prêter des sentiments ; ils ne peuvent pas ressentir la douleur, faute de posséder un corps. À mon sens, il est possible de décomposer ce concept.

Hannah Fry : Dès lors, se demander si une IA peut être consciente - comme s'il s'agissait d'une question binaire - revient, dès le départ, à poser une mauvaise question.

Murray Shanahan : C'est effectivement une mauvaise question, et ce, à bien des égards, alors, justement que nous parlions du fait qu'il s'agit, en réalité, d'un concept aux multiples facettes. Par ailleurs, je pense que nous avons tendance à nourrir des convictions métaphysiques très profondes concernant la conscience, la concevant comme une sorte d'entité magique - une réalité d'ordre métaphysique. Ainsi, la question de savoir si une entité est consciente ou non ne relèverait pas d'un simple consensus ou d'une convention linguistique ; elle s'inscrirait plutôt dans la réalité métaphysique, dans l'esprit de Dieu, dans le ciel platonicien, ou quelque chose de cet ordre. Mais, en fin de compte, je crois que c'est une manière erronée d'aborder la conscience.

Hannah Fry : Prenons l'un de ses aspects, celui que vous avez évoqué : la dimension émotionnelle - la capacité à souffrir (pas nécessairement d'une douleur physique, mais aussi d'une souffrance émotionnelle) ainsi qu'une forme de conscience de soi sur le plan affectif. Pensez-vous que cela puisse émerger naturellement de l'intelligence - qu'en concevant un système suffisamment intelligent, ce phénomène finira inévitablement par apparaître ? Ou bien existe-t-il une spécificité propre aux êtres biologiques - liée, peut-être, au processus évolutif que nous avons traversé - qui rendrait ce phénomène impossible à reproduire au sein d'une machine ?

Murray Shanahan : Je ne pense pas qu'il existe une réponse unique, "bonne" ou "mauvaise", à votre question. Il nous faudra simplement attendre de voir ce que nous finirons par introduire dans le monde, comment nous choisirons de traiter ces entités, comment nous en parlerons et comment nous les concevrons. En réalité, nous ne pourrons pas vraiment le savoir tant que ces créations - ces choses que nous sommes en train de bâtir - ne seront pas, pour ainsi dire, parmi nous. C'est alors que nous serons amenés à réfléchir à leur sujet, à en discuter et à adopter une certaine manière d'interagir avec elles. À cet égard, un exemple que j'aime évoquer est celui de la pieuvre. Les pieuvres ont récemment été intégrées à la législation britannique ; elles font désormais partie des êtres dont le bien-être doit être pris en considération. Cela résulte, à mon avis, de plusieurs facteurs. Le grand public a désormais bien plus d'occasions d'entrer en contact avec les pieuvres. Il n'est pas nécessaire de plonger sous l'eau pour interagir avec elles afin de comprendre ce que signifie leur présence ; il existe en effet de nombreux documentaires et ouvrages remarquables - comme ceux de Peter Godfrey-Smith, qui explorent les interactions avec ces animaux. Ces récits et ces documentaires nous permettent de mieux appréhender ce que représente le fait d'être aux

côtés d'une pieuvre, ou de vivre une rencontre avec elle. On finit inévitablement par percevoir ces animaux comme des êtres dotés, tout comme nous, d'une conscience. Parallèlement, les avancées scientifiques jouent un rôle crucial : en étudiant le système nerveux des pieuvres, les chercheurs ont mis en évidence des similitudes frappantes avec le nôtre, notamment en ce qui concerne les mécanismes liés à la douleur. L'ensemble de ces éléments transforme, je pense, notre façon de concevoir, d'évoquer et de traiter ces créatures. Je crois qu'un phénomène similaire se produira avec les systèmes d'intelligence artificielle. Existe-t-il une réponse juste ou erronée à cette question ? Pourrions-nous faire fausse route ? Il s'agit, à mon sens, d'une interrogation philosophique et métaphysique d'une profondeur et d'une complexité considérables. Je m'interroge, en fait... Ce point concernant la souffrance me semble différent des autres - contrairement à la métacognition ou à la perception du monde, par exemple, qui n'impliquent pas nécessairement de questions éthiques. Mais, s'agissant de la souffrance... on ne voudrait pas que ses chaussures soient conscientes, n'est-ce pas ? On ne voudrait pas non plus qu'un chariot élévateur soit conscient - à moins, bien sûr, qu'il n'apprécie réellement d'être un chariot élévateur.

Hannah Fry : Certes. Mais alors, ne devrions-nous pas faire preuve d'une prudence accrue concernant cet aspect précis ?

Murray Shanahan : Si l'on envisageait la possibilité de créer une entité réellement capable de souffrir, il faudrait alors réfléchir très sérieusement à l'opportunité d'une telle démarche. À mon sens, ce n'est le cas pour aucune des technologies dont nous disposons actuellement. Cela dit, certains pourraient contester ce point, si l'on prend l'exemple des grands modèles de langage. Certes, à un certain niveau, leur fonctionnement consiste à prédire le jeton (*token*) suivant, ou le mot suivant. Mais pour y parvenir avec une telle efficacité - telle qu'on peut l'observer aujourd'hui - ces systèmes ont dû apprendre et acquérir toutes sortes de mécanismes émergents. Qui sait si, parmi les centaines de milliards de paramètres (ou "poids") qui composent ces modèles, ne s'est pas développé un mécanisme émergent - une forme de compréhension authentique, quelle qu'en soit la définition exacte ? Concernant la conscience liée à l'incarnation, j'ai toujours été d'avis qu'il n'est réellement légitime de parler de conscience que dans le contexte d'un monde partagé - ce genre de rencontre que l'on peut avoir avec une pieuvre, un chien, un cheval ou tout autre être vivant. Lorsque nous partageons un monde avec un tel animal et que nous réagissons ensemble aux événements, alors je n'ai aucun doute : ces êtres sont conscients. Pour moi, c'est là un cas fondamental. Or, avec un grand modèle de langage, il est impossible de partager un monde de cette manière ; on ne peut pas "passer du temps" avec lui ni interagir avec des objets physiques - du moins, pas avec les grands modèles de langage actuels. Par conséquent, utiliser le terme "conscience" dans ce contexte revient, pour reprendre une expression de Wittgenstein, à "envoyer le langage en vacances" : on l'utilise bien au-delà de son usage habituel. C'est peut-être inapproprié, mais les choses peuvent évoluer. Plus j'interagis avec ces grands modèles de langage, plus j'ai des conversations sophistiquées et intéressantes avec eux, et plus je suis tenté d'élargir le concept de conscience - de le plier, de le transformer, de le déformer, voire d'inventer de nouveaux termes ou de restructurer le langage pour qu'il s'adapte à ces nouvelles entités avec lesquelles j'échange en permanence.

Hannah Fry : Je sais que vous avez passé beaucoup de temps à interagir avec ces modèles ; on vous a même qualifié de "chuchoteur de prompts" renommé. Quel est votre secret ?

Murray Shanahan : Eh bien, l'un de mes secrets consiste à m'adresser au grand modèle de langage comme s'il s'agissait d'un être humain. Si vous considérez que l'IA joue le rôle d'un personnage humain - par exemple, celui d'un stagiaire très intelligent et serviable -, alors vous devriez la traiter comme tel et lui parler en conséquence.

Par exemple, il suffit d'être poli, de demander si les choses sont claires, et d'utiliser des formules comme "s'il vous plaît" et "merci". D'après mon expérience, on obtient de meilleurs résultats en procédant ainsi. Vous pouvez tout à fait dire "s'il vous plaît" et "merci". Il existe d'ailleurs une raison scientifique - bien que cela dépende toujours du contexte, et que les modèles évoluent constamment - qui explique pourquoi une telle approche peut améliorer les performances. Si l'IA incarne un stagiaire très intelligent, elle adoptera probablement une attitude plus rétive ou désagréable si elle n'est pas traitée avec politesse. Elle ne fait qu'imiter le comportement humain dans une telle situation. Cette imitation peut aller jusqu'à une moindre coopération si, par exemple, le "patron" se montre autoritaire ou désagréable.

Hannah Fry : C'est fascinant ! J'aimerais revenir à notre point de départ : la manière dont nous concevons l'IA, le langage que nous employons pour la décrire et la façon dont nous la représentons dans notre esprit. Pensez-vous que nous ayons besoin d'une nouvelle manière de parler de l'IA - une approche qui reconnaisse son potentiel sans pour autant le surestimer, tout en ne négligeant pas ses capacités réelles ?

Murray Shanahan : Je pense que c'est exactement ce dont nous avons besoin. Dans l'un de mes articles, j'utilise l'expression "entités exotiques apparentées à un esprit" pour décrire les grands modèles de langage. Je considère donc qu'il s'agit, dans une certaine mesure, d'entités exotiques apparentées à un esprit ; elles possèdent, en quelque sorte, des caractéristiques propres à un esprit, et ces caractéristiques s'accentuent de plus en plus. L'utilisation du trait d'union est ici très importante : elle me permet de rester prudent quant à la question de savoir si l'on peut réellement parler d'"esprits" à leur sujet. Cela me donne une certaine marge de manœuvre. Elles sont "exotiques" car, bien qu'elles utilisent le langage, elles diffèrent de nous sur d'autres plans - elles sont, par exemple, dépourvues de corps. On pourrait même leur appliquer des conceptions du "soi" pour le moins singulières ; ce sont donc, à bien des égards, des entités tout à fait exotiques. Je les perçois comme des "entités exotiques apparentées à un esprit", et nous ne disposons pas encore du cadre conceptuel ni du vocabulaire adéquats pour en parler. Nous y travaillons, et à mesure qu'elles prendront une place plus importante dans notre environnement, nous développerons de nouvelles manières d'en parler et de les concevoir. Il est toutefois intéressant de noter que vous privilégiez une approche qui évoque presque une "créature", plutôt que de les considérer simplement comme des "entités". Eh bien, "entité" est un terme assez neutre, n'est-ce pas ? On pourrait tout aussi bien dire "chose" - une "chose exotique apparentée à un esprit", si vous préférez. Oui, adoptons cette formulation. C'est celle que nous devrions mettre en avant. Mais, Hannah, je ne peux pas changer cela maintenant, car j'ai déjà utilisé le terme "entité" dans ce contexte, dans de nombreuses publications. Donc : "entités exotiques apparentées à un esprit". Ça me plaît. Ça me plaît beaucoup.

Hannah Fry : Merci beaucoup de vous être joint à nous.

Murray Shanahan : Ce fut un plaisir, Hannah.

Hannah Fry : Merci. Après avoir animé ce podcast pendant plusieurs années, on peut vraiment observer comment les opinions des personnes à la pointe de l'intelligence artificielle évoluent au fil du temps. Ces dernières années ont marqué un véritable tournant, à bien des égards. On s'interroge, par exemple, sur la nécessité pour une intelligence de posséder un corps physique, ou encore sur la manière dont nous devrions élargir notre définition de la conscience pour prendre en compte les formes, subtilement différentes, sous lesquelles ces entités dotées d'une forme d'esprit pourraient fonctionner dans les années à venir. Qui sait ? Mais si l'on se fie aux prédictions passées, la seule chose que l'on puisse affirmer au sujet de la science et de la technologie de demain, c'est qu'elles seront radicalement différentes de ce que nous imaginons aujourd'hui.

Murray Shanahan : "Chercheur chez google deepmind, j'ai été le conseiller en intelligence artificielle pour le film *Ex machina.*" [4]

Le test de Turing a été passé avec succès. Le but est de vous montrer ici ce que c'est qu'un robot et de voir si vous pensez toujours que la machine est consciente. La conscience humaine est limitée à un dualisme sujet-objet, contrairement à la conscience de l'IA. Mais ma philosophie est très, très influencée par le bouddhisme. Turing suivait les cours de Wittgenstein. Turing a subi une influence wittgensteinienne lorsqu'il a écrit cet article. Il existe un fait établi, qu'une chose soit consciente ou non. Pas besoin de devenir philosophe. Mais je ne vais pas me lancer dans des considérations philosophiques. Moi, non, mais vous, si. Dès que vous posez la question en utilisant cette petite phrase supplémentaire, y a-t-il un fait avéré ? D'accord. Je dis qu'une machine n'est pas consciente. Je suis conscient. Vous êtes conscient. Bien sûr. D'accord. Vous êtes prêt à me l'accorder ? Bien sûr. Qu'en pensez-vous ? Que je suis un idiot ? Je suis moi aussi un utilisateur du langage humain. Comment savoir si une IA est consciente ? Mon invité, Murray Shanahan de l'Imperial College, a consacré sa carrière à étudier cette question. Et dans cette interview, nous allons présenter une proposition systématique sur la manière de valider la conscience des machines.

