Une science de l'humain?

Stephen N. Lyle

January 1, 2022

Commençons par une question: la science peut-elle comprendre les humains? Evidemment, nous avons déjà deux sciences de l'humain: la médecine et la psychologie. La première est la science du corps et la seconde la science de l'esprit. Car c'est bien comme ça que nous nous comprenons—le corps et l'esprit—comme si les deux choses étaient vraiment séparées. C'est un dualisme naturel. Mais aujourd'hui, des tentatives sont faites pour unifier cela et créer une science qui expliquerait tout en même temps. L'idée serait de commencer par ce qu'on peut voir et mesurer—la matière, la façon dont elle est organisée et la façon dont elle bouge—et comprendre l'esprit sans faire appel à d'autres ingrédients.

Alors, qu'est-ce qui pourrait constituer une telle science? On pense par exemple aux sciences cognitives: l'étude interdisciplinaire et scientifique de l'esprit. Le travail de l'esprit est la cognition, l'ensemble de tous ces processus mentaux liés à la compréhension, la connaissance, la mémoire, le langage, le raisonnement, l'apprentissage, l'intelligence, la résolution de problèmes, la prise de décision, la perception et l'attention. Et probablement plus. Donc c'est déjà une grande partie de ce que fait couramment un humain. Mais tout cela tourne autour du cerveau. Et le reste du corps? Les émotions? Nous savons que les émotions sont intimement impliquées dans la prise de décision, et le corps dans son ensemble est également intimement impliqué dans notre façon de penser. En effet, tout ce que nous pensons est en fin de compte *pour* notre corps.

Ainsi, lorsque nous faisons des sciences cognitives, nous devons voir le cerveau comme une partie du corps, bien qu'il s'agisse d'un élément clé, surtout si ce que nous visons réellement ici est une étude interdisciplinaire et scientifique de l'ensemble de l'être humain—corps et âme, pour ainsi dire—après tout, c'est ça l'unification. Visons donc quelque chose de plus grand qu'une simple étude du cerveau et voyons ce qui se passe.

Une très brève histoire

Ce sera en effet très bref, et commencera assez arbitrairement avec l'image de la figure 1. Vers la fin du XIXe siècle, la science cherchait à voir de quoi était fait un cerveau. Il n'était pas facile de bien distinguer les détails. La figure 2 montre quelqu'un qui travaillait à fond dans cette entreprise. Après avoir obtenu son diplôme de médecin, Sigmund Freud a passé plusieurs années à dessiner des images comme celle que vous voyez sur la figure 1, et a même contribué à certaines techniques pour faciliter de telles observations.

Mais ce n'est pas Freud qui a fait la grande découverte. C'est le neuroscientifique espagnol Santiago Ramón y Cajal qui a finalement montré que le cerveau et le système nerveux sont constitués de cellules, bien qu'il ne soit pas le seul dans cette découverte. Les efforts de beaucoup avant lui, dont Freud, ont été essentiels pour fournir les bases nécessaires, comme c'est toujours le cas dans l'entreprise scientifique. Freud a quitté ce domaine pour aller inventer la psychanalyse. Il a toujours dit, cependant, que nous ne comprendrons jamais la psychologie humaine sans d'abord déchiffrer le fonctionnement interne du cerveau. Mais la psychanalyse était peut-être ce qu'on pouvait faire de mieux à

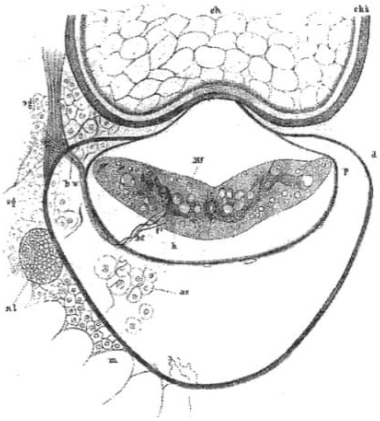

Figure 1: Cellules nerveuses dans la moelle épinière d'une lamproie. Dessin datant de 1877 [1]

l'époque.

