gros objet conducteur émergeant de l'eau. Elles réagissent alors par une attaque explosive. Ce comportement littéralement choquant m'est apparu lorsque j'ai tenté de capturer une grosse anguille à l'aide d'une épuisette dont le bord et la poignée étaient en métal. Elle s'est retournée et a bondi hors de l'eau tandis que sa mâchoire inférieure restait appuyée contre la poignée métallique et qu'elle émettait une longue série de décharges de forte tension (heureusement, je portais des gants de protection). C'est un comportement défensif que j'ai retrouvé chez toutes les anguilles électriques testées.

En analysant les phénomènes électriques qui ont accompagné le bond de mon anguille, de nombreuses pièces du puzzle biologique et historique se sont mises en place. Si les anguilles interprètent les petits objets conducteurs comme des proies comestibles, elles interprètent comme de gros animaux menaçants les grands objets conducteurs partiellement submergés qui s'approchent. *A priori*, elles pourraient fuir à la nage, mais il se trouve que pendant la saison sèche, elles se voient souvent piégées dans des mares à la portée de prédateurs tels que caïmans ou félins. Ajoutez à ce scénario le fait que les anguilles ne peuvent pas diriger leurs impulsions électriques et vous aurez la recette d'un étonnant mode de défense expliquant le récit de von Humboldt.

DES TESTS AVEC DES BRAS DE ZOMBIE

Pour en savoir plus, je me suis procuré des bras de zombie postiches, que l'on vend aux États-Unis lors des fêtes d'Halloween. Après en avoir retiré le faux sang, j'y ai placé des diodes électroluminescentes de façon à imiter le passage des voies nerveuses, puis j'ai présenté cette nouvelle cible à mes anguilles. Cela a déclenché des bonds défensifs, et les diodes ont d'autant plus clignoté que l'anguille s'élevait loin de l'eau. Comment expliquer cela ?

Répondre à cette question implique de déterminer le circuit électrique équivalent à l'ensemble gymnote+gros objet conducteur, afin de pouvoir déterminer la tension, ou force électromotrice, de l'organe électrique. Je devais également calculer la résistance électrique des matériaux du circuit. J'ai mesuré chaque variable successivement, en commençant par l'organe électrique de l'anguille. Longue d'un peu plus d'un mètre, la plus grosse anguille de mon laboratoire avait un potentiel électrique de 382 volts et une résistance interne de seulement 450 ohms, ce qui, en l'absence d'autres résistances, donne des intensités de courant de près de un ampère. Clairement, une anguille électrique est plus violente qu'un Taser !

Lorsque l'anguille bondit hors de l'eau en appuyant sa mâchoire inférieure contre une cible, le courant qui circule habituellement de la tête vers la queue s'affaiblit, car l'air, mauvais conducteur, tend à l'interrompre, de sorte que c'est le courant à travers la cible qui augmente. Cela explique les impressionnants bonds du gymnote : plus ils sont hauts, plus l'épaisseur d'air isolant la tête de la queue est grande, et plus l'intensité du courant dans la cible est importante.

Pour préciser les choses, j'ai dû résoudre un exercice de physique élémentaire : calculer le courant électrique dans un circuit comportant deux résistances en parallèle. Il fallait pour cela déterminer chacune de ces résistances. J'ai obtenu la première – celle du corps de l'anguille – en faisant attaquer des gymnotes une plaque métallique reliée à un voltmètre. L'autre résistance à déterminer était celle du bras postiche servant de cible. Après avoir précisé toutes les autres variables du problème, je suis parvenu à estimer la résistance électrique complexe s'établissant entre la mâchoire de l'anguille électrique, une cible vivante et l'eau baignant l'ensemble.