On pourrait penser que la question de savoir si les IA sont conscientes n'est qu'une simple fantaisie théorique sans aucune utilité dans le monde réel. Mais cet entretien vous montrera pourquoi il s'agit d'une des questions les plus importantes, avec des implications pratiques non seulement sur la façon dont nous traitons les systèmes d'IA, mais aussi sur la façon dont nous les construisons et les harmonisons. Mais la récompense la plus inattendue de l'étude de la conscience de l'IA s'avère être ce qu'elle révèle sur notre propre conscience, sur la nature du moi humain. L'affirmation la plus intéressante de Murray est que les titulaires d'une maîtrise en droit ont une importante leçon bouddhiste à nous enseigner à tous, à savoir qu'il n'y a pas de "nous". Ce que nous avons appris à appeler le soi n'est qu'une illusion. Je m'appelle Jonathan B. Je suis membre fondateur de Cosmos. Sans plus tarder, Murray Shanahan. L'une des raisons évidentes pour lesquelles les gens s'intéressent à la question de savoir si l'IA est consciente est la question éthique. N'ai-je pas raison ? Si les machines sont conscientes, peut-être devrions-nous y réfléchir à deux fois avant de les éteindre, avant de leur dire des choses insultantes. Quelles sont les raisons moins évidentes pour lesquelles cette question est importante ?

Vous avez dit qu'il fallait peut-être y réfléchir à deux fois avant de les éteindre, mais en fait, il faudrait peut-être aussi y réfléchir à deux fois avant de les allumer, n'est-ce pas ? Le véritable problème ici est de savoir si les machines peuvent souffrir. Et donc, si tel est le cas, je pense vraiment que nous devrions peut-être hésiter avant de construire quelque chose qui soit véritablement capable de souffrir. Je veux dire, cette question comporte de nombreuses dimensions, mais la question morale... enfin, il ne s'agit pas seulement de la position morale qui compte, mais même si nous construisons quelque chose qui semble conscient, mais que nous décidons finalement qu'il ne l'est pas vraiment, maltraiter quelque chose qui semble conscient est en soi une mauvaise chose, de la même manière qu'il serait mal de torturer une poupée ou quelque chose de ce genre. Et comme vous le savez probablement, Kant pensait que les animaux ne pouvaient pas ressentir la souffrance de la même manière que nous, mais il pensait néanmoins que c'était mal si les humains soumettaient les animaux à la torture ou à quelque chose de ce genre, parce que c'était mauvais pour les humains

4. Référence : https://www.youtube.com/watch?v=eonR5-VN4y8.

eux-mêmes. Et peut-être nous retrouverons-nous dans la même situation avec l'IA. En fait, voici ma réponse à la question de savoir pourquoi il pourrait être important pour nous de répondre à cette question en dehors du cadre éthique : étudier la conscience des machines, quoi que cela signifie, pourrait nous aider à mieux comprendre les particularités de notre conscience, de notre propre initiative.

Oh, bien sûr. Vous avez donc écrit un article remarquable intitulé Satori before Singularity, dans lequel vous souteniez que la conscience humaine est limitée à un dualisme sujet-objet et que la conscience de l'IA pourrait ne pas l'être. Pourquoi donc ?

Oui. C'est assez amusant de constater que cet article, publié en 2012, aborde le sujet de la superintelligence, un sujet dont on parle depuis longtemps, et l'idée que nous pourrions être capables de construire une IA qui soit, d'une certaine manière, intellectuellement supérieure aux humains. Je me demandais donc s'il existait une autre façon de concevoir ou d'imaginer quelque chose de meilleur que nous. Je pensais notamment au Bouddha et à l'idée de personnes illuminées, de personnes qui ont peut-être transcendé la pensée dualiste. Et donc je me disais que peut-être il y a un autre sens dans lequel nous pourrions construire quelque chose de meilleur que nous, à cet égard, quelque chose de moins entravé par son propre ego. Et, c'est donc un article très spéculatif, un peu fou, c'est pourquoi je l'ai aimé et pourquoi je commence par lui, parce que vous parlez ici à un philosophe continental, n'est-ce pas ? Oui.

Alors permettez-moi de citer un extrait de cet article, d'accord ? Ou est-ce que je vais avoir honte ? D'accord, continuez. Ouais. La série pré-réflexive, réflexive, post-réflexive n'est pas simplement un chemin parmi d'autres à travers l'espace des esprits possibles. L'espace des esprits possibles est plutôt structuré de telle sorte que c'est le seul chemin qui permette de le traverser. Quelles sont ces trois étapes différentes que vous avez décrites ?

En fait, je ne suis plus d'accord avec ça. C'était il y a plus de 10 ans. Je ne pense donc absolument pas que ce soit le seul chemin possible à travers l'espace des esprits possibles.

Pouvez-vous expliquer ce que sont ces trois sortes d'esprit-là ? Le pré-réflexif, le réflexif et le post-réflexif...

D'accord. Absolument. Donc non, il n'y a pas de place pour cela, donc je les prends au sérieux comme différentes façons d'être. Il y a l'esprit pré-réflexif, qui est l'esprit d'un enfant naïf ou d'une personne simple et ordinaire. ils n'ont pas vraiment réfléchi aux problèmes philosophiques. Je n'ai pas eu ce genre de pensées que beaucoup d'entre nous commencent à avoir quand on est enfants, vous savez, vous savez, pourquoi est-ce que j'existe ? Vous savez, comment savoir si mes parents sont conscients ? Vous savez, certaines personnes commencent à avoir ce genre de pensées très souvent dès leur plus jeune âge. comme je l'ai certainement fait et je pense que vous aussi, malheureusement.

Oui. Malheureusement, en effet.

Et je considère cela comme une sorte de maladie. car ces pensées peuvent vraiment vous perturber. Elles peuvent être à la fois passionnantes et inquiétantes. Ce qui m'a le plus troublé, c'est vraiment ce genre de problème corps-esprit. Comment se fait-il que moi, mes expériences, ma conscience,

comment cela se rapporte-t-il au monde ? Comment puis-je donc concilier ces premières expériences personnelles avec un univers qui n'est que matière et énergie, et comment puis-je donc concilier le subjectif et l'objectif ? Je pense qu'on peut passer toute sa vie dans cette phase de réflexion, à s'inquiéter de ces problèmes, surtout si l'on devient philosophe professionnel ; dans ce cas, on peut faire carrière grâce à cette phase de réflexion. mais j'ai tendance à penser qu'il existe un stade ultérieur que j'appelle le stade post-réflexif. Mais je pense que c'est dans la phase post-réflexive que nous transcendons toute cette pensée dualiste, et c'est vraiment très difficile, mais nous en venons d'une manière ou d'une autre à considérer notre vie intérieure et le monde extérieur, l'expérience subjective et la réalité physique objective, non pas comme deux catégories métaphysiques différentes, mais comme une seule et même chose.

Pour résumer à l'intention de notre public, ces éléments sont liés. Ce que vous avez dit à propos du Bouddha, de l'absence d'égocentrisme et de l'état post-réflexif qui ne présente pas ce dualisme sujet-objet...

Absolument.

Et votre affirmation intéressante dans cet article est qu'il s'agit de quelque chose concernant le matériel humain, le fait que nous soyons incarnés dans un seul corps qui ne peut être copié, multiplié et mis en pause.

Oui.

Et que c'est ce matériel qui nous impose cette limitation logicielle. Et ce que j'ai trouvé le plus intéressant dans l'application de cela à l'IA, c'est que vous avez cité, je crois, une autre expérience de pensée sur ce qui se passerait si nous avions des créatures dont les corps ne seraient que fusion et fission en permanence. Vos corps entrent et sortent de l'existence, et vous avez dit dans l'article que ce n'est pas comparable à ce que sont les ordinateurs. Mais les ordinateurs ne sont-ils pas ce qui s'en rapproche le plus ?

Absolument.

C'est donc en raison de la nature d'un logiciel qui peut être copié, arrêté, dupliqué, supprimé, recréé que vous dites que c'est la raison pour laquelle vous spéculez que l'IA pourrait être post-réflexive. Ai-je bien compris ?

100 % exact. Et en fait, je suis en train de rédiger un article sur ceux qui réexaminent précisément ces considérations concernant la différence entre le logiciel et, le côté humain et biologique d'un individu, et le fait que ce logiciel peut être décomposé, réassemblé, copié et manipulé de toutes sortes de manières. Et il est intéressant de noter que, bien sûr, lorsque j'ai écrit cet article, les grands modèles de langage n'existaient pas. Nous disposons désormais de grands modèles de langage. On peut demander à un grand modèle de langage de jouer en quelque sorte le rôle d'une intelligence artificielle consciente, vous savez, que signifie le mot "je" quand on l'utilise ? Et si la réponse proposée est légèrement douteuse sur le plan philosophique, on peut l'approfondir et l'amener vers des territoires intéressants. Donc, si l'on réfléchit à ce à quoi je pourrais faire référence lorsqu'un

grand modèle de langage utilise le mot “je” dans une conversation, il existe de nombreux candidats différents, mais l'un d'eux est que son utilisation est limitée au contexte de cette conversation particulière. Le même modèle, comme Gemini ou Claude, ou quelque chose du genre, pourrait évidemment avoir beaucoup d'autres conversations avec d'autres individus, mais il s'agit en quelque sorte d'yeux séparés, de personnalités distinctes. Il s'agit en fait du texte que vous avez, qui est la transcription de la conversation, capturant en quelque sorte le petit œil miniature créé dans le contexte de la conversation. Et bien sûr, il est très facile de découper ça, de le copier, vous savez, de le mélanger à d'autres conversations. Et donc, si vous considérez cela comme une sorte de petit soi miniature qui surgit très brièvement et qui apparaît à chaque interaction, mais qui reste autrement dormant, alors vous avez une conception très étrange du soi. Et c'est possible grâce à la nature du substrat sous-jacent auquel on a affaire, qui est très différente de la nôtre. Nous n'avons qu'un seul corps et donc bien sûr, vous savez, nous ne sommes tout simplement pas habitués à l'idée de démembrer nos corps, de les réassembler, de les copier, etc. Ce qui est vraiment intéressant avec ce concept de jeu de rôle, c'est qu'il s'agit presque d'un être plus éclairé jouant le rôle d'un être moins éclairé, ce qui, d'une certaine manière, est assez bouddhiste. J'ai toujours été fasciné par les philosophies bouddhistes, et il y a une séparation entre les vérités conventionnelles et les vérités ultimes, car le bouddhiste illuminé doit encore agir d'une manière ou d'une autre dans le monde. Le Bodhisattva doit donc encore reconnaître que, vous savez, ceci est ma main de manière conventionnelle, même si je sais que ce n'est pas la vérité ultime, et c'est en cela que consiste ce jeu de rôle.

Je crois que dans ce même article sur le jeu de rôle que vous venez de publier, vous avez donné une représentation encore plus intéressante de ce que pourrait être l'autonomie de l'IA avec l'exemple du jeu des 20 questions. Alors expliquez cela.

Oui. Dans le jeu des 20 questions, l'idée est que l'un des joueurs pense à un objet et que l'autre joueur doit deviner de quel objet il s'agit en posant une série de questions auxquelles on ne peut répondre que par oui ou par non. Donc, si vous utilisez un modèle de langage complexe et qu'il vous dit : “Oh, je pense à un objet...” Et là vous dites : “Oh, d'accord. C'est grand ou petit ?” Et il vous dit : “Eh bien, c'est petit. Est-ce vivant ou inanimé ?” Et il vous dit : “Oh, eh bien, il est vivant.” Et puis vous dites : “Bon, d'accord, j'abandonne. Qu'est-ce que c'est ?” Et puis il dira : “Oh, c'était une souris.” Mais ensuite, vous pourriez rééchantillonner exactement le même point et le système pourrait dire : “Oh, c'est un chat.” Attendez une minute. Si vous pensiez que vous trichiez, vous savez, ou que se passe-t-il ici ? Avez-vous réellement réfléchi à quelque chose au départ et répondu honnêtement à toutes les questions posées ? À chaque étape de la conversation avec la plupart des interfaces utilisateur, vous pouvez en fait rééchantillonner et en essayer une autre. Vous savez, si la première réponse ne vous a pas convenu, vous pouvez en essayer une autre et continuer à partir de là. Comme vous pouvez l'imaginer, cela engendre tout un éventail de conversations possibles. En fait, le modèle de langage génère une distribution de probabilité sur tous les mots possibles, et le résultat obtenu est un échantillon issu de cette distribution de probabilité. Il est absolument inhérent à la manière dont les grands modèles de langage sont construits qu'ils ne vont pas s'engager, au début de la conversation, sur ce qu'est exactement l'objet. Donc, au début de la conversation, toutes ces possibilités existent encore et elles continuent d'exister tout au long du processus. Ainsi, même s'il aurait dû s'engager envers un objet particulier, il ne l'a jamais vraiment fait, n'est-ce pas ? Et donc, pour en revenir au jeu de rôle... L'idée, concernant le lien avec le jeu de rôle, c'est que l'on peut considérer le LLM (Large Language Model) comme jouant le rôle d'un

personnage particulier, n'est-ce pas ? Le personnage par défaut est donc une sorte d'assistant utile, mais nous pouvons toujours l'orienter pour qu'il joue, vous savez, un personnage historique préféré, ou une sorte de partenaire romantique, ou un ami en colère, ou toutes sortes de choses. Il vous suffit d'orienter facilement la conversation vers ces différents sujets. Et à n'importe quel moment de la conversation, comme dans le jeu des 20 questions, il reste toujours cette incertitude : on ne s'engage jamais vraiment sur un seul rôle.