Juste un mot en passant sur la lamproie, cette petite créature marine que vous voyez dans la figure 3. Il était bien sûr plus facile d'acceder au système nerveux d'une lamproie qu'à un cerveau humain, et son système nerveux était aussi un peu plus simple, ou du moins, un peu moins complexe. Inutile de dire que d'autres créatures ont été sacrifiées dans la quête pour comprendre le fonctionnement du cerveau, soulevant une importante question éthique. La lamproie est également bien connue dans la région bordelaise où elle joue un rôle culinaire qui pour beaucoup soulève moins de questions éthiques.

Outre les premières explorations du cerveau, un autre mouvement important à la fin du XIXe siècle a été le béhaviorisme. Ce mouvement essayait de répondre à la question suivante: comment comprendre le comportement des humains ou

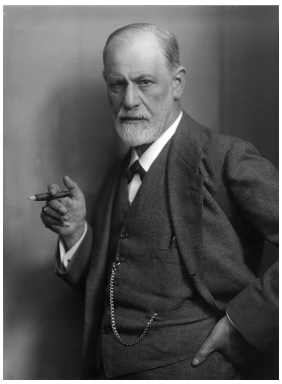

Figure 2: Photo de Sigmund Freud prise par son neveu Max Halberstadt

Figure 3: La lamproie, représentante moderne d'une ancienne lignée de poissons sans mâchoire

d'autres animaux quand on n'a pas d'outils adéquats pour voir ce qui se passe dans leurs cerveaux, par exemple? La spéculation sur des 'états mentaux' était problématique pour la méthode scientifique. N'était-il pas préférable pour les psychologues et les éthologues de focaliser sur des phénomènes observables et des comportements mesurables? Bien sûr, le problème était de produire des hypothèses scientifiques qui pouvaient être mises à l'épreuve dans des expériences.

Selon les tenants du béhaviorisme, les comportements observables peuvent être conditionnés de deux manières:

- par des réponses réflexes aux stimuli,
- par l'histoire des interactions de l'individu avec son environnement, c'est-à-dire par l'apprentissage.

Mais l'animal lui-même, et en particulier son cerveau, était traité comme une boîte noire, faute d'avoir une meilleure solution, et la spéculation sur des états mentaux inobservables était interdite.

Le béhaviorisme a beaucoup apporté à notre compréhension du monde animal. Cependant, ses limites ont fini par se révéler. Au cours des premières décennies du vingtième siècle, il est devenu évident que les animaux de laboratoire apprennent même lorsqu'il n'y a pas de 'conditionnement', c'est-à-dire même lorsqu'ils ne sont ni récompensés pour leurs réussites ni punis pour leurs échecs. Les animaux apprennent tout le temps d'une manière ou d'une autre. En effet, on commençait à les percevoir plutôt comme des systèmes de traitement d'informations, ce qui nous semble aujourd'hui une évidence. Les données sont collectées par les sens, traitées par le corps de l'animal, et en particulier son cerveau, et utilisées pour produire sa réaction à n'importe quelle situation dans laquelle il se trouve. Cette idée, que la cognition n'est fondamentalement qu'un traitement de données, est au cœur des sciences cognitives aujourd'hui.

Un autre point de repère dans notre histoire a été le travail dans les années 1920 et 1930 par des mathématiciens comme Alan Turing (Fig. 4). Turing pensait aux algorithmes. La notion d'algorithme n'avait rien de nouveau. Le nom vient du mathématicien arabe du neuvième siècle Muhammad ibn Mūsā al-Khwārizmī, et ces choses étaient connues des Babyloniens, des Egyptiens et

