

> 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  $\Phi$  (prononcée «fi») –, plus il est conscient. En revanche, s'il n'a aucun pouvoir causal intrinsèque, son  $\Phi$  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  $\Phi$  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 ?

## 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  $\Phi$  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