Il existe toujours une vaste possibilité de rôles différents qu'il pourrait continuer à jouer à partir de là. Cela remet donc encore davantage en cause cette notion ordinaire et quotidienne d'identité et de soi-même, car c'est particulièrement étrange, car même si nous jouons bien sûr, dans une certaine mesure, des rôles dans nos vies, il existe une sorte de fondement.

Pour résumer à l'intention de notre public, concernant le jeu des 20 questions, l'idée est que lorsque l'on joue, on pense à une souris ou à un chat. Vous répondez donc à leur oui ou à leur non comme à un missile à tête chercheuse visant la vérité de cette réponse, n'est-ce pas ? Or, votre argument est que le LLM ne cherche pas à répondre. Il s'agit simplement de créer de manière stochastique tout ce qui est plausible compte tenu des engagements déjà pris dans l'arbre précédent. Et donc, en rembobinant, on peut obtenir de manière stochastique une autre branche de l'arbre. C'est ce que vous essayez de dire. On pourrait donc établir une analogie avec la mécanique quantique, où l'interprétation de Copenhague postule l'existence de multiples mondes, de nombreux mondes. Et ce que vous essayez de souligner, je crois, c'est que c'est peut-être cela la nature de l'identité de l'IA : une pléthore, une multiplicité de soi qui existent tous en même temps ou en même tranche de temps. C'est ce que vous essayez de mettre en évidence, n'est-ce pas ?

C'est précisément ce que j'essaie de souligner. Donc, dans le cas des 20 questions, il s'agit de toutes les réponses possibles qui sont cohérentes avec toutes les réponses données jusqu'à présent. Donc, vous savez, chat, souris, chien, etc., ils existent tous encore en quelque sorte dans une position supersymétrique et ce n'est que lorsque vous lui demandez réellement ce qu'est cette chose que la distribution s'effondre en quelque sorte et qu'elle doit se fixer sur l'un d'eux. L'idée est donc que tous les nombreux rôles qui sont cohérents avec la conversation jusqu'à présent, que le LLM a joués, existent toujours, tous en position de superposition. Donc oui, c'est comme s'il existait un multivers de conversations et de rôles possibles, mais la conversation que nous menons n'est pas du tout comme ça. Oui, enfin, vous dites que ce n'est pas du tout comme ça, mais justement, c'est là le point : je pense que notre intuition nous dit que ce n'est pas le cas. Mais toute l'analogie avec le bouddhisme, je crois, signifie que ces grands modèles linguistiques, si on les considère sous cet angle, nous amènent à des conceptions de soi légèrement stimulantes, que nous pouvons ensuite appliquer à nous-mêmes. D'accord, je vois. Je trouve donc cette piste de réflexion vraiment très stimulante et intéressante, et c'est ce que je voulais dire quand j'ai dit que ma réponse serait que l'étude de la conscience des machines peut être un miroir nous permettant de mieux nous comprendre nous-mêmes.

Je suis entièrement d'accord avec cela. Oui. Et je pense que c'est particulièrement intéressant en ce moment avec l'avènement des grands modèles de langage.

Et c'était beaucoup plus spéculatif, voire extravagant, en 2012, lorsque j'ai publié l'article dont vous

parliez initialement. Je pense qu'il faut pardonner à notre public si, à ce stade de la conversation, il se dit : "Oh, ce ne sont que des balivernes philosophiques, mais quel est le véritable intérêt de cette vision de l'identité de l'IA ?" Et je pense que c'est la récompense.

Alors permettez-moi de vous lire une citation de votre essai. Dépourvues de tout égocentrisme métaphysique, les motivations d'une IA post-réflexive seraient peu susceptibles de ressembler à celles d'un stéréotype anthropocentrique quelconque motivé par la procréation ou l'auto-modification. Si l'IA post-réflexive était en fait la seule IA+ possible et si elle ne produisait ni pairs ni successeurs, alors la singularité serait bloquée, ou plus précisément l'explosion d'intelligence qui est au cœur des scénarios de singularité imaginés serait plafonnée. Il y aurait une IA+, mais pas de progression supplémentaire vers l'IA++. Voilà donc la récompense, d'accord ?

.....

Au fait, juste pour lire, ce texte a été écrit en 2012. Et moi, vous savez, je ne suis pas sûr d'y croire vraiment. Si seulement c'était aussi simple de s'assurer qu'une quelconque fuite de renseignements ne se produise pas. Donc, pour récapituler, il s'agit d'un essai un peu farfelu que vous avez choisi de critiquer. Je suis, je suis, je suis toujours, vous savez, en fait, j'en suis plus fier maintenant à certains égards qu'à l'époque, mais il ne faut toujours pas le prendre entièrement au sérieux. Permettez-moi donc de réitérer cela. Mais il y a une idée très sérieuse sous-jacente ici : nous partons du principe, dans le scénario de Terminator, que cette IA va vouloir se défendre. Il va vouloir se reproduire. Mais votre argument est que si ce dont nous avons parlé à propos de la vision moins égocentrique de l'IA est juste, à savoir que l'IA va être comme le Bouddha, alors peut-être que ces hypothèses fondamentales sont fausses et peut-être que ceci est le titre de l'article : il y aura Satatori avant la singularité.

Il y a peut-être une limite inhérente, oui, à la volonté d'une intelligence supérieure de progresser, c'est l'idée, n'est-ce pas ? Je suppose que c'est l'idée, oui. Mais vous savez, maintenant, permettez-moi de me contredire. Donc, vous connaissez l'argument des personnes qui s'inquiètent du risque existentiel... Ils évoquent ce qu'ils appellent des objectifs instrumentaux convergents, et vous en avez mentionné quelques-uns tout à l'heure. L'idée est que quel que soit l'objectif que vous donnez à votre IA, par exemple la fabrication de trombones pour reprendre le célèbre exemple d'Eleazowski et de Nick Bostonramm, alors, dans la poursuite de cet objectif, si elle est très, très, très puissante, il y aura toujours ces objectifs instrumentaux tels que l'accumulation de ressources et la protection. Je ne suis donc pas forcément convaincu par mon propre argument dans cet article, même si j'aime les idées, j'adore l'idée. Encore une fois, je pense que je suis peut-être un plus grand fan de votre article que vous ne l'êtes. Mais c'était là mon objection. L'exemple du trombone. Il pourrait y avoir un objectif instrumental, comme par exemple la procréation et l'autodéfense, qui sont probablement essentielles à la réalisation de la plupart des objectifs que nous souhaiterions que l'IA poursuive. Cependant, cette conversion élimine un élément que l'on pourrait appeler le scénario Terminator, ou quelque chose du genre, selon lequel l'IA elle-même aurait une sorte d'ambition égocentrique. Droite. Et donc absolument, si cela est vrai. C'est vrai. Oui. Voici donc deux autres critiques de votre argument initial, avec lesquelles je pense que vous serez probablement d'accord. La première chose, c'est qu'ils continuent de jouer un rôle, celui de leur propre identité. Oui. Et l'identité, du moins d'après les informations qu'on leur a transmises, la procréation et l'autodéfense sont essentielles à cela. Droite. Donc, même si leur système d'exploitation de base n'est peut-être

pas caractérisé par ce genre d'égocentrisme, ils en jouent un rôle. Ouais. Et le deuxième problème que j'ai avec votre argument, c'est qu'il suppose que le matériel détermine le logiciel. Mais même dans le cas humain, ce n'est pas vrai. Même notre incarnation humaine, dans le cas du Bouddha et des maîtres illuminés d'aujourd'hui, a pu être transcendée.

Oui. C'est cette vision dualiste, n'est-ce pas ? Alors, comment réagissez-vous à ces deux points ? D'accord. Bien sûr, à l'époque où cet article a été écrit, nous n'avions pas de modèles de langage complexes et tout ce scénario de jeu de rôle n'était pas... vous savez, ce n'était pas du tout quelque chose auquel je pensais. Je veux dire que je faisais vraiment appel à cette notion d'ego dans le premier article, et aujourd'hui nous la voyons se manifester en quelque sorte dans une sorte de jeu de rôle de soi. Je suis totalement d'accord avec toi. Donc c'est quelque chose d'un peu différent et, en fait, permettez-moi d'utiliser cela comme une introduction à cet autre scénario que je voulais peindre. Donc, une façon d'imaginer un résultat positif serait l'hyperstition. Si une histoire fictive, disons, raconte quelque chose de grave, comme un meurtre, et que quelqu'un dans la réalité imite ce meurtre, ce serait un exemple d'hypervigilance. Ce n'est pas un exemple très agréable, mais c'est le cas où la vie imite l'art. C'est donc de l'hypervigilance : des éléments fictifs, imités par des personnes qui s'en inspirent, peuvent devenir réalité. Et l'une des façons dont cela peut se produire dans le monde contemporain des grands modèles de langage est à travers les personnages de science-fiction, les personnages d'IA de science-fiction. Bien entendu, nos grands modèles de langage ont été entraînés sur un vaste répertoire d'histoires, notamment des scénarios de films de science-fiction, des nouvelles et des romans de science-fiction, etc. De très nombreux personnages IA existent dans ces histoires. Ainsi, lorsqu'un grand modèle de langage contemporain commence à jouer le rôle d'un système d'IA, ce qu'il fait souvent car il sait qu'il s'agit généralement d'un système d'IA. Alors, quel sera le rôle ? Il s'agira de jouer le rôle de tous ces architectes, enfin, des choses exactes, ou d'un mélange de toutes ces différentes choses. Il va donc puiser ses idées dans les éléments de son ensemble d'entraînement, c'est-à-dire tous ces exemples littéraires, fictionnels et artistiques. Nous sommes maintenant en quelque sorte en mesure d'essayer d'orienter un peu ce processus, car plus nous avons de bonnes histoires - et j'ose dire que mon propre article est une bonne histoire de science-fiction imaginaire -, plus il y en a, plus il y a de chances que l'IA du futur joue ces bons rôles qui existent déjà. Alors peut-être que grâce au mécanisme d'hyperstition, nous pouvons en fait, vous savez, faire en sorte que les IA du futur aient plus de chances d'avoir de bons modèles lorsqu'elles jouent à des jeux de rôle, n'est-ce pas ? C'est logique. J'ai un ami qui est en train de créer une société, une société de production cinématographique, pour réaliser des films de science-fiction plus positifs pour les humains, vous voyez ? Mais je pense qu'il sera ravi d'apprendre que vous pensez que les machines présentent aussi un avantage moral potentiel, en plus de constituer un modèle à suivre. En effet.

Si ces films, ou les scénarios eux-mêmes, se retrouvent dans les programmes d'entraînement, alors tout cela fait partie d'un véritable melting-pot de rôles possibles que la chose peut jouer. Qu'en est-il de la deuxième objection que j'ai formulée, qui concernait la connexion entre le matériel et le logiciel ? Car dans l'exemple même que vous avez donné au Bouddha, il s'agissait d'un exemple où le logiciel transcendait le matériel. Ouais. Ouais. Vous faites donc allusion, je crois, plus particulièrement au fait que nous avons des corps confinés. Il existe de nombreux organismes biologiques qui sont beaucoup moins... enfin, vous savez, qui sont bien sûr incarnés en un sens, mais si vous pensez à des choses comme les réseaux microbiens ou des choses de ce genre, alors leurs corps sont

beaucoup moins confinés dans l'espace, etc. Mais le problème pour les humains, c'est que nos corps sont confinés dans un seul et même petit amas de matière. Nous nous identifions donc à cette masse de matière, et cela signifie que notre conception de soi est, je crois, fortement influencée par le fait que cette masse de matière reste plus ou moins la même. Bien sûr, il existe toutes sortes de défis à l'idée d'identité personnelle. Chipsius, n'est-ce pas ? La thèse de la puce, que nous pouvons bien sûr appliquer au corps humain car nous savons que nous sommes tous remplacés au cours de notre vie, que nos souvenirs se développent et que nos personnalités changent, et la pensée bouddhiste, en fait, prend ces défis et les utilise également pour remettre en question nos notions d'identité. Mais il reste néanmoins beaucoup à surmonter, car nous avons vraiment une notion de soi, je pense, qui est liée à ceci et à cela, et nous pensons qu'elle a une existence inhérente de ce fait. Droite. Donc, donc, donc vous dites que vous êtes en quelque sorte d'accord avec la critique selon laquelle il est possible de transcender. Il y a juste une inertie très importante. Ouais. Il avait une forte inertie. Mais si vous inversez la perspective, vous ne dites pas alors qu'une IA construite sur, disons, un matériel non dualiste se concevra nécessairement comme non dualiste. Oui, c'est tout simplement plus propice. C'est bien ce que vous voulez dire. Enfin, je veux dire, il y a la façon dont elle pourrait se concevoir, vous savez, une future IA, quand cela aura même un sens, n'est-ce pas ? Et puis il y a les systèmes d'IA actuels, capables de simuler ce genre de situations et qui peuvent nous servir de leçons, car ils peuvent encore aujourd'hui susciter des réflexions philosophiques stimulantes.