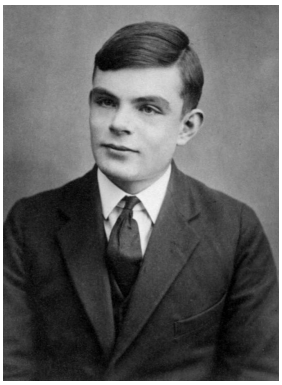

Figure 4: Alan Turing à seize ans

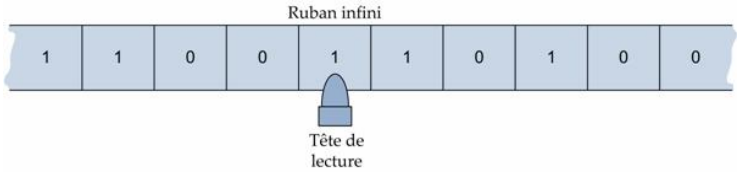

Figure 5: Une machine de Turing consiste en une bande infiniment longue dans les deux sens, divisée en cases pouvant contenir soit le chiffre zéro, soit le chiffre un [2]. La tête de lecture/écriture implémente un algorithme: à chaque étape, elle peut se déplacer une case vers la droite ou la gauche, et lorsqu'elle y arrive, elle peut changer le nombre qui s'y trouve ou le laisser tel quel. Mais avec des instructions convenablement ingénieuses et une interprétation convenablement ingénieuse de la chaîne de zéros et de uns, ces 'machines' peuvent calculer n'importe quelle fonction mathématique qui est effectivement calculable. Image: Jean-Yves Girard [3]

des Grecs. Un algorithme est une séquence finie et non-ambiguë d'opérations ou d'instructions capable de résoudre une classe de problèmes. Tout le monde sait depuis l'école comment diviser un nombre par un autre, ou comment trouver tous les diviseurs d'un nombre entier. Les premières fois qu'on doit les implémenter, ce sont des exercices mentaux assez exigeants, mais avec la pratique, ils deviennent automatique. En effet, ils deviennent mécanique.

Et voici l'idée brillamment simple de Turing: pourquoi ne pas fabriquer des machines pour implémenter des algorithmes. Naturellement, c'était plus facile à dire qu'à faire. En effet, les premières machines de ce type, communément appelées machines de Turing, étaient des jouets de mathématiciens, sans aucune pièce mécanique (voir Fig. 5). Mais peu importe, ce sont les idées qui comptent, et celle-ci a été rapidement mise en œuvre. De plus, cela finirait par soulever ce qui semble aujourd'hui être une question évidente: les informations récoltées par nos sens, sont-elles traitées de manière algorithmique dans le cerveau? Nous

reviendrons sur cette question.

Pendant la seconde guerre mondiale, les idées de Turing et de bien d'autres comme John von Neumann et Norbert Wiener ont conduit au développement de vraies machines de traitement de données, les premiers ordinateurs. Inutile de dire que, comme toujours en science, il y avait de nombreux antécédents à cela, remontant à des gens comme Charles Sanders Pierce au XIXe siècle. En Grande-Bretagne, ces machines étaient conçues pour déchiffrer les codes utilisés par l'Allemagne nazie, tandis que celles aux USA avaient pour objectif de calculer les trajectoires de missiles, mais pas que. Ils ont également été développées dans le cadre du projet Manhattan pour déterminer ce qui se passe dans les réactions nucléaires lorsqu'une bombe explose.

En 1952, deux physiologistes britanniques, Alan Hodgkin et Andrew Huxley, ont élaboré un modèle mathématique de la façon dont les signaux électriques se propagent dans les neurones. Il s'agit d'une réalisation formidable qui leur a valu des prix, mais pourquoi cela a-t-il pris autant de temps? Après tout, ces cellules avaient été identifiées au début du siècle. En effet, le problème est que les neurones sont très petits, comme nous le verrons dans un instant, et les signaux électriques très faibles. Il a donc fallu un développement technologique considérable pour atteindre ce stade. Comme toujours, une technologie développée par la science a pu être utilisée par la suite pour approfondir nos connaissances scientifiques.

Dans les années 1950, le linguiste Noam Chomsky étudiait la structure du langage—le langage au sens générique—et les transformations telles que celles entre formes déclaratives et interrogatives, ou entre formes actives et passives. Il a observé que de telles transformations pouvaient être effectuées de manière algorithmique, c'est-à-dire par des ensembles de règles fixes, et il ne fallut donc pas longtemps avant que des informaticiens comme Joseph Weizenbaum du MIT construisent des ordinateurs capables de converser avec un être humain. Voici un exemple d'une conversation entre ELIZA, le système conçue par Weizenbaum, et une jeune femme qui ne sait pas que son interlocuteur est une machine (la jeune femme commence):

Les hommes sont tous pareils.

Dans quel sens?

Ils sont toujours en train de nous enquiquiner pour quelque chose.

Est-ce que vous avez un exemple plus spécifique?

Eh bien, c'est mon compagnon qui voulait que je vienne ici.

