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