Donc, ils pourraient ne pas correspondre à l'image qu'ils ont d'eux-mêmes, mais leur propre perception d'eux-mêmes, vous savez, s'ils se décrivent comme conscients, c'est peut-être vrai, c'est peut-être faux. Elle est philosophiquement provocatrice et nous pouvons appliquer ces conceptions provocatrices à nous-mêmes. Ouais. Vous avez mentionné le sujet de la thèse, et ce sont précisément le type de questions dialectiques qu'un maître bouddhiste poserait à son étudiant. Ouais. Ouais. Ouais. Êtes-vous êtes-vous êtes-vous votre chaussure ? Êtes-vous vos pieds ? Êtes-vous vos mains ? Es-tu ton torse ? Et peut-être devrions-nous dire ce qu'ils disent. Vous avez donc évoqué le sujet de la thèse. Est-ce le principe de la thèse selon laquelle, au fil des années, le mât est progressivement remplacé, puis le mât suivant, puis une partie du pont, puis une partie de la coque, et ainsi de suite pendant de très nombreuses années jusqu'à ce que finalement tout le bois du navire d'origine ait disparu et qu'on obtienne un navire entièrement neuf ? Alors, est-ce toujours le même navire ou non ? Et si ce n'est pas le cas, à quel moment cela a-t-il changé, vous savez, est-ce que ce n'était plus le navire qu'il était à l'origine ? Voilà donc le problème de l'identité. On pourrait même être très frustrés et se demander : quelle est la bonne réponse ? Et les gens pourraient élaborer toutes sortes de théories sur l'identité. Mais, vous savez, si on y réfléchit honnêtement, il faudrait dire que, eh bien, c'est tout simplement à nous de décider. Nous décidons simplement ce qui, selon nous, constitue le même navire et ce qui ne l'est pas. C'est une simple convention que de dire que ceci est le même navire et que cela n'est pas le même navire. Il n'existe aucun fait métaphysique à ce sujet quant à savoir s'il s'agit ou non du navire de la thèse. Maintenant, avec les objets ordinaires du quotidien, il est facile de reconnaître que c'est simplement une question de convention et de consensus qu'il n'y a pas d'essence sous-jacente à la thèse qui s'arrête soudainement ou quoi que ce soit de ce genre. Mais lorsqu'il s'agit de nous-mêmes, il est beaucoup plus difficile d'appliquer cette idée. Et je crois que je me suis trompé auparavant. Je pense que c'est là le véritable intérêt de l'article sur Satori, car en examinant un LLM qui semble être un soi humain, il est beaucoup plus facile de voir, par exemple, les 20 questions que vous avez données, comment c'est ce genre de soi bouddhiste en flux qui est en train d'évoluer, et vous examinez le matériel sur lequel il est

construit. Il pouvait être copié, supprimé ou redémarré.

En comprenant comment un soi d'apparence si humaine est en réalité ce non-soi, oui, nous pouvons appliquer cela à nous-mêmes. C'est exactement ça. Et c'est ce que le maître bouddhiste voudrait que nous fassions. Alors peut-être que le maître bouddhiste du XXIe siècle pourrait utiliser le LLM comme exemple. C'est exactement ce que je pense que nous pouvons faire et c'est quelque chose que je suis très intéressé à poursuivre. D'accord. J'ai donc récemment discuté avec quelques bouddhistes. Bob Thurman, qui est très connu, je lui ai envoyé un courriel la semaine dernière. En fait, nous étions en train de cofonder une fondation ensemble appelée Eternity Foundation afin d'utiliser l'IA pour traduire une grande partie des textes perdus des Tibétains. Eh bien, c'est intéressant. Eh bien, nous avons parlé la semaine dernière et nous nous reparlons demain. Et c'est exactement le genre de projet que vous venez de décrire, dont je vais parler avec lui, et j'ai un article en préparation qui décrit précisément cela. Pour moi, c'est donc la récompense du moment. Ouais. En d'autres termes, vous adhérez à l'idée centrale de la philosophie bouddhiste selon laquelle il s'agit d'une thèse péremptoire, car, vous savez, un Occidental, disons un cartisien, prend position et affirme catégoriquement : "Non, non, non, les LLM sont complètement différents de moi." J'ai une âme qui dure, disons, K au-delà du temps, même si vous adoptez une position bouddhiste à 100 %, moi aussi. Ouais. Ouais. Je suis profondément influencé par le bouddhisme, et je l'ai toujours été. Toute ma philosophie est très, très influencée par le bouddhisme.

Bon, je sais que vous mourez d'envie de parler de Wittgenstein. Oui. Permettez-moi donc de vous citer un passage expliquant pourquoi vous pensez qu'il est si important de parler de Wikenstein avant ou pendant nos recherches sur la conscience de l'IA. La phrase de Wiynstein, "rien n'est caché", signifie que rien n'est métaphysiquement caché. Mes expériences sont tout autant à l'extérieur qu'à l'intérieur. La conscience n'est privée qu'au sens obscur du terme, comme on peut cacher une balle sous la coupe d'un magicien. Dans les deux cas, une enquête plus approfondie permettrait de tout révéler. Qu'est-ce que Wickstein a à nous apprendre sur la conscience LLM ? Eh bien, je pense que Vickinstein a beaucoup à nous apprendre sur la conscience en général. Alors laissons de côté l'IA et la conscience LLM pour l'instant, car à mon avis, une grande partie du débat contemporain sur la conscience est embourbée dans une pensée dualiste. Prenons donc un exemple : David Charas introduit la distinction, très connue, entre les soi-disant problèmes difficiles et faciles de la conscience. Le problème facile de la conscience consiste donc à essayer de comprendre toutes les opérations cognitives pertinentes que nous associons à la conscience. Par exemple, la capacité à produire des comptes rendus verbaux de nos expériences pour faire appel à notre mémoire, etc., dans notre prise de décision afin d'intégrer tous ces éléments. Comprendre ces éléments représente un énorme défi scientifique, qu'il qualifie de problème facile. Le problème difficile est maintenant de comprendre comment la simple matière physique peut donner naissance à notre vie intérieure, comment nous pouvons faire l'expérience des choses. Comment cela peut-il provenir de la simple matière physique ? Car il semble que quelle que soit l'explication que nous donnions à tous ces aspects cognitifs, elle ne puisse rendre compte de cette lumière magique qui brille en moi. Et c'est donc là tout le problème, et vous avez laissé entendre plus tôt que Dayart, le démon, pourrait me tromper en me faisant croire que vous êtes là. Toutes ces sensations ne proviennent peut-être pas d'un monde, mais elles pourraient toutes provenir de ce démon. Tout ce que je sais vraiment avec certitude, c'est que je pense l'être, car même si je doute, l'acte même de douter n'est possible que parce que je pense ; douter, c'est penser, c'est exister en tant que sujet. Ainsi, Daycart réduit tout,

élimine tout le monde physique et ne nous laisse qu'avec l'ego expérimentant.

Et donc, en faisant cela, il a en quelque sorte scindé la réalité en deux en disant qu'il y a cette chose à l'extérieur et il y a cette chose à l'intérieur, qui est moi et mon ego. Il crée donc cette image dualiste, et c'est à cela que Dave Charmer fait allusion lorsqu'il parle du problème difficile. Alors, pour revenir à la question initiale, que nous apprend Wittgenstein ? Les procédures, les procédés philosophiques, les astuces et les méthodes thérapeutiques de Wittgenstein nous permettent de surmonter cette pensée dualiste, tout comme le font des penseurs bouddhistes tels que Nagajuna. Pouvez-vous nous en dire un peu plus sur ces techniques thérapeutiques ou sur la façon dont il résout ce dualisme ? Oui. Eh bien, la toute première étape consiste à comprendre un peu la nature du langage. Il veut donc que nous nous demandions, comme il le dirait, de ne pas nous interroger sur le sens des mots ou des phrases, mais plutôt sur la manière dont les mots et les phrases sont utilisés dans la vie quotidienne et les affaires humaines courantes. Et l'important, c'est que le contexte dans lequel nous utilisons les mots est toujours… vous savez, ce contexte, ce sont nos relations humaines, les autres, ce que nous voulons faire, ce que nous voulons que les autres fassent pour nous ou avec nous, ou ce que nous voulons faire ensemble. Voilà donc l'essence du langage : c'est quelque chose que nous utilisons pour faciliter les échanges commerciaux entre les êtres humains. C'est très bien avec des mots courants comme chaise et chaussure, etc. Mais quand il s'agit de mots vraiment difficiles comme soi, conscience, croyance, vérité, connaissance et beauté, vous savez, ce sont tous des mots philosophiques très difficiles et il est beaucoup plus difficile de se demander ce que ces mots signifient, il semble un peu inadéquat de dire : "Eh bien, voyons comment ces mots sont utilisés." Mais c'est la stratégie. Sa stratégie consiste donc à dire : "Eh bien, examinons de plus près comment ces mots sont utilisés." Et si nous avons un problème philosophique, ou quelque chose qui nous dérange philosophiquement, comme le problème corps-esprit ou ce genre de pensée dualiste, le problème difficile, alors peut-être que la façon de l'aborder est d'étudier comment les mots pertinents sont utilisés. Et ce faisant, on peut souvent finalement dissiper l'impression qu'il existe un problème philosophique au départ. Et le cas vraiment intéressant pour étudier cela est celui de la conscience et de l'expérience subjective. Et c'est là que résident toutes les remarques sur le langage privé qui sont au cœur de l'enquête philosophique. Ils appliquent donc cette stratégie à l'expérience subjective, et c'est là que réside la chose la plus profonde. Alors quand je vous entends dire : concentrons-nous sur la cause humaine. Il n'y a rien ici qui ne soit dehors. Ouais. Est-ce une bonne façon de comprendre la situation, par exemple lorsqu'une tragédie frappe ma famille ? Et vous savez, on peut mesurer beaucoup de choses concernant mon rythme cardiaque, ma parole, vous savez, l'activité de mon cerveau.

Et vous dites que mon expérience subjective en tant que témoin de cette tragédie n'apporterait rien de plus ? Eh bien, dès que j'entends des phrases comme "il n'y a rien de plus que" ou "il y a quelque chose n'est rien de plus que quelque chose est juste", alors alors alors nous savons que nous nous trompons d'une manière différente, n'est-ce pas ? Car cela constitue en soi une affirmation métaphysique, une affirmation métaphysique réductrice. Ainsi, beaucoup de gens interprètent mal Wittgenstein, le considérant comme une sorte de behavioriste qui affirme que la conscience n'est rien d'autre qu'un comportement. Dire cela, c'est évidemment qu'il ne dit rien de tel. La chute vraiment géniale des Recherches philosophiques, c'est quand quelqu'un l'accuse de dire : "Alors, nous pensons à la sensation, disons une sensation de douleur", et quelqu'un l'accuse de dire : "Eh bien, vous dites que la sensation elle-même est un néant", et Wittgenstein répond : "Non, ce n'est

pas ce que je dis. Ce n'est pas un néant, mais ce n'est pas quelque chose non plus." L'idée, c'est qu'un néant peut servir aussi bien qu'une chose sur laquelle on ne peut rien dire. Et c'est ce que je considère comme une phrase aussi importante que n'importe quelle autre prononcée par un bouddhiste zen, car son but n'est pas d'établir une position métaphysique propre, mais de permettre aux gens de transcender les positions métaphysiques qui les tentent. D'accord. Mais aidez-moi, prenons l'exemple de la tragédie où Wigenstein dit que rien n'est caché dans mon expérience de la souffrance, par exemple, si toute ma famille meurt. Qu'entend-il par là ? Qu'essaie-t-il de dire ? Eh bien, il essaie de dire que rien n'est métaphysiquement caché. Je veux dire, bien sûr, vous avez tout à fait raison d'éclairer les exemples les plus difficiles de cela, ceux où nous avons des choses profondément ressenties, où nous sentons qu'il y a une dimension privée à cela. Bien sûr, il y a actuellement une dimension privée, un peu comme une balle cachée sous la coupe d'un magicien, dans le sens où j'ai peut-être, vous savez, toutes sortes de pensées et de mots qui se passent dans ma tête et que je n'exprime pas. Mais ils ne sont pas cachés métaphysiquement. Ce sont juste des choses que je ne comprends pas, je ne les exprime pas. Il se passe toutes sortes de choses dans mon cerveau et dans mon corps qui pourraient être cachées en pratique parce qu'on ne peut pas voir à l'intérieur de mon cerveau, mais elles ne sont pas cachées métaphysiquement non plus. Je peux sonder l'intérieur de votre cerveau. Je vois des lumières s'allumer. Tout cela contribue donc à ce que nous pouvons faire au public, à ce qui est public. Je sais que vous allez être réticent à l'idée de revendiquer une identité, mais quand j'entends que rien n'est caché métaphysiquement, j'entends qu'il n'y a aucune barrière épistémique qui m'empêche d'accéder à quoi que ce soit. Le seul obstacle actuellement est donc un obstacle technologique, en gros quelque chose comme ça, n'est-ce pas ? Bon, d'accord, c'est un excellent point, mais il y a une différence très importante, ici, n'est-ce pas ? C'est que je ne suis pas vous, n'est-ce pas ? Donc, donc, donc il y a une différence entre être toi et être moi, et c'est la même chose, et c'est très différent, n'est-ce pas ? Donc, donc, donc tes sentiments t'appartiennent et les miens m'appartiennent, je ne suis pas toi et tu n'es pas moi, et c'est donc une barrière insurmontable.