Je me suis aussi servi de mon propre bras afin de déterminer la dernière variable et de pouvoir tester les résultats de toutes les mesures précédentes. J'ai fait appel à une toute petite anguille présentant une force électromotrice de seulement 198 volts et une résistance interne de 960 ohms, et construit un dispositif capable de mesurer le courant à travers mon bras pendant l'attaque du gymnote. C'est ainsi que j'ai résolu l'exercice de physique que je m'étais posé et éprouvé dans ma chair à quel point les anguilles électriques sont efficaces quand elles attaquent… ■

TROIS ESPÈCES D'ANGUILLES ÉLECTRIQUES

Autour de David de Santana, de la Smithsonian Institution aux États-Unis, une équipe de chercheurs a étudié l'anatomie et l'ADN de 107 anguilles électriques capturées au Brésil, au Suriname, en Guyane française et au Guyana. Les résultats révèlent l'existence de trois populations génétiquement distinctes, donc de trois espèces au lieu d'une seule reconnue précédemment, *Electrophorus electricus*. Celle-ci vit au nord, principalement au Guyana et au Suriname ; *Electrophorus varii* est disséminé le long des basses terres amazoniennes jusqu'au nord du Brésil et en Guyane française ; et *Electrophorus voltai* vit au sud du bassin amazonien, au Brésil.

Les chercheurs ont mis en évidence des différences anatomiques subtiles entre ces espèces, qui se seraient différenciées avec la séparation des trois populations, quand le grand bassin amazonien s'est formé il y a quelque 3 millions d'années. Ils ont aussi établi que *E. voltai* peut produire des tensions de 860 volts, bien plus que les 650 volts cités jusqu'à présent pour les anguilles électriques. C'est un record absolu dans le monde vivant.

BIBLIOGRAPHIE

C. David de Santana *et al.*, **Unexpected species diversity in electric eels with a description of the strongest living bioelectricity generator,** *Nature Communications*, vol. 10, article 4000, 2019.

K. C. Catania, **Power transfer to a Human during an electric eel's shocking leap,** *Current Biology*, vol. 27(18), pp. 2887-2891, 2017.

K. C. Catania, **Electric eels use high-voltage to track fast-moving prey,** *Nature Communications*, vol. 6, article 8638, 2015.

K. C. Catania, **The shocking predatory strike of the electric eel,** *Science*, vol. 346, pp. 1231-1234, 2014.

L'ESSENTIEL

> Deux théories s'opposent sur la nature de la conscience : celle de « l'espace de travail neuronal global » et celle de « l'information intégrée ».

> L'une implique que les machines deviendront conscientes lorsqu'elles auront des capacités suffisamment sophistiquées, l'autre non.

> Ces théories sont actuellement testées par une collaboration internationale.

L'AUTEUR

CHRISTOF KOCH
directeur scientifique et président de l'institut Allen pour les sciences du cerveau, à Seattle, aux États-Unis

Les machines peuvent-elles être conscientes?

Nul doute que les robots et ordinateurs deviendront de plus en plus « intelligents ». Mais auront-ils une subjectivité et un sentiment d'exister ? Cette question est autrement plus débattue.

Dans les décennies à venir, les progrès rapides des algorithmes d'apprentissage engendreront des machines d'une intelligence comparable à la nôtre. Capables de parler et de raisonner, elles auront leur place dans une myriade de domaines tels que l'économie, la politique et, inévitablement, la guerre. La naissance d'une véritable intelligence artificielle (IA) aura un impact profond sur l'avenir de l'humanité et conditionnera l'existence même d'un tel avenir.

Prenez par exemple la citation suivante : « Encore aujourd'hui, des recherches sont en cours pour mieux comprendre ce que les nouveaux programmes d'IA seront capables de faire, tout en restant dans les limites de l'intelligence d'aujourd'hui. La plupart des programmes d'IA actuellement programmés se limitent principalement à prendre des décisions simples ou à effectuer des opérations simples sur des quantités relativement faibles de données. »

Peut-être avez-vous eu l'impression que quelque chose clochait dans ce paragraphe ? Cette citation est l'œuvre de GPT-2, un robot linguistique que j'ai testé l'été dernier. Développé par OpenAI, un institut basé à San Francisco qui promeut les IA « vertueuses », GPT-2 est un algorithme d'apprentissage fondé sur un réseau de neurones artificiels. Ses entrailles contiennent plus d'un milliard de connexions simulant des synapses, les points de jonction entre neurones.

IL NE COMPREND RIEN, ET POURTANT…

La tâche du réseau est apparemment stupide : confronté à un texte de départ arbitraire, il doit prédire le mot suivant. Il ne « comprend » pas les textes comme le ferait un humain. Mais durant sa phase d'apprentissage, il a dévoré des quantités astronomiques de textes – huit millions de pages >

> internet en tout – et ajusté ses connexions internes pour mieux anticiper des suites de mots.