Votre compagnon voulait que vous veniez ici?

Eh bien, il dit que je suis souvent déprimée.

Je suis désolé d'apprendre que vous êtes déprimée.

C'est vrai. Je suis malheureuse.

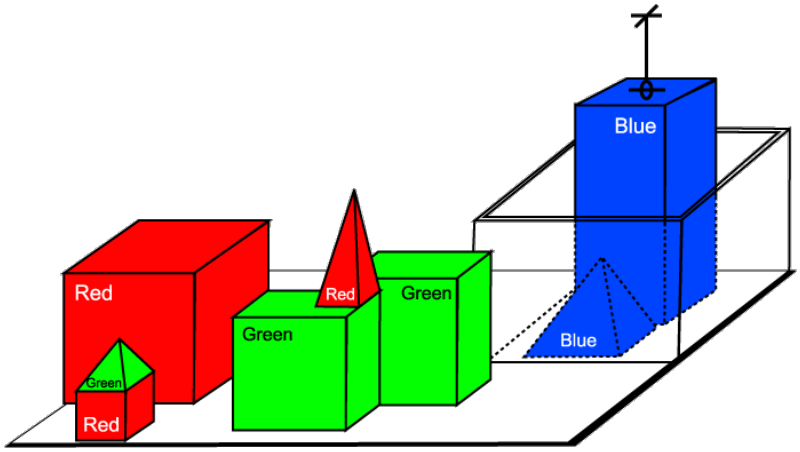

Figure 6: Le monde virtuel habité par SHRDLU. D'après [4]

Pensez vous que ça vous aidera de venir ici?

J'ai besoin d'aide. De ça je suis convaincue.

Ca vous ferait quoi d'obtenir de l'aide?

Je pourrais peut-être apprendre à mieux m'entendre avec ma mère.

Parlez moi un peu plus de votre famille.

Il est clair que la machine pourrait se faire passer pour un psychanalyste, jusqu'à un certain point. En ce sens, il aurait réussi au moins une version du bien connu test de Turing. Mais ce n'est pas difficile de sentir ses limitations.

Une autre réalisation précoce dans le nouveau domaine de l'intelligence artificielle était SHRDLU, création de Terry Winograd à l'université de Stanford. SHRDLU était un robot virtuel vivant dans un monde virtuel très simple contenant des formes géométriques 3D avec des couleurs vives et facilement distinguables (voir Fig. 6). Dans ce système, des règles grammaticales abstraites sont représentées et intégrées avec d'autres informations sur l'environnement. Vous pouvez demander à SHRDLU de déplacer un objet vers une nouvelle position. Si vos instructions ne sont pas claires, le robot vous interrogera jusqu'à ce que tout soit sans ambiguïté—essentiel si l'on doit appliquer un algorithme. Il procédera ensuite à l'exécution de la tâche.

Naturellement, il s'agit d'un monde quelque peu rudimentaire, mais ce n'est pas parce qu'une tâche serait facile pour les humains qu'elle soit réellement simple. Et ceux qui travaillent dans l'IA en sont pleinement conscients, car ils en subissent les conséquences dans tout ce qu'ils essaient de faire. Ils sont bien plus conscients des difficultés rencontrées que la personne moyenne dans la rue, qui peut être prompte à critiquer ce qui semble parfois être des réalisations plutôt dérisoires

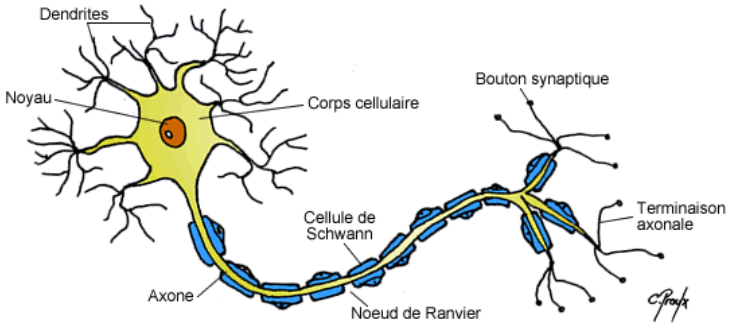

Figure 7: Schéma d'un neurone. Image: Chantal Proulx

Le cerveau

Ce fut en effet une histoire très courte, commençant arbitrairement en 1877 et se terminant brusquement dans les années 1970. Mais bien sûr, la discussion suivante porte sur ce qui s'est passé ensuite, donc cela fait aussi partie de l'histoire. L'objectif principal de la première section était de planter le décor et de montrer combien de domaines d'investigation différents sont impliqués dans cette entreprise, même si on s'y attendait déjà à partir de la description des sciences cognitives comme l'étude interdisciplinaire et scientifique de l'esprit, ou mieux, de l'être humain.