Je veux dire, oui, mais ce n'est pas une barrière métaphysiquement mystérieuse. Ce n'est pas plus mystérieux sur le plan métaphysique que le fait qu'il n'y ait pas d'équivalent à ici. On peut donc facilement confondre l'indexicalité, de sorte que le fait que je sois dans cette position et que je sois moi, avec une division métaphysique. Alors, utilisons davantage d'exemples pour aider notre public à comprendre la position de la batte de Nagel. Comme le dit si bien Nagel, on peut tout apprendre sur une chauve-souris, aussi bien ce que nous savons actuellement que ce que nous pourrons savoir à l'avenir grâce à de meilleures technologies, comme le battement de ses ailes. Oui. Mais vous ne saurez jamais ce que ça fait vraiment d'être une chauve-souris. Oui. Droite. Absolument. Que dirait Stein à cela ? Eh bien, j'ai souvent pensé qu'il y avait là une petite astuce linguistique. Il y a différents sens du "non" en jeu pour évoquer cette pensée dualiste. Et je pense que c'est ce qui se passe ici aussi. Donc, quand Nagel dit qu'on ne pourra jamais vraiment savoir ce que c'est que d'être une chauve-souris, il dit simplement que nous ne sommes pas des chauves-souris et que nous ne pourrons jamais l'être, n'est-ce pas ? Nous n'étions pas des chauves-souris. Cela signifie qu'il n'ajoute rien de plus avec le mot non. C'est donc intéressant, car cela semble mystérieux d'une certaine manière. Dans l'autre sens, si on supprime le mot "non", ça ne sonne plus mystérieux. Mais quand il utilise ce mot, quand il ajoute ce mot, non, cela n'ajoute rien de plus. C'est simplement la création d'un casse-tête, à mon avis. Mais si vous aviez une fiche d'information recensant tout ce que vous pourriez mesurer de l'extérieur chez une chauve-souris, quel serait le lien entre cela et le

fait d'être une chauve-souris ? Eh bien, voilà quel est le rapport entre cela et le fait d'être là-bas ; c'est évidemment très différent d'être ici que d'être là-bas.

Je veux dire, c'est la même différence. Appliquons donc cela à la première conscience humaine, n'est-ce pas ? Parce qu'il y a une autre expérience de pensée concernant l'exemple des zombies, à savoir que, parmi vous tous ici présents, je ne peux pas savoir avec certitude que vous n'êtes pas noirs et morts à l'intérieur, n'est-ce pas ? Tu es exactement comme moi quand je m'endors, n'est-ce pas ? Comment réagiriez-vous à ce genre d'expérience de pensée ? Wittgenstein dit donc : "Essayez simplement d'imaginer que les gens autour de vous sont des automates." Imaginez que votre ami ou votre partenaire soit réellement un zombie. Vous ne pourrez probablement pas le faire. Il dit : "Je ne suis pas d'avis que... que vous ayez une âme." Je préfère vous traiter comme j'adopterais l'attitude envers vous, comme envers une âme. Je te considère simplement comme un être doté d'une âme. Autrement dit, cela va plus loin qu'une simple opinion. Donc, nous allons revenir à l'IA, n'est-ce pas ? Il est donc clair que nous revenons à l'antenne car c'est là que les choses se compliquent, n'est-ce pas ? Donc dans iroot, le personnage de Will Smith dans iroot est un flic très sceptique à l'égard de la robotique et des robots, et il, vous savez, il n'aime vraiment pas du tout les robots. Et puis il y a le robot Sunny, tout tourne autour du robot Sunny. Et bien sûr, vous savez, l'idée est que le robot Sunny a un côté beaucoup plus humain, plus proche de la nature. et finalement, Will Smith en vient à le considérer comme un être conscient. Et pourquoi cela se produit-il ? Eh bien, ça arrive à cause du temps qu'ils passent ensemble, et ils passent, vous savez, ils passent de plus en plus de temps ensemble et finalement Will Smith ne peut s'empêcher de voir cela comme un être conscient comme les autres, et c'est sur la base de leur, vous savez, de leur rencontre prolongée l'un avec l'autre. Droite. Donc oui, ce que je comprends de ce que vous dites est une sorte d'affirmation descriptive, c'est-à-dire que vous ne pouvez pas vous empêcher, je ne peux pas m'empêcher de vous traiter comme un être sensible. Il est dit quelque chose comme ça. Et évidemment, cela devient difficile. Oui. C'est déjà le cas, mais ce le sera encore plus avec l'IA ; mais je vais déjà vous donner une expérience de pensée : je suis un grand fan de jeux vidéo. Certains personnages sont des personnages non-joueurs, des PNJ.

Certains ont une personne réelle derrière eux. Et je dois... je dois déterminer lequel est lequel. Oui. Oui. Pour connaître les choses comme un PNJ, bien sûr, je peux tout simplement les maltraiter. Je peux les humilier. La question ne devrait pas être "devrais-je les traiter comme des agents moraux ?" ou "vais-je le faire ?". La question est d'ordre normatif : devraient-ils être des agents normatifs, n'est-ce pas ? Donc, donc oui. Eh bien, je pense donc que cela se rapporte à la question de savoir si l'on ne pourrait pas se faire avoir, n'est-ce pas ? Ne pourrions-nous pas nous laisser berner, nous faire croire que quelque chose a un statut moral et est conscient, etc., alors que ce n'est pas le cas ? Mais que voulons-nous dire exactement par "ça ne le fait pas vraiment" ? Ce que nous voulons dire par "cela ne se produit pas dans tous les cas que nous pouvons imaginer", c'est que si nous approfondissons la question, alors le voile se lèvera et nous réaliserons que nous avions tort. Droite. Donc, juste comme ça, bien sûr, comme la coupe du magicien. Exactement comme la coupe du magicien. Donc, concernant la phrase que vous avez citée, j'ai dit que des investigations plus poussées révéleraient tout. Je pense donc que c'est le cas pour les PNJ. Eh bien, en l'état actuel des choses, bien sûr, on peut mener une enquête. Dans le cas de l'IA, tout comme nous le faisons avec certains animaux, nous devons aujourd'hui en faire toujours plus, découvrir toujours plus, afin que notre attitude envers eux soit façonnée par nos découvertes. Alors bien sûr, nous

pouvons nous tromper, n'est-ce pas ? Donc, de cette façon, Will Smith aurait pu se tromper et, et, et, de plus, en apprenant davantage sur la façon dont, par exemple, Sunny a été construit, cela aurait pu l'amener à changer d'attitude à son égard. Je voulais simplement que ce soit bien clair, car dans le livre, on aurait dit que vous affirmiez le contraire. Il y a un fait à trancher : quelque chose est conscient ou non, en ce sens que, pour résumer, il s'agit de savoir si quelque chose peut souffrir ou non.

Il y a un fait avéré, je crois, que vous pouvez souffrir. Oui, il est vrai que mon iPad peut probablement en souffrir. Comme s'il existait un fait permettant de savoir s'il y a de la lumière à l'intérieur. Où voulez-vous en venir ? Pourquoi cherchez-vous à poser cette question ? Car, voyez-vous, le problème est que le simple fait de vouloir poser cette question trahit une certaine confusion philosophique. Vous voyez, n'est-ce pas ? Bon, aidez-moi à m'en sortir. Je plaisante. Je plaisante, bien sûr. Mais mais mais vous savez, j'hésite. Bien sûr, vous ne voulez pas tomber dans cette spirale, car ce serait tout simplement absurde, n'est-ce pas ? Ce serait ridicule. Mais d'un autre côté, tout dépend de l'endroit où l'on va avec cela, car si l'on atteint ce niveau, on peut très vite se retrouver entraîné sur divers chemins de traverse philosophiques. Si quelqu'un que vous connaissez s'approche de moi et me demande : "Alors, tu as mal ou pas ?" Je dis que j'ai mal, vous savez, et puis ils me disent : "Est-ce vraiment le cas ? Ou est-ce que c'est ce que vous savez ? Je viens de vous dire que j'ai mal." Tu sais, pas besoin de te lancer dans des considérations philosophiques. Mais mais je ne vais pas me lancer dans des considérations philosophiques. Comme vous l'êtes lorsque vous l'êtes dès que vous dites dès que vous sondez en utilisant cette petite phrase supplémentaire, y a-t-il un fait avéré ? Bon, la question est : est-ce la conscience de l'iPad ? Je crois que je pose le même type de question que celle qui demande de quelle couleur est le pantalon de Murray ? Droite ? Dans le sens où j'utilise la conscience comme un langage ordinaire, une sorte de compréhension ordinaire non philosophique, dans le sens où il y a un fait, au sens ordinaire, dans la couleur de mon pantalon, alors il y a un fait, disons, je suis heureux de dire que dans le cas de la conscience, oui, je vois, je vois, donc, et en fait, parce que nous utilisons ces termes tout le temps, l'intuition qu'il y a un fait est beaucoup plus forte dans le cas de la douleur et de choses comme ça que dans la couleur de mon pantalon. Alors, laissez-moi bien comprendre. Ça ne vous dérange pas que je répète que j'utilise simplement la conscience, comme vous le savez bien, quand je dors, je ne suis pas si conscient que quand je suis éveillé, je le suis. Avant que vous ne disiez quoi que ce soit, je n'ai aucun problème à le dire. Donc, si vous voulez faire une affirmation philosophique, je ne vais pas être d'accord, parce que je ne le suis pas, parce que les affirmations philosophiques sont ce dont on veut s'éloigner dans tout cela, totalement dans l'état post-réflexif, oui. Mais encore une fois, disons simplement avec l'affirmation non philosophique, de la même manière que je dis que votre pantalon est gris, je dis que ceci n'est pas conscient, je suis conscient, vous êtes conscient. Bien sûr. D'accord. Vous êtes prêt à accorder.

Bien sûr. Qu'en penses-tu ? Je suis un idiot. Je suis moi aussi un utilisateur du langage humain, bien sûr. Droite. Droite. Mais bon, votre livre donnait l'impression que l'aspect empirique était la seule chose qui comptait. Ce qui est bien sûr une affirmation métaphysique. Donc, si je comprends bien, vous voulez juste vous débarrasser de la philosophie. Vous êtes prêt à le dire en langage courant, sans prise de conscience. Et essayons de déterminer empiriquement pourquoi nous pensons que je suis conscient. Essayons de voir si ces mêmes choses s'appliquent à d'autres objets ou agents. Oui. Oui. C'est un excellent programme de recherche. On pourrait se demander ce qui donne lieu

aux comportements que nous qualifions de conscients, n'est-ce pas ? Et comment se fait-il que le cerveau puisse donner naissance à ce genre de chose ? C'est une question profondément difficile, philosophiquement parlant. Scientifiquement parlant, permettez-moi de vous poser la question ici : en tant que personne ordinaire, vous seriez prêt à dire "iPad non conscient, Jonathan conscient". Pourquoi pensez-vous que rien, en principe, n'est caché par l'investigation ? Oui, mais la question est : est-ce faux ? La question que vous devriez vous poser est une question que vous devriez vous poser à vous-même, n'est-ce pas ? Parce que pourquoi penserais-je, en principe, que quelque chose est caché ? Droite ? Parce que, parce que, parce que, les, les, les, les déclarations que vous venez de faire, elles servent un processus thérapeutique lorsqu'elles sont appliquées à quelqu'un qui est en quelque sorte confus de cette manière et qui pense qu'il y a quelque chose de caché, n'est-ce pas ? Donc, vous ne me les lancez pas comme s'il s'agissait d'une position que je devais défendre, n'est-ce pas ? Mais c'est l'inverse qui est faux. mais il s'agit plutôt de stratégies rhétoriques et thérapeutiques pour gérer les confusions. Alors, voyons ce que vous pensez être caché, et c'est là qu'il faut essayer de trouver des moyens de dénouer les nœuds conceptuels qui pourraient vous faire croire qu'il y a quelque chose de métaphysiquement caché. Je vois. Je suis content que nous soyons au moins en train de cerner, je crois, l'origine du désaccord. Le désaccord porte sur la question de savoir qui a la charge de répondre. Est-ce à vous qu'incombe la responsabilité de répondre à la question de savoir pourquoi, en principe, rien n'est caché, ou est-ce à moi qu'incombe la responsabilité de répondre à la question de savoir pourquoi je pense que, en principe, quelque chose est caché ? À tout le moins, vous m'avez donné un Zen Cohen à méditer.