J'ai écrit les premières phrases de l'article que vous êtes en train de lire, puis je les ai « injectées » dans l'algorithme en lui demandant de composer une suite. Il a notamment craché le paragraphe cité. Certes, ce texte ressemble aux efforts d'un étudiant de première année pour se rappeler un cours d'introduction à l'apprentissage automatique durant lequel il aurait rêvassé. Mais le résultat contient malgré tout les mots et les phrases clés – pas si mal, vraiment !

Les successeurs de ces robots risquent de déclencher un raz-de-marée de faux articles et reportages, qui viendront polluer internet. Ce ne sera qu'un exemple de plus de programmes exécutant des prouesses que l'on croyait réservées aux humains : jouer à des jeux de stratégie en temps réel, traduire du texte, recommander des livres et des films, reconnaître les gens sur des images ou des vidéos…

Les algorithmes sauront-ils un jour écrire un chef-d'œuvre aussi abouti que *À la recherche du temps perdu*, de Marcel Proust ? Difficile à dire, mais les prémices sont là. Rappelez-vous comme les premiers logiciels de traduction et de conversation étaient faciles à tourner en dérision, tant ils manquaient de finesse et de précision. Mais avec l'invention des réseaux neuronaux profonds et la mise en place d'infrastructures de calcul surpuissantes par les entreprises du numérique, les ordinateurs se sont améliorés sans relâche, jusqu'à ce que leurs productions n'aient plus rien de ridicule.

Comme on le voit avec le jeu de go, les échecs et le poker, les algorithmes d'aujourd'hui sont capables de surpasser les humains (en novembre 2019, Lee Sedol, un des plus grands joueurs de go de l'histoire, a décidé de prendre sa retraite après avoir perdu plusieurs fois contre l'algorithme AlphaGo ; il a déclaré que celui-ci était une entité ne pouvant plus être vaincue). Si bien que notre rire se fige : sommes-nous comme l'apprenti sorcier de Goethe, ayant invoqué des esprits serviables que nous ne pouvons plus contrôler ?

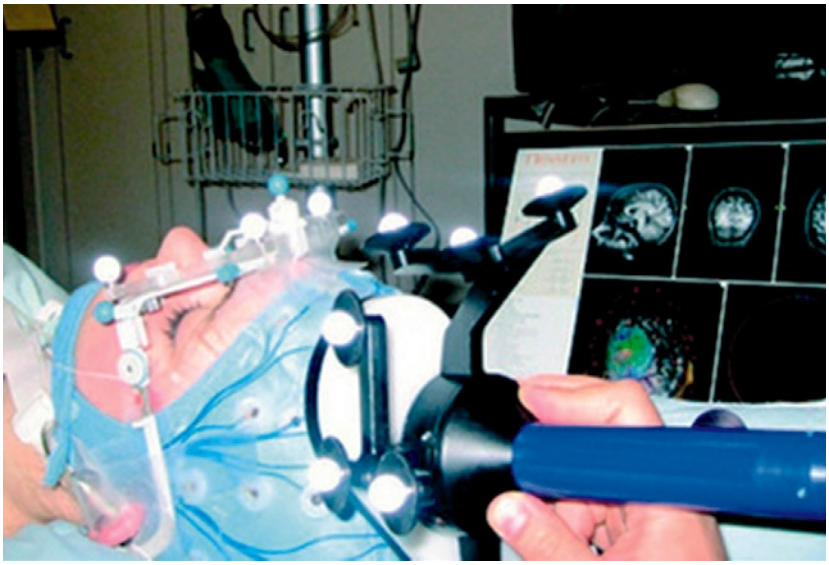
L'INTELLIGENCE N'EST PAS LA CONSCIENCE

Bien que les experts ne s'accordent pas sur la nature exacte de l'intelligence, la plupart d'entre eux estiment que, tôt ou tard, les ordinateurs atteindront ce que l'on appelle l'intelligence artificielle générale dans le jargon (c'est-à-dire qu'ils deviendront capables de réaliser n'importe quelle tâche intellectuelle accessible à un humain). Mais auront-ils pour autant une forme de sentiment d'eux-mêmes ? En d'autres termes, pourront-ils un jour être conscients ?

