

> 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 événements 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

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



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.

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 accèdera 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

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

expériences conscientes spécifiques – l'épicode 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.

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.