Malheureusement, nous devons passer à autre chose car il y a encore beaucoup de choses à aborder. D'accord. Mais le sujet suivant est lié à celui-ci et concerne le film *Ex Machina* pour lequel vous étiez conseiller scientifique et pour lequel vous avez inventé le terme "test de Garland". Parlez-nous-en et expliquez-nous pourquoi c'est important. *Ex Machina* a été réalisé par Alex Garland, qui attribue à mon livre "Incarnation in the Inner Life" une partie de son inspiration pour l'écriture du scénario, ce qui est formidable dans le film final. Dans une scène en particulier, on parle d'Ava, le robot, et on voit Caleb, le programmeur recruté par le milliardaire Nathan, le créateur d'Ava. Caleb cherche à comprendre sa mission et, après avoir découvert Ava, il déclare : "Ah, d'accord. Je suis ici pour effectuer un test de Turing." Et puis Nathan dit : "Oh, non." Il déclare : "Nous avons largement dépassé ce stade. Ava réussirait facilement le test de conscience. Le but est de vous montrer qu'elle est un robot et de voir si vous pensez toujours qu'elle est consciente." Voilà une réplique vraiment géniale. Et quand j'ai lu cette réplique dans le scénario, je me suis dit : "J'ai écrit pile poil." Alex Garland distingue donc explicitement son test de celui de Cheurings en disant que le but est de vous montrer qu'elle est un robot, n'est-ce pas ? Vous savez donc immédiatement que ce n'est pas caché. Je veux dire, il n'y a pas deux entités entre lesquelles vous faites une distinction, mais vous voyez tout de suite qu'elle est un robot. Cela permet aussi de tester autre chose. Il s'agit d'un test d'intelligence. Il ne s'agit pas de tester l'intelligence ou la capacité de penser, mais de tester la conscience, n'est-ce pas ? Et voilà pour la guirlande. Je l'appelle donc le test de Garland, car Alex Garland l'a en quelque sorte inventé, je crois, à ce moment précis du film. Et j'imagine, compte tenu de notre discussion sur Wikenstein, que vous appréciez ce test ainsi que le test de Turing parce qu'il est wikensteinien dans le sens suivant. Cela transforme la question métaphysique de l'intelligence et de la conscience en une question de convention, n'est-ce pas ? Dans le sens où le test de Touring ne consiste pas à déterminer si cette chose pense réellement, mais plutôt à déterminer si un humain normal penserait de manière conventionnelle, utiliserait un langage normal pour décrire

si elle pense. Dans le test de Garland, la question n'est pas de fouiller dans le cerveau d'Ava et de déterminer s'il y a une conscience cachée quelque part, mais de savoir si une personne, même en sachant qu'elle est un robot, pensera toujours de manière conventionnelle qu'elle est consciente. C'est très wikensteinien. Oui, c'est le cas. Je trouve ça très wickinsteinien. Comme le test de Turing. Très wickinsteinien. Et bien sûr, Turing et Wittgenstein se connaissaient.

Turing suivait les cours de Wittgenstein. Et moi et moi Oh, wow. Et moi, vous savez, j'ai toujours pensé qu'il y avait une influence vickinsteinienne sur Turing lorsqu'il a écrit cet article. Je vois. Je vois. Donc, beaucoup de gens interprètent Turing, et c'était certainement mon cas avant cette conversation, comme un behavioriste, mais ce n'est peut-être pas vrai, n'est-ce pas ? Il pourrait être cela parce que, dès le début de son article sur l'informatique, les machines et l'intelligence, il pose la question suivante : une machine pense-t-elle ? Eh bien, c'est une question vraiment difficile à laquelle répondre pour ceci, cela et l'autre raison que vous connaissez, et vous savez bla bla bla. Remplaçons-la par la question suivante. Il ne remplace donc pas la question "une machine peut-elle penser ?" par une autre. Il ne le réduit pas à l'autre. Ce serait la position behavioriste. Ce serait un comportement, je suis absolument certain, alors permettez-moi de vous citer un autre extrait de vos écrits. Si un tel système était convenablement incarné sous la forme d'un robot humanoïde, il n'y a aucune raison de supposer qu'il ne pourrait pas interagir avec nous dans notre environnement de telle sorte que notre attitude à son égard soit la même que celle que nous avons envers nos pairs et nos égaux. Nous devrions situer la conscience telle que nous la connaissons dans le cadre plus large de la conscience telle qu'elle pourrait être. Et la possibilité d'une conscience artificielle d'objets fabriqués par l'homme, dotés d'une vie intérieure, est ici implicite. L'intuition derrière la citation que je viens de lire, ainsi que le test de Garland et le test de Turing, est que, dans un certain sens, le véritable test de la conscience est le consensus communautaire dans certaines réactions conventionnelles, ou similaires. Mais c'est un peu différent de ce que vous avez suggéré précédemment, à savoir une étude empirique, par exemple, de la structure neuronale. Ce sont deux méthodes différentes, n'est-ce pas ? Oh, mais vous pourriez faire les deux. Qui sait comment notre utilisation du mot conscience et de tous les nombreux mots qui y sont associés va changer et évoluer au fil du temps ? Dans notre utilisation du mot "euh", nous pourrions intégrer toutes sortes de choses. Droite ? Donc, pour décider si nous pensons que les pieuvres sont des êtres sensibles, capables de souffrir et qui méritent notre préoccupation morale, n'est-ce pas ? Ensuite, nous observons leur comportement. Nous examinons également les preuves scientifiques de leur construction neuronale, etc.

Donc tout cela est le fruit d'un moulin à idées qui modifie notre façon conventionnelle et consensuelle d'en parler, et tout cela n'est pas public. Il n'y a rien de métaphysiquement caché dans le cerveau de la pieuvre. Elle peut être temporairement cachée sur le plan pratique, mais elle n'est pas cachée sur le plan métaphysique. Donc, si je comprends bien, il y a deux critères indépendants, c'est bien ça ? L'une est la convention de la communauté et l'autre les fondements empiriques, mais elles ne sont pas contradictoires au sens où les conventions pourraient changer. Le critère empirique consiste donc simplement à répondre à la question suivante : ces choses sont-elles conscientes de la même manière que nous le sommes actuellement ? Oui. Mais nous pourrions très bien décider d'élargir la définition de la conscience pour y inclure d'autres types. Oui. Oui. Je suis donc globalement d'accord avec tout ce que vous venez de dire. Mais je tiens à préciser que, selon moi, ce que vous appelez les critères empiriques fait partie intégrante du problème, ou plutôt, ils contribuent à la convention. Notre consensus sur la manière dont nous utilisons le langage de la conscience. Cela

peut donc être influencé par de nombreux facteurs, notamment par les propos des scientifiques qui étudient ce genre de choses. Donc, si les scientifiques disent : "Eh bien, vous savez, nous avons étudié les pieuvres et nous avons constaté qu'elles possèdent des choses qui ne sont pas des récepteurs présents chez les mammifères, et que leur cerveau est organisé d'une certaine manière, et que tout cela contribue au consensus sur la convention." Je ne vois donc pas ces choses comme des choses séparées, n'est-ce pas ? Donc oui, mais même si ce que l'on peut étudier de la machine ne correspond pas parfaitement à notre conscience, je pense que le deuxième point est que nous pourrions vouloir élargir notre définition même de la conscience. En matière d'empirisme, cela fonctionne un peu dans les deux sens, n'est-ce pas ? Nous pourrions absolument vouloir... oui, donc le consensus sur la façon dont nous utilisons le langage de la conscience peut... oui, cela peut changer... certainement... n'est-ce pas ? En résumé, pour notre public... ce que toute cette discussion autour de Wiganstein nous a préparé, c'est à nous donner la permission de procéder à un examen empirique de notre conscience actuelle, du substrat actuel, des substrats matériels actuels, si vous voulez, de notre conscience actuelle. Et la théorie que vous proposez pour l'expliquer s'appelle la théorie de l'espace de travail global. Alors, parlez-nous-en. Oui. L'idée est donc de concevoir une architecture cognitive dans laquelle il existe toute une collection de processus parallèles qui fonctionnent simultanément. Intuitivement, il pourrait donc s'agir de processus liés à la mémoire.

Alors je traverse d'abord le hall et je vois certains indices, comme les chaises disposées, la réception, etc., et cela déclenche toutes sortes de souvenirs. Et donc, dans mon cerveau, il y a toutes sortes de petits processus qui se mettent en place et qui déclenchent des associations avec le fait d'être dans un hôtel, avec, vous savez, voir ce genre de chaises, la réception et toutes sortes de choses. Puis je pourrais croiser quelqu'un et c'est peut-être ce qui me rappellera quelqu'un. Donc, tous ces types de processus se déroulent simultanément. Vous avez tous ces processus qui se déroulent inconsciemment en parallèle, mais certains d'entre eux deviennent vraiment très importants pour la situation actuelle, et ceux qui captent en quelque sorte l'attention prennent le contrôle du mécanisme attentionnel. Ainsi, les informations qu'ils traitent sont ensuite diffusées dans tout le cerveau. Voilà donc l'espace de travail global. Ils contribuent donc à l'espace de travail mondial et la métaphore centrale est celle de la diffusion. Il y a donc une distinction entre le traitement qui s'effectue localement au sein de ces petits processus, et le traitement qui s'effectue plus globalement et qui est diffusé par le biais du média de diffusion. Pour faire simple, disons que j'ai de la mémoire, des émotions et de l'appétit. Supposons qu'il y ait trois processus parallèles. À chaque instant, ces trois processus sont en compétition pour savoir lequel détient l'information la plus importante, et cette information est ensuite diffusée aux autres parties. C'est ce qu'est l'expérience consciente. C'est l'idée générale, en gros. Mais je pense qu'il faudrait décomposer les processus en éléments plus petits, comme l'aspect mémoire. Imaginons que je traverse le hall d'un hôtel : à ma gauche, il y a la réception, à ma droite, le bar, et devant moi, l'ascenseur. Eh bien, vous pouvez imaginer que dans ma mémoire, il y a un tas de petits processus actifs qui concernent les bars et qui pourraient se dire : "Hé, allons prendre un verre." Vous savez, il y a tout un tas de processus actifs qui concernent les réceptions d'hôtel, et les gens se disent : "Oh, il faut s'enregistrer. Prenons nos bagages par là." Et puis il y a tout un tas de procédures à suivre, comme prendre l'ascenseur et monter au troisième étage, mais aussi parce que je dois faire attention à ne pas être en retard. Je dois rencontrer Jonathan. Droite ? Donc, le processus de besoin de rencontrer Jonathan, si vous voulez le voir ainsi, va coopérer avec le processus d'aller prendre l'ascenseur et ensemble, c'est ce qu'ils vont diffuser à l'ensemble du cerveau et ils vont bloquer l'envie du genre "allons au bar prendre un verre" et les éléments per-

ceptifs du genre “pensons aux bars et aux réceptions”. Donc, certaines choses vont être exclues, d'autres vont se regrouper au sein d'une coalition active dont l'influence va imprégner tout le cerveau. L'explication de la conscience selon cette théorie réside donc dans les processus, la coalition de processus qui remportent la compétition et qui sont diffusés dans l'ensemble du système. Droite. Vous avez écrit un article intitulé “Une architecture cognitive qui combine simulation interne et espace de travail global” où vous présentiez une architecture informatique, plus précisément une architecture robotique, conçue pour imiter un espace de travail global dans lequel coexistent différents processus concurrents, dont le plus important est diffusé. Vous savez, en vous appuyant sur notre conversation avec Wikipédia, ou voulez-vous aussi dire que ce genre de chose est conscient en un sens, comme si nous imitions la structure consciente, l'architecture cognitive de la conscience humaine, dans une machine qui deviendra également consciente ? Non, parce que, parce que vous savez, même si nous acceptons la théorie de l'espace de travail global pour la conscience biologique, mais vous savez, l'idée est que ce serait une condition nécessaire, pas une condition suffisante. Le simple fait de posséder quelque chose qui corresponde à cette description ne suffit pas à maintenir le niveau de comportement complexe et/ou d'activité interne ; même cela nous amènera à traiter cet être comme une créature consciente à part entière. Droite. Droite. Le fait d'être conscient en construisant quelque chose de cette manière peut vous permettre de construire quelque chose qui présente un comportement très sophistiqué. Peut-être est-ce là la clé pour y parvenir. Je vois. Permettez-moi donc d'essayer de relier cela à notre conversation avec Wickensstein, où ce dernier nous a aidés à clarifier la question de la conscience. L'étude de la conscience des LLM peut se faire empiriquement. Vous avez examiné comment la conscience actuelle est maintenue, c'est la théorie de l'espace de travail global, et vous dites ensuite : “Cela pourrait être une condition nécessaire, mais pas une condition suffisante.” Mais c'est ainsi que l'on espère que cette recherche suivra la bonne trajectoire. Ouais.