Par « conscience » ou « sentiment subjectif », j'entends une qualité inhérente à toute expérience – par exemple, la saveur d'une madeleine, une rage de dents, l'étirement du temps quand on s'ennuie ou le sentiment de vitalité et d'anxiété juste avant une compétition. En paraphrasant le philosophe Thomas Nagel, auteur d'un article titré *Quel effet cela fait-il d'être une chauve-souris ?*, nous pourrions dire qu'un système est conscient si cela « fait quelque chose » *d'être* ce système.

Considérez l'embarras que vous ressentez lorsque vous vous apercevez que vous venez de commettre une gaffe, parce que votre interlocuteur a perçu comme une insulte ce qui n'était pour vous qu'un trait d'humour. Les ordinateurs éprouveront-ils un jour une telle émotion ? Et lorsque vous êtes au téléphone depuis un temps interminable, écoutant une voix synthétique répéter en boucle : « Nous sommes désolés de vous faire attendre », le logiciel pourra-t-il se sentir gêné de vous maintenir dans cet enfer du service client ?

Il ne fait guère de doute que notre intelligence et nos expériences résultent des capacités naturelles de notre cerveau, fondées sur des enchaînements de causes et d'effets. Cette prémisse a extrêmement bien servi la science au cours des derniers siècles. Le cerveau humain, cet organe d'à peine un kilo et demi à la texture comparable au tofu, est de loin le morceau de matière active organisée le plus complexe connu de l'Univers. Mais il obéit aux mêmes lois physiques que les pierres, les arbres et les étoiles. Rien n'échappe à ces lois. Nous ne comprenons pas encore tout à fait les mécanismes causaux en jeu, mais nous les expérimentons tous les jours : un groupe de neurones s'allume pendant que vous voyez des couleurs, tandis que l'activation de cellules nerveuses dans une région voisine du cortex est associée à une humeur joviale. Lorsqu'un neurochirurgien stimule ces neurones à l'aide d'une



La théorie de l'information intégrée a inspiré la construction d'un appareil pour mesurer le niveau de conscience chez des patients incapables de communiquer. Il est ici testé par l'équipe de Steven Laureys, au Coma science group de l'université et du CHU de Liège.

© Coma Science group

LA FRANCE EST-ELLE CONSCIENTE ?

Selon la théorie de l'information intégrée, le degré de conscience d'un système dépend de son pouvoir causal intrinsèque – qui quantifie la capacité de ses sous-systèmes à s'influencer mutuellement. Mais alors, tout ensemble doté de parties en interaction n'est-il pas conscient ? N'est-ce pas le cas par exemple de l'ensemble que vous formez avec votre interlocuteur au téléphone, voire de celui formé par la population d'un pays entier ? Non, car la théorie postule qu'à l'intérieur d'un système, la conscience apparaît précisément là où le pouvoir causal intrinsèque est maximum. Si vous parlez avec quelqu'un, vous allez vous influencer mutuellement, mais de façon très inférieure à ce qui se passe dans le cerveau de chacun. Il n'y aura donc pas formation d'un « superesprit » – pour cela, il faudrait augmenter drastiquement les connexions entre vos deux cerveaux. Il en va de même pour un pays comme la France et ses 67 millions d'habitants : même s'il y a de multiples interactions entre ces derniers, le pouvoir causal intrinsèque reste maximal à l'intérieur du cerveau de chacun. C'est donc là que surgit la conscience.

électrode, le sujet voit des couleurs dans le premier cas, éclate de rire dans le second. Inversement, quand le cerveau est sous anesthésie, ces expériences disparaissent.

Compte tenu de cela, l'évolution de l'intelligence artificielle débouchera-t-elle sur une conscience artificielle?

DEUX THÉORIES ANTAGONISTES

Cette question mène à deux possibilités fondamentalement différentes. Selon la première, les machines vraiment intelligentes seront sensibles et dotées de conscience: elles parleront, raisonneront, s'occuperont d'elles-mêmes, seront capables d'introspection... Une idée dans l'air du temps, comme on le voit dans des romans et des films tels que *Blade Runner*, *Her* ou *Ex Machina*.