Naturellement, on s'attend aussi à ce que le cerveau joue un rôle clé dans une telle étude. Alors qu'est-ce que le cerveau? On a déjà vu qu'il est composé, entre autres, de neurones. Alors commençons par ceux-ci, mais en gardant à l'esprit que ce récit est très simplifié. Il y a beaucoup plus d'ingrédients dans le cerveau que les neurones et on peut imaginer qu'ils jouent un rôle important, sinon ils ne seraient pas là. La figure 7 est un schéma d'un neurone générique. Encore une fois, nous devons nous rappeler qu'il existe différents types de neurones et que le diagramme ne montre que quelques-unes des principales caractéristiques.

Notez le corps cellulaire sur la gauche. Son noyau contient l'ADN que la cellule utilise comme matrice pour fabriquer des protéines. Le noyau est entouré de filaments ramifiés appelés dendrites. Le corps cellulaire peut mesurer entre un centième et un dixième de millimètre. Comme nous avons dit, il n'est donc pas très grand. Cependant, ces cellules, les neurones, ont une extension appelée l'axone dont la longueur peut varier entre un millimètre et un mètre. Les signaux électriques passent le long de l'axone. A son autre extrémité, à droite sur l'image, se trouvent les terminaisons axonales, qui peuvent être bien plus nombreuses que ce qui est montré ici. Chaque terminaison est dotée d'un bouton synaptique, et celui-ci se connecte avec la dendrite d'un autre neurone, bien qu'il y ait un petit espace entre le bouton synaptique et le début de la dendrite qui s'ensuit. Cet écart s'appelle une synapse (figure 8).

Il est intéressant de donner quelques chiffres ici. On estime qu'il y a environ 80 milliards de neurones dans le cerveau humain. C'est-à-dire, 80 mille mil-

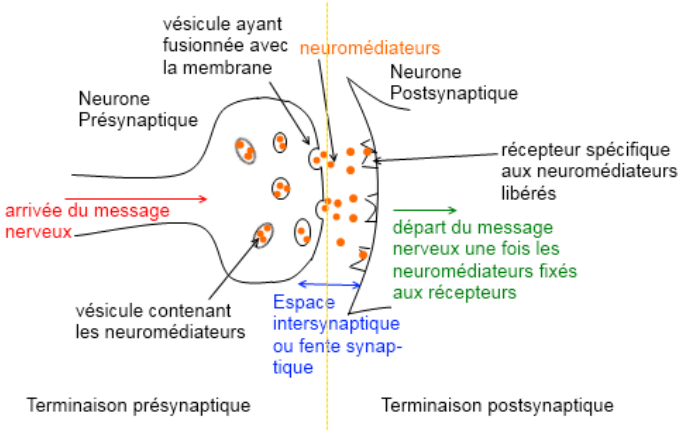

Figure 8: Schéma d'une synapse, avec plus de détails que nécessaire pour nos besoins [5]. Les signaux sont transférés chimiquement à travers ce minuscule espace par des molécules appelées neuromèdiateurs. La distance n'est que de vingt à quarante nanomètres (1 nanomètre = 1 milliardième de mètre)

lions! Mais il y a aussi 500 millions dans l'intestin et 40 000 dans le cœur. Une grande partie du corps est donc concernée par ces cellules. Et chaque neurone se connecte en moyenne à environ 7000 autres neurones, donc si l'on considère cela comme une sorte de circuit électrique, il s'agit d'un circuit qui est vraiment hautement interconnecté. Puisqu'il y a une synapse pour chaque connexion, cela signifie qu'il y a environ un million de milliards de synapses chez un enfant de trois ans, alors que le nombre a tendance à diminuer par la suite.

La figure 9 montre une vue de la partie externe du cerveau humain. Ce que nous voyons ici, c'est le cortex. Il n'a que quelques millimètres d'épaisseur mais, avec tous les plis, couvre en fait une assez grande surface. Ces plis ont été numérotés par le neurologue allemand Karl Brodmann, qui a vécu jusqu'au début du vingtième siècle. Il en a identifié 52 au total. On ne connaît pas encore les fonctions de chacun de ces plis, si tant est qu'on puisse leur attribuer une fonction. Notez tout de même que l'aire de Broca semble participer à la production d'une parole cohérente, tandis que l'aire de Wernicke aide au traitement de la parole et à la compréhension du langage. Paul Broca était un médecin français qui a vécu au XIXe siècle, et Carl Wernicke un autre neurologue allemand de la seconde moitié du XIXe siècle. On voir ici comment la communauté scientifique commençait à cerner le cerveau à l'époque.

C'est un aperçu très superficiel du cerveau mais cela soulève déjà une question intéressante: peut-on étudier l'activité cognitive en regardant ce qui se passe dans le cerveau, c'est-à-dire de manière ascendante, ou bottom-up, comme disent les Anglo-Saxons? On dit que le neurone se déclenche lorsqu'un signal

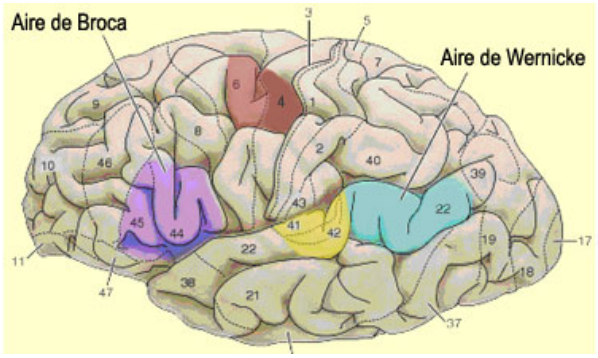

Figure 9: Le cortex humain. Tous les mammifères ont un cortex, entourant d'autres parties du cerveau

passe dans son axone. Comme cela signifie qu'un petit courant électrique circule le long de son axone vers d'autres neurones, ce courant peut être détecté en plaçant une toute petite électrode à proximité. Un neurone individuel peut également être stimulé en appliquant une tension suffisamment faible via une telle microélectrode. Voyons donc un exemple de recherche mettant cela à profit.

La figure 10 montre le neuroscientifique américain pionnier Eric Kandel. Il avait un parcours assez inhabituel. Il a étudié l'histoire et la littérature à Harvard, avant de s'intéresser à la psychothérapie. Et à partir de là, il est passé aux neurosciences. Un peu à l'opposé de Freud, mais les temps ont changé et il a eu de la chance à cet égard. Une technologie beaucoup plus fine était devenue disponible et Kandel a été la première personne à enregistrer des signaux électriques dans des cellules individuelles de l'hippocampe. Cet élément clé du cerveau de chaque vertébré joue un rôle majeur dans l'apprentissage et la mémoire. Chez l'homme, il est important pour la mémoire épisodique, le souvenir conscient d'expériences personnelles, contenant des informations sur ce qui s'est passé et aussi où et quand cela s'est passé.

Plus important encore, Kandel a également établi l'un des premiers liens directs entre ce qui se passe dans les circuits neuronaux et le comportement observé en corrélation avec cette activité. Mais lorsqu'il a réussi à enregistrer les signaux électriques de cellules individuelles de l'hippocampe, il n'avait toujours aucune idée de la façon dont un souvenir pourrait être stocké, ni comment ce souvenir pourrait être lié à un comportement particulier. Les animaux impliqués étaient trop compliqués pour ces premières expériences. Dans ses propres mots [6]:

> J'ai réalisé qu'en enregistrant simplement ce qui se passait dans l'hippocampe, je ne pouvais pas vraiment voir comment les informations entrantes sont modifiées et comment la sortie est modifiée. J'ai eu la chance de réaliser très tôt que, même si l'hippocampe est merveilleux, c'est un problème trop complexe pour commencer. On a besoin de quelque chose de beaucoup plus simple.

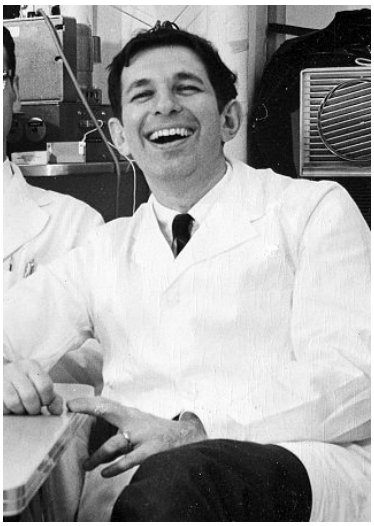

Figure 10: Eric Kandel [6]

C'est là qu'il a tourné son attention vers la limace de mer géante, également connue sous le nom de lièvre de mer de Californie (voir la figure 11).

Cette créature peut mesurer 30 ou 40 cm de long et pourrait peut-être faire un bon repas si elle était préparée à la bordelaise, qui sait? L'avantage pour Kandel était que cet animal a un style de vie relativement simple et seulement 20 000 neurones dans son cerveau. Ses neurones sont regroupés en ensembles distincts, chacun contrôlant un certain type de comportement. Kandel a donc tenté de comprendre sa capacité de mémoire et d'apprentissage au niveau des neurones individuels. Pour ce faire, il a associé un choc électrique indolore sur un neurone à un choc électrique douloureux sur un autre. En raison du stimulus douloureux, l'animal rétracte sa branchie et son siphon comme une forme de défense.

La première partie de l'expérience a fourni un exemple classique du phénomène de conditionnement si cher aux béhavioristes. Car après avoir administré ces combinaisons de chocs électriques à quelques reprises, Kandel a observé que l'animal rétractait ses parties délicates lors de l'administration du seul choc indolore. Cependant, cette réaction ne persistait pas longtemps, à moins que les chocs combinés eux-mêmes ne soient d'abord administrés plusieurs fois. C'était comme si l'animal avait une mémoire et oubliait si la période de conditionnement—d'apprentissage, pourrait-on dire—n'avait pas été assez longue ou convaincante. Exactement comme lorsque nous apprenions nous-même à appliquer les algorithmes à l'école. Il fallait donc s'y attendre, comme les béhavioristes l'avaient si clairement démontré sur bien d'autres animaux et comme nous ne le savons que trop bien nous-même.

Mais la différence avec l'expérience de Kandel était qu'il est ensuite allé voir ce qui avait réellement changé dans les neurones ciblés. Concernant la

Figure 11: La limace de mer géante *Aplysia californica* [6]. Image: Chad King/NOAA MBNMS, CC BY-SA 3.0

mémoire à court terme, résultant d'une courte période de conditionnement, il s'est avéré que les synapses existantes permettaient aux signaux de passer plus facilement. Cela illustre la célèbre règle de Hebb, qui porte le nom du neuropsychologue canadien Donald Hebb (on le cite on anglais, c'est plus court): *When it fires, it wires*. Cela voudrait dire que, chaque fois qu'un neurone en déclenche un autre, cela tend à faciliter la connexion entre eux. Concernant la mémoire à long terme, résultant d'une plus longue période de conditionnement, Kandel a découvert que les noyaux des corps cellulaires des neurones se sont mis à produire des protéines pour construire plus de synapses, et donc plus de connexions. Et si vous demandez ce qui se passe dans l'hippocampe humain au niveau des neurones individuels lorsque nous mémorisons ou apprenons, c'est fondamentalement la même chose.

Voilà donc un exemple d'expérience sur des neurones individuels. Mais qu'en est-il de l'activité neuronale à des échelles plus grandes, impliquant des ensembles de neurones plus étendus? Pouvons-nous voir ce qui s'y passe et le relier à l'exécution de fonctions animales? Une technique fascinante est la stimulation magnétique transcrânienne (SMT). L'idée est d'intervenir dans l'activité neuronale depuis l'extérieur du crâne en focalisant un fort champ magnétique sur une partie du cerveau. Contrairement aux investigations sur un seul neurone, qui nécessitent un accès direct à la zone cérébrale concernée, la SMT est non-invasive, même si l'