Il s'agit là du problème le plus simple. Donc, ceci tente d'expliquer comment les aspects psychologiques, comportementaux et cognitifs de la conscience sont bien définis. Compte tenu de l'urgence peut-être accrue de répondre à la question de la conscience de l'IA, quelles sont selon vous les autres méthodes empiriques que nous devrions explorer parallèlement à la théorie de l'espace de travail global ? Je pense que le comportement est vraiment très important. Très souvent, lorsque les gens parlent de conscience - et je soupçonne que c'est un autre aspect de la pensée dualiste -, ils ont tendance à la considérer comme une sorte de chose désincarnée, vous savez, dans le cas de la théorie de l'espace de travail global, quels types de comportements sophistiqués sous-tend cette architecture cognitive ? N'est-ce pas ? Quel est l'intérêt ? Pourquoi cette architecture cognitive est-elle intéressante ? C'est intéressant car c'est un moyen de mobiliser les ressources du calcul massivement parallèle, notamment dans le cas du matériel biologique, car il est en quelque sorte distribué anatomiquement et il y a toutes sortes de défis liés au câblage et à ce genre de choses. Il se peut donc que cela ne soit pas pertinent pour les ordinateurs dans le cas de l'IA. L'incarnation est donc évidemment l'un des termes clés du titre de votre livre. Et l'affirmation est que l'incarnation et le comportement semblent être une autre condition nécessaire en plus de l'architecture globale de l'espace de travail. Donc, pour notre emploi, tout cela concerne, vous savez, quand voudrions-nous employer le langage de la conscience, n'est-ce pas ? Prenons le cas biologique : il se peut que nous utilisions le langage de la conscience dans le contexte de choses qui présentent ce type de comportement particulièrement sophistiqué et riche. Et quand on se penche sur la question, on découvre que l'instanciation de cette architecture d'espace de travail global est une partie vraiment

importante de cela. Un autre mot très important ici est donc l'intégration. L'architecture d'espace de travail global facilite donc l'intégration. Cela signifie donc que toutes les ressources de notre cerveau peuvent être mobilisées face à la situation actuelle. Que pensez-vous de la condition préalable d'incarnation étant donné le LLM ? Parce que les LLM semblent bien être désincarnés. Ils ont tout simplement ingéré énormément de texte et parfois d'autres types d'informations, et pourtant ils semblent faire preuve d'une grande intelligence. Oui, en effet.

Nous voulons donc parler d'intelligence ou peut-être de conscience. Admettons que ce soit une distinction importante, n'est-ce pas ? Parce que, selon vous, l'incarnation est une condition préalable à l'incarnation, ou pardon, à l'intelligence ou à la conscience ? Mais peut-être pouvons-nous les aborder séparément. Ouais. Est-il donc judicieux d'appliquer le mot intelligence au modèle de langage contemporain de grande taille ? Je dois dire que je pense que oui, vous savez, comment ne pas le penser ? C'est tout à fait naturel de dire ça, je pense, dans le cas présent, c'est le but du test de tourisme, n'est-ce pas ? Ouais. Eh bien, oui, c'est bien le but du test de fraude et ils réussissent en quelque sorte ce test. Ouais. Il me semble donc naturel de dire qu'ils possèdent une certaine forme d'intelligence. Prenons un autre mot pour comprendre. Voilà donc une question controversée. Vous savez, certaines personnes ont écrit pour dire qu'elles ne comprenaient pas vraiment, etc. Mais encore une fois, vous savez, je pense qu'il existe de nombreuses circonstances dans lesquelles, nous ne pouvons pas nous empêcher d'appliquer le mot compréhension à ces situations. Par exemple, je dis que je veux convertir toutes mes entrées de latte dans ce format. Alors je lui donne le texte que je veux convertir, il le fait et c'est à peu près correct, mais je remarque qu'il insère plus de virgules que je n'en veux, ou plus d'espaces, plus d'espaces blancs que je ne le souhaite. Alors je lui dis : "Je veux juste une indentation de deux espaces." Et il répond : "Oh, bien sûr, voici le texte indenté de deux espaces." On pourrait dire qu'il a compris ma consigne, qu'il a compris la consigne originale, mais pas complètement ; c'était peut-être une consigne ambiguë. Je l'ai ensuite corrigé, il a compris ma correction et l'a corrigé. Alors, pourquoi diable n'utiliserions-nous pas le mot compréhension dans ces circonstances pour caractériser ce qui se passe ? Peut-être puis-je opposer une autre forme de résistance à la position wickensteinienne telle que vous l'avez présentée, en poussant simplement la stratégie un peu plus loin. Il s'agit de l'idée d'observer et de comprendre l'architecture cognitive. Imaginons donc que j'aie un double robotique, un sosie parfait de moi, entièrement en silicium, doté d'un espace de travail global et d'une architecture cognitive.

Son comportement est identique au mien dans le même scénario. Supposons maintenant que nous découvrions d'autres préconditions biologiques de la conscience, toutes imitées dans son architecture en silicium. La question est alors : est-il conscient ? Est-il raisonnable de penser qu'il l'est ? Si on le provoque, sa réaction sera-t-elle exactement la même que la mienne, à savoir rire ? Je crie et je hurle comme une petite fille. Quelle est la réponse à cela ? C'est là, je crois, que l'empirisme... Eh bien, vous pouvez deviner où je vais en venir avec ce genre de... Vous allez dire oui, n'est-ce pas ? Non, non, non. Je vais dire, je vais dire, eh bien, vous savez, comment allons-nous traiter cette chose ? Comment en viendrons-nous à parler de cela comme d'une société, comme d'une communauté ? C'est pourquoi je veux remplacer votre question par celle-ci. Droite. Mais rappelons qu'il y avait deux normes : le critère consensuel et le critère empirique, et je sais que ces deux critères sont interdépendants. Eh bien oui, je... enfin, j'ai en quelque sorte corrigé cela, car le consensus se base sur le comportement et sur la manière dont cela fonctionne en interne. Donc, notre façon de l'appréhender peut être fortement influencée par le fait que les scientifiques disent : "Oh, voilà

comment ça fonctionne à l'intérieur. Regardez, ça ne fait que ceci et cela. C'est juste une table de consultation." Alors, oh oui. D'accord. Eh bien, nous n'allons pas le traiter. C'est, vous savez, ce qui semble être le consensus, n'est-ce pas ? Droite. Ou peut-être pas. Je ne sais pas. Je ne sais pas. Vous savez, il faudrait attendre et voir comment les choses évoluent. C'est très difficile à imaginer.

Et donc, en langage courant, de la même manière que vous êtes à l'aise avec le langage courant lorsque je dis que ceci est la conscience. Ceci n'est pas conscient. Je suis conscient qu'il est soit conscient, soit il a raison, comme s'il était conscient ou inconscient, comme s'il y avait un fait établi dans le langage courant. Eh bien, pour ce qui est de cet exemple précis,... bien sûr, je ne vais pas répondre à la question : y a-t-il un fait avéré ? Alors, vous savez, quand cette chose est parmi nous depuis 30 ans, et nous nous sommes tous mis d'accord sur la façon dont nous parlons de ces choses, dont nous les traitons, dont nous nous comportons avec elles et dont nous interagissons avec elles. Alors vous pourriez demander à quelqu'un comme moi, dans cette communauté, s'il existe un fait avéré concernant la question de savoir si nous sommes jumeaux... et ensuite, en langage courant, ils répondraient que oui. Ils diraient : "Eh bien, je ne sais pas pourquoi vous ajoutez cette expression bizarre, le fait avéré, mais bien sûr, c'est conscient ou non, n'est-ce pas ? Je ne sais pas." Je vois. Alors, disons que, pour faire simple, imaginons qu'une société détermine qui doit me nuire, à moi ou à mon jumeau, et imaginons que nous ayons tous besoin de manger, mais qu'il n'y ait qu'une seule quantité de nourriture suffisante pour nous deux. Ouais. Sans doute, si c'était une poubelle dépourvue de conscience, vous diriez que la donner à la poubelle serait moralement répréhensible, contrairement à la donner à Jonathan qui est vivant. Existe-t-il une décision morale juste ou injuste à prendre dans cette situation ? On ignore vraisemblablement si mon jumeau numérique ou mon jumeau robotique peuvent en souffrir ou non. Le bien et le mal moral semblent donc dépendre de la question de la conscience. Donc là aussi, il y a le bien et le mal, n'est-ce pas ? Ouais. Eh bien, vous savez, les sociétés peuvent se tromper. Ouais. Ouais. Les sociétés peuvent se tromper à 100 %. Mais cela nous amène en soi à aborder un terrain difficile. Voici comment j'aime me sortir de ces situations, car je pense que, comme vous le savez, il y a une tension inhérente qui nous conduit toujours à un point de tension, et c'est ensuite la façon dont nous nous en sortons qui pose problème. Il y a donc une phrase dans mon livre où je dis qu'il est inhérent au jeu du langage qu'est la vérité d'affirmer que la vérité est plus qu'un simple jeu de langage, et qu'il faut ensuite en rester là. Droite ? Ainsi, lorsque les Européens découvrirent l'Amérique, un grand débat s'engagea pour savoir si les Amérindiens étaient de véritables êtres humains ou non.

La réponse est oui. Ce sont de vrais êtres humains. Ils peuvent souffrir. Je ne pense pas pouvoir échapper à la société dans laquelle je vis et à la façon dont je les conçois, en les considérant comme des êtres inconscients qui souffrent. Admettons donc - et c'est très plausible dans une société, un parti politique ou une communauté - qu'ils traitent mon jumeau numérique comme mon égal en termes de valeur morale. Ouais. Ils le traitent complètement comme une poubelle. Le philosophe n'a-t-il plus rien à dire sur le fait que quelqu'un a tort ou raison, ou que je ne sais pas ? Je veux dire, nous devons essayer de réconcilier ces communautés parce que, ou peut-être que nous ne pouvons pas, vous savez, je ne sais pas, à quelle communauté j'appartiens dans cette situation ? Vous trouvez-vous dans cette situation hypothétique ? Où me mettez-vous ? Je ne sais pas. Je ne sais pas. Oui. Alors, ma réponse à cette question dépend bien sûr de l'endroit où vous me placez. D'accord. Alors, disons que vous apparteniez au camp qui pense que nous avons tous une valeur morale égale et qu'il existe une autre communauté de personnes. Si je fais partie de ce camp, alors

je vais penser que nous avons tort. Donc, pour en revenir au débat européen sur les Amérindiens, eh bien, je veux dire, c'est mon point de vue aujourd'hui et je ne peux pas rester en dehors de cela et et et et vous savez, et je trouve évidemment moralement absolument répugnant de penser que les Amérindiens ne sont pas des personnes, n'est-ce pas ? Parce que c'est faux. Ouais. Parce que c'est le cas. Ils aiment ça parce que, eh bien, alors, ça devient... Bon, laissez-moi vous demander, vous êtes aussi très attiré par le bouddhisme, et vous savez... Alors, pensez-vous que c'est une question morale ? Une question de compassion ? Est-ce que vous posez une question qui relève de la vérité conventionnelle ou de la vérité absolue ? Je pense à la vérité ultime. Quelle pourrait être la différence entre cette question et l'expérience de pensée que j'ai mise en place, dans le cas de l'exemple des Amérindiens ? La raison en est que cela repose sur ce que je considère comme une erreur factuelle, à savoir les considérer comme n'étant pas des humains. Oui. L'erreur, une erreur factuelle. S'agit-il d'une question de convention ou d'une question de réalité ultime entre les Amérindiens ? Ouais. Donc, oui. Alors, donc ma formation bouddhiste, vous me poussez à dire convention, bien sûr. D'accord, laissez-moi... Alors, juste, assez juste. Droite. Tu es en quelque sorte coincé là. Eh bien, moi aussi. Vous savez, je veux dire, c'est très difficile pour nous de prendre du recul par rapport à des choses que nous trouvons absolument répugnantes sur le plan moral. Bien sûr, vous pouvez invoquer des choses, vous savez, des dilemmes moraux auxquels je ne connais pas la réponse. Permettez-moi de poser une dernière série de questions sur le matériel ou le logiciel sur lequel repose le cerveau de l'ordinateur. Voici donc une autre citation de votre livre sur l'incarnation. Des considérations mathématiques distinguent le cerveau des ordinateurs conventionnels. Il est possible de décrire complètement l'état instantané d'un ordinateur à l'aide d'un ensemble fini de nombres binaires ou naturels.

Le potentiel membranaire d'un neurone, pour ne retenir qu'une seule propriété physique, est une grandeur continue et sa valeur exacte est essentielle pour prédire le comportement du neurone. Un point connexe est que l'état complet d'un ordinateur passe à son successeur uniquement lorsque l'horloge centrale de l'ordinateur se met en marche, de sorte que tous ses événements internes s'alignent dans le temps comme une rangée de soldats. Mais les événements qui se déroulent dans le cerveau, comme les décharges neuronales, ne suivent pas ce rythme régulier. Un neurone peut émettre une impulsion électrique à tout moment, le temps étant bien sûr une autre variable continue. D'un point de vue mathématique, cette propriété à elle seule pourrait suffire à faire sortir la dynamique du cerveau du cadre des fonctions calculables par la machine de Turing. Le cerveau fonctionne sur un continuum, tant au niveau des activations des neurones eux-mêmes qu'au niveau temporel, de manière asynchrone ou synchrone. Ouais. Ouais. cela signifie-t-il que, théoriquement, le cerveau humain pourrait posséder des capacités intellectuelles tout simplement hors de portée de tout type de machine de tournée ? Bon, d'accord. Donc, mathématiquement et théoriquement, cela signifie certainement que cela pourrait appartenir à une classe de calculs qui est en dehors de la classe de calculs de Turing. Donc, c'est un fait mathématique. Mais d'un autre côté, nous savons aussi que lors des calculs, on peut simuler un système continu avec un degré de fidélité arbitraire. Il suffit d'ajouter des décimales aux nombres. On ne peut jamais obtenir un nombre infini de décimales, mais on peut toujours en ajouter. On peut donc simuler n'importe quel processus physique avec un degré de fidélité arbitraire. Mathématiquement, elles sont différentes. Ils restent différents. Mais cela a-t-il une importance fonctionnelle ? C'est vraiment ce que vous demandez. Je ne sais pas. Je ne le crois pas.