D'un point de vue scientifique, cette hypothèse s'appuie sur l'une des principales théories de la conscience, dite de «l'espace de travail neuronal global» (GNW, pour *global neuronal workspace*). Cette théorie considère que l'origine de la conscience réside dans certaines caractéristiques architecturales du cerveau. Elle s'inspire en cela de l'«architecture du tableau noir», développée par les sciences informatiques dans les années 1970: le principe est que des programmes spécialisés accèdent à un répertoire commun d'informations, appelé tableau noir ou espace de travail central. Les psychologues ont postulé qu'un espace de ce type existe dans le cerveau et qu'il est essentiel à la cognition humaine. Sa capacité est faible, de sorte qu'à un moment donné, il ne contient qu'une seule perception, qu'une pensée unique ou qu'un seul souvenir. Dans cet espace de travail, les nouvelles informations entrent en concurrence avec les anciennes et les remplacent.

Le neuroscientifique Stanislas Dehaene et le biologiste moléculaire Jean-Pierre Changeux, tous deux au Collège de France, à Paris, ont transposé ces idées à l'architecture du cortex, la couche externe du cerveau. Ils ont postulé que l'espace de travail repose sur un réseau de neurones dits «pyramidaux», qui relient des régions corticales éloignées, en particulier les zones associatives préfrontales, pariétotemporales et médianes (cingulaires).

Une grande partie de l'activité cérébrale reste localisée et de ce fait n'accède pas à la conscience. C'est le cas des modules neuronaux qui contrôlent la posture du corps ou la direction du regard. Mais lorsque l'activité d'une ou plusieurs régions dépasse un seuil critique – disons, lorsqu'on présente à quelqu'un l'image d'une friandise appétissante –, elle déclenche une vague d'excitation neuronale qui se propage à travers l'espace de travail, dans tout le cerveau. Ce signal devient alors accessible à une multitude de processus auxiliaires, tels que le langage, la planification, le circuit de la récompense, la mémoire à long terme et le stockage dans une mémoire tampon à court terme.

Ce serait le fait de diffuser cette information à l'échelle globale qui la rendrait consciente. Ainsi, l'expérience subjective de la friandise appétissante résulterait de l'action de neurones pyramidaux qui ordonnent à la région motrice du cerveau de se saisir de la friandise, tandis que d'autres modules neuronaux anticipent une récompense sous la forme d'une poussée de dopamine, causée par le petit *shoot* de sucre et matières grasses à venir.

En d'autres termes, les états conscients découleraient de la façon dont l'algorithme de >

1 milliard de synapses virtuelles: c'est ce que contient l'architecture du logiciel linguistique GPT-2

> l'espace de travail traite les entrées sensorielles, les sorties motrices et les variables internes liées à la mémoire, à la motivation et aux attentes. La conscience, ce serait ce traitement global. Pour les tenants de cette théorie, la conscience artificielle n'est pas exclue.

LA CONDITION DE LA CONSCIENCE

La seconde hypothèse repose sur la théorie de l'information intégrée (IIT, pour *integrated information theory*), qui adopte une approche plus fondamentale pour expliquer la conscience. Giulio Tononi, psychiatre et neuroscientifique à l'université du Wisconsin-Madison, en est le promoteur principal, avec d'autres collaborateurs, dont moi-même. Le principe est de partir de la nature de l'expérience consciente et d'en déduire les propriétés du système physique qui en constitue le support. Par exemple, chaque expérience forme un tout unique – lorsque je vois une image, je n'ai pas deux expériences conscientes juxtaposées, une de la partie gauche de mon champ visuel et une de la partie droite, mais une seule – et cela doit se traduire d'une façon ou d'une autre dans le système physique sous-jacent.

En particulier, l'information doit y être traitée de façon intégrée, c'est-à-dire que les éventuels sous-systèmes doivent interagir et s'influencer mutuellement. On parle de « pouvoir causal intrinsèque » et on quantifie l'importance de ces influences mutuelles par une mesure mathématique que nous appelons l'« information intégrée ».

Ainsi, le pouvoir causal intrinsèque n'est pas une notion floue et éthérée, mais une quantité susceptible d'être évaluée avec précision. On peut la déterminer pour tout type de mécanisme, par exemple un neurone qui émet des potentiels d'action affectant les cellules connectées en aval (*via* des synapses), ou un circuit électronique composé de transistors, de résistances et de fils. Elle est d'autant plus élevée qu'un système est capable d'influencer, à partir de son état actuel, ses propres valeurs d'entrée et de sortie. L'état de ce système est alors marqué par son passé et porteur de son avenir.

Notre théorie stipule que tout mécanisme doté d'un tel pouvoir est conscient. Plus un système a une valeur importante d'information intégrée – représentée par la lettre grecque Φ (prononcée « fi ») –, plus il est conscient. En revanche, s'il n'a aucun pouvoir causal intrinsèque, son Φ est nul et il ne ressent rien.

Au sein du cortex, la quantité d'information intégrée est énorme, car on y trouve un vaste réseau de neurones hétérogènes et étroitement interconnectés (qui s'influencent donc mutuellement). La théorie a inspiré la construction d'un appareil de mesure de la conscience, en cours d'évaluation clinique (*voir la photo page 46*). Cet instrument serait précieux pour déterminer si certains patients sont conscients mais incapables de communiquer, ou totalement inconscients – par exemple dans les états végétatifs persistants, l'anesthésie et le *locked-in syndrome*.

Pour les ordinateurs programmables, en revanche, la situation est différente, comme l'ont montré des travaux qui ont analysé leur structure à l'échelle des composants métalliques – les transistors, fils et diodes qui servent de substrat physique à tout calcul. Leur pouvoir causal intrinsèque et leur Φ sont infimes, notamment parce que leur connectique est très différente de celle du cerveau : chaque composant n'est connecté qu'à quelques autres (là où un seul neurone forme, en moyenne, 10000 synapses), et il y a bien moins de rétroactions que dans nos systèmes neuronaux. Cela reste vrai quel que soit le logiciel qui tourne sur le processeur : que ce logiciel calcule les impôts ou simule le cerveau, cela ne change rien.

Dans le film *Ex Machina*, un robot humanoïde nommé Ava s'éveille à la conscience. Un scénario réaliste ?

Ex_Machina © Universal Pictures, 2014 (capture d'écran).

CE QU'ON EST, ET NON CE QU'ON FAIT

Le point clé est que deux réseaux qui effectuent la même opération d'entrée-sortie, mais dont les circuits sont configurés différemment, sont susceptibles de posséder des quantités différentes d'information intégrée. La grandeur Φ peut être très faible pour un circuit et très élevée pour l'autre. Bien qu'ils soient identiques de l'extérieur, l'un des réseaux expérimente quelque chose, tandis que son homologue zombie ne ressent rien. La différence se situe sous le capot, dans le câblage interne. En un mot, la conscience résulte de ce qu'on *est*, et non de ce qu'on *fait*.

La théorie de l'espace de travail neuronal global et la théorie de l'information intégrée

ont certaines conséquences drastiquement opposées. Imaginons par exemple que l'on simule le connectome – le câblage synaptique exact de tout le système nerveux. Les anatomistes ont déjà cartographié les connectomes de quelques vers. Ils travaillent actuellement sur celui de la drosophile, la petite mouche qui tourne autour des fruits trop mûrs, et prévoient de s'attaquer à la souris au cours de la prochaine décennie. Supposons qu'à l'avenir il soit possible de scanner un cerveau humain entier, avec ses près de 100 milliards de neurones et ses 100 000 milliards de synapses, après la mort de son propriétaire, puis de simuler son fonctionnement sur un ordinateur avancé, peut-être une machine quantique. Si le modèle est suffisamment fidèle, cette simulation se réveillera et se comportera comme un simulacre numérique de la personne décédée. Ainsi, elle parlera et accédera à ses souvenirs, à ses envies, à ses peurs et aux autres traits de sa personnalité. Mais sera-t-elle consciente?