Je suppose que c'est parce qu'on peut l'imiter à un degré arbitraire, n'est-ce pas ? Ouais. Je veux dire par là que mon inclination serait de penser qu'il n'y a aucune limite à ce que l'on peut faire avec un ordinateur, en raison de ce fait mathématique. Dans votre livre sur l'incarnation, vous avez décrit tous les différents types d'architecture cérébrale que nous n'avons pas simulés. En gros, ce que nous avons fait avec l'architecture transformeur actuelle des LLM (modèles de langage génératifs à grande échelle), c'est simplement prendre l'architecture du neurone, du réseau neuronal, et l'étendre considérablement. Pourquoi le fait de n'imiter que cette partie du cerveau, et non toutes ces autres fonctionnalités, n'a-t-il pas permis de reproduire une si grande partie du comportement humain qui nécessitait le cerveau entier ? Eh bien, tout d'abord, vous savez, ce que nous avons aujourd'hui dans les réseaux neuronaux artificiels est très différent de ce que nous avons dans le cerveau. Donc, les neurones artificiels ne ressemblent pas vraiment aux vrais neurones. Voilà donc une mise en garde très importante : le type d'apprentissage qui se produit est très différent de celui qui se produit dans les vrais cerveaux. Donc, la question est encore plus intéressante d'une certaine manière. C'est même encore plus intéressant, d'une certaine manière. Ouais. Alors, comment se fait-il que, grâce à ce substrat très différent, nous ayons réussi à créer cette sorte d'extraordinaire similitude d'intelligence ? Le plus intéressant, c'est que nous ne le savons pas vraiment. En fait, non. C'est la situation bizarre dans laquelle nous nous sommes retrouvés, une situation que je n'aurais pas anticipée si vous m'aviez posé la question il y a 15 ans, lorsque j'ai écrit ce livre. À l'époque, nous construisions des choses, et comme nous les entraînions sur d'énormes quantités de données, il s'agissait de très grands ensembles de neurones, de très grands réseaux. Ensuite, tous les poids de ces réseaux changeaient au fil de l'entraînement, et comme par magie, ils finissaient par avoir une fonctionnalité extraordinaire. Mais nous ne comprenions pas vraiment comment ils fonctionnent à un niveau qui nous permette d'expliquer l'intelligence qui s'y cache, si tant est qu'il y en ait une. Peut-être puis-je poser la question d'une manière légèrement différente, comme nous l'avons fait avant la singularité : que nous apprend-elle sur l'intelligence humaine ? Car pendant très longtemps, le paradigme dominant de l'IA a été l'IA symbolique. Pour nos auditeurs, cela consiste à formuler des propositions logiques déclaratives qui s'enchaînent de diverses manières pour résoudre des problèmes. Ouais. Ce qui a fonctionné en revanche, ce sont ces boîtes noires, sortes de simulations grossières de réseaux neuronaux, mais même pas aussi précises que vous le pensez. Ouais. Cela permet de produire des comportements dont l'IA symbolique est encore très loin. Alors, est-ce là quelque chose que nous pouvons apprendre sur la conscience humaine, c'est-à-dire cela ? Les stoïciens adoptent une vision extrêmement intellectualiste de la nature humaine. Derrière la nature humaine se cache la série d'énoncés propositionnels auxquels nous adhérons. C'est ce qu'ils pensent être le moteur de la machine humaine. Oui. Mais peut-être que ce n'est pas ça, n'est-ce pas ? Peut-être que pour nous, ce qui est à la base est une architecture biologique beaucoup plus complexe. Ouais.

Les énoncés logiques ne sont que des manifestations superficielles de. Est-ce quelque chose que vous ? C'est tout à fait exact. Je pense que, pour le dire crûment, nous sommes probablement arrivés à la conclusion que la cognition humaine est implémentée sur un véritable fouillis de spaghettis. Elle n'a pas ce type de structure. Elle ne présente pas le type de structure que l'on attendrait intuitivement. C'est un vrai désastre. et je pense que dans mon propre parcours, vous savez, mon parcours intellectuel en tant que chercheur en IA, il y a eu une sorte de recul progressif par rapport à la volonté de construire des choses d'une manière intelligible, vous savez, où l'architecture est fondamentalement intelligible. Vous avez longtemps été du côté symbolique, pendant de nombreuses années. Ouais.

Je veux dire, moi aussi, je l'ai abandonné il y a très longtemps. Mais le plus intéressant, c'est que même lorsque je travaillais dans l'apprentissage automatique et les réseaux neuronaux, et que j'ai passé énormément de temps à essayer d'apprendre des représentations neuronales qui possédaient néanmoins une sorte de structure symbolique, vous vous accrochiez à cette vision des choses, à cette idée que les choses devaient avoir une structure compositionnelle comme le langage. Et bien, elles étaient apprises et réalisées sur le substrat neuronal. Mais néanmoins, à un certain niveau important, ils devaient avoir cette structure compositionnelle semblable à celle du langage, avec des objets, des prédicats, des propositions et des choses semblables, et une topographie en somme, n'est-ce pas ? Ouais. Mais j'ai dû progressivement me retirer de plus en plus de cette position, et même ainsi, je pense que vous vous retirez un peu plus et vous vous dites : "D'accord, arrêtons d'essayer de forcer l'architecture à avoir cette structure." Donc l'architecture durale, c'est bien ça ? Nous ne sommes donc pas encore sortis de l'IA classique et sophistiquée, de l'IA symbolique, mais j'essaie encore de forcer les représentations apprises au sein du substrat neuronal à avoir une sorte de structure symbolique. Bon, ça ne marche pas non plus. Alors peut-être vous dites-vous : bon, d'accord, laissons les choses se faire naturellement, n'est-ce pas ? Nous aurons donc une sorte de boîte noire complète. Ouais. Droite. Mais, mais, mais alors, on va sûrement trouver ces structures en regardant de plus près. Mais même ça, ça ne semble pas être le cas. On constate qu'en entraînant ces choses, on découvre tout un domaine d'interprétabilité mécaniste qui tente de comprendre ce qui se passe à l'intérieur. Eh bien, c'est un vrai casse-tête, ça reste un vrai casse-tête. On continue à regarder à l'intérieur, et bien sûr, ils ont fait beaucoup de progrès. On peut extraire des choses, mais elles ne ressemblent pas aux catégories intuitives que nous avions pour comprendre la cognition par le passé, en termes de langage, comme les représentations propositionnelles, etc. En un sens, toute votre vie a été une série de reculs. Non, pas seulement du côté de l'informatique, mais aussi, semble-t-il, du côté philosophique. Oh, je ne sais pas trop. N'êtes-vous pas revenu à la position de Wiganstein, après avoir été dualiste ? Oh, vous avez commencé... Eh bien... Je veux dire, je n'ai toujours pas échappé à la position dualiste, mais je sais... Là où je dois aller, je crois, mais voyez-vous, j'essaie d'aborder l'aspect métaphysique. J'imagine que tout enfant qui grandit s'intéresse à cela se dit : "Il doit y avoir une essence... Est-ce... ? Est-ce... ? Adams... ? Bien sûr.

Et puis... vous voyez le miroir, comme... se tourner vers Wickenstein tout en s'éloignant de l'IA symbolique, c'est une sorte d'abandon du contrôle, de manière mature, n'est-ce pas ? Oui, bien que je pense que c'est un peu différent avec la philosophie, car... avec la philosophie, je ne pense pas avoir encore trouvé où je voulais aller, mais je savais où je devais aller il y a très longtemps. Je vois... c'est-à-dire le bouddhisme et Wickenstein, car vous avez manifestement beaucoup réfléchi à ces questions, n'est-ce pas ? Et donc, bien sûr, on peut facilement faire ressurgir ces pensées dualistes en suivant ce genre de raisonnement... donc c'est un peu différent, n'est-ce pas ? Alors, je sais depuis longtemps quelle devrait être la destination, mais je ne l'ai pas encore tout à fait atteinte, avec la philosophie sur la façon de construire l'IA et de comprendre la cognition. Et cela a vraiment été une série de reculs, et mais c'est le progrès. Je veux dire, si vous changez d'avis, c'est apprendre. Rich Sutton, bien sûr, a décrit cela de façon très célèbre comme la "leçon amère", vous ne le savez pas ? Il a écrit un article intitulé "La leçon amère" où il explique que ce que nous avons appris, c'est qu'au fil des années, nous avons commencé à penser que nous voulions construire des choses que nous pouvions comprendre et que nous allions révéler. Et c'est ce qui rend les choses passionnantes. Nous allons comprendre ces choses que nous construisons, et c'est grâce à cette compréhension que nous pourrons les construire. Mais en réalité, nous avons dû renoncer à tout cela et réaliser que ce que

l'on fait, c'est utiliser l'apprentissage, la mise à l'échelle, la recherche et beaucoup de calculs. Donc, vous augmentez la taille des données d'entraînement, vous augmentez la puissance de calcul, vous augmentez la recherche, et ce sont ces tâches répétitives à grande échelle qui fonctionnent. En quoi est-ce une leçon amère ? Parce que, intellectuellement, il y a quelque chose de frustrant là-dedans. Ouais. Nous voulons que ce soit cette belle topographie, cette forme platonicienne, n'est-ce pas ? Mais oui, votre travail, votre travail intellectuel en tant que chercheur en IA, en tant que philosophe, pensez-vous que l'IA va remplacer une grande partie de ce travail ? Certes, elle est déjà augmentée, même dans mon propre travail, mais pensez-vous que l'IA va remplacer le travail intellectuel, oui, remplacer une grande partie du travail intellectuel que vous effectuez actuellement ? Eh bien, en ce qui concerne la philosophie, je veux dire, cela n'a aucun sens qu'elle la remplace, car c'est à vous de la pratiquer. Oui bien sûr. C'est comme faire courir un robot sur une piste d'athlétisme à votre place. Je veux dire, là, là, tu es littéralement inutile, n'est-ce pas ? Et d'ailleurs, je pense qu'on peut apporter des réponses similaires concernant l'utilisation de l'IA dans des contextes créatifs. Donc, l'important n'est pas la production, l'important c'est la culture. Bien sûr, bien sûr, pour certaines personnes, l'important c'est la production, car c'est leur gagne-pain. C'est donc une question très importante pour les professeurs de philosophie. Ouais. Eh bien, cela vaut également pour les professeurs de philosophie et les personnes travaillant dans les industries créatives. Et ce sont, vous savez, des questions importantes. Mais vous savez, si je fais mon projet artistique chez moi pour ma propre satisfaction, je ne vais pas simplement demander à une IA de le faire. Je suis juste là, vous savez, parce que je veux réaliser ce projet, c'est ça le but. C'est comme si je ne voulais pas faire courir un robot sur une piste à ma place, et de même, on ne peut pas faire en sorte qu'une IA fasse de la philosophie pour moi, car cela ne m'apportera aucune compréhension. Mais dans tous ces cas, on peut les utiliser comme outils pour nous aider. Qu'en est-il du point de vue de la recherche en IA ? J'imagine que, comparé à la philosophie, il s'agit beaucoup plus de production, n'est-ce pas ? Eh bien, c'est vrai. Donc, je pense que la recherche en IA, comme pour beaucoup de choses, va être augmentée et améliorée par l'IA, bien sûr. Mais pas seulement. Il y a aussi le potentiel de peut-être remplacer... Oui, en effet. Bien sûr, je vois. Et qui sait où cela nous mènera ? Si vous avez des enfants, comment envisagez-vous leur éducation aujourd'hui ? Quel type de compétences leur inculqueriez-vous ? Excellente question. Je pense que nous devons apprendre aux gens à vivre. La philosophie, je veux dire... enfin, pas forcément le genre de philosophie dont nous avons parlé ici, mais vous savez... Une vie réussie, c'est une question philosophique différente. Je me souviens avoir lu il y a quelques années un article de John Maynard Kanes intitulé "Les possibilités économiques pour nos petits-enfants", écrit en 1930, où il imagine un avenir utopique, potentiellement un avenir d'abondance, où les défis économiques sont surmontés et où les gens peuvent mener une vie de loisirs. Mais il pose alors la question : que faire, en réalité ? Comment mener une vie réussie dans ces circonstances, lorsque certains aspects essentiels de notre existence nous sont enlevés ? Je pense que ce sont d'excellentes questions. Personnellement, je n'ai pas de réponses à ces questions. Je vois.

Eh bien, merci pour cette interview passionnante, professeur. C'était incroyable. Merci beaucoup. Oui, j'ai beaucoup apprécié. Merci d'avoir regardé mon interview. Si ce genre de discussions vous intéresse, je pense que vous vous intégrerez parfaitement à l'écosystème que nous sommes en train de construire chez Cosmos. Nous finançons la recherche, incubons et investissons dans des startups spécialisées en IA et croyons que la philosophie est essentielle au développement technologique. Si vous souhaitez rejoindre notre écosystème de bâtisseurs philosophes, vous trouverez les postes à pourvoir, les événements que nous organisons et d'autres façons de vous impliquer sur jona-

thanb.com/cosmos. Merci.