Si l'imitation du fonctionnement du cerveau est tout ce qui compte pour créer la conscience, comme c'est postulé par la théorie de l'espace de travail neuronal global, la personne simulée sera consciente, réincarnée dans un ordinateur. Ce qui n'est pas sans rappeler l'un des fantasmes favoris de la science-fiction: télécharger son connectome dans le *cloud* pour continuer à vivre dans l'au-delà numérique…

Mais la théorie de l'information intégrée propose une interprétation radicalement différente de cette situation: le simulacre ressentira la même chose que n'importe quel logiciel basique – c'est-à-dire rien. Il agira comme une personne, mais sans aucun sentiment. Il ne sera qu'une sorte de zombie (sans le moindre goût pour la chair humaine, rassurez-vous).

Selon cette théorie, les pouvoirs causaux intrinsèques du cerveau sont nécessaires pour créer la conscience. Et ils ne peuvent être simulés: ils doivent faire partie intégrante du système physique sous-jacent.

C'est un peu comme si vous simuliez une pluie diluvienne ou un trou noir sur un ordinateur: vous ne risquez ni d'être mouillé ni d'être avalé par une énorme force gravitationnelle. Tout simplement parce qu'une simulation n'a pas le pouvoir causal de condenser la vapeur d'eau ou de courber l'espace-temps.

Il n'y a toutefois rien de magique dans le cerveau humain et, en principe, il serait possible de créer une conscience équivalente à la nôtre en allant au-delà d'une simple simulation. Il faudrait alors construire du matériel dit «neuromorphe», fondé sur une architecture similaire à celle du système nerveux.

L'éventuel ressenti d'un connectome simulé n'est pas la seule différence entre les deux théories. Celles-ci ne localisent pas au même endroit le substrat physique de certaines expériences conscientes spécifiques – l'épicentre de ce substrat étant à l'arrière du cortex pour l'une, à l'avant pour l'autre. Cette prédiction et d'autres sont actuellement testées dans le cadre d'une collaboration à grande échelle, impliquant six laboratoires aux États-Unis, en Europe et en Chine, et qui vient de recevoir un financement de 4,5 millions d'euros de la fondation Templeton World Charity.

Si l'imitation des fonctions cérébrales suffit pour créer la conscience, la simulation numérique d'un cerveau créera une sorte de réincarnation de la personne

DES SUJETS À PART ENTIÈRE

La question d'une éventuelle sensibilité des machines importe aussi pour des raisons éthiques. Si les ordinateurs font l'expérience de la vie par leurs propres sens, ils cessent d'être un simple moyen mis au service d'un but déterminé par les humains. Ils doivent être respectés pour eux-mêmes.

Selon la théorie de l'espace de travail global, les machines intelligentes pourraient bien passer de simples objets à sujets à part entière – chacune existerait comme un «je», doté de son propre point de vue. Une idée que l'on retrouve dans les séries télévisées *Black Mirror* et *Westworld*. Et lorsque les ordinateurs auront atteint des capacités cognitives rivalisant avec celles de l'humanité, ils feront valoir leurs droits juridiques et politiques – le droit de ne pas être jetés, de ne pas voir leur mémoire effacée, d'éviter la douleur et la dégradation. L'alternative, incarnée par la théorie de l'information intégrée, est qu'aussi perfectionnés soient-ils, les ordinateurs resteront des coquilles vides, semblables à des fantômes. Il leur manquera ce à quoi nous attachons le plus de valeur: le sentiment de la vie elle-même. ■

BIBLIOGRAPHIE

C. Koch, **The Feeling of Life Itself : Why Consciousness Is Widespread but Can't Be Computed**, MIT Press, 2019.

L. Naccache, **Observer la conscience**, *Pour la Science*, n° 500, pp. 75-80, juin 2019.

C. Koch, **Mesurer la conscience est enfin possible**, *Pour la Science*, n° 483, pp. 26-34, janvier 2018.

S. Dehaene *et al.*, **What is consciousness, and could machines have it ?**, *Science*, vol. 358, pp. 486-492, 2017.

UN

NE MEURT JAMAIS.

EN TRIANT VOS JOURNAUX, MAGAZINES, CARNETS, ENVELOPPES, PROSPECTUS ET TOUS VOS AUTRES PAPIERS, VOUS AGISSEZ POUR UN MONDE PLUS DURABLE. PLUS D'INFORMATIONS SUR LE RECYCLAGE SUR TRIERCESTDONNER.FR

CITEO

Donnons ensemble une nouvelle vie à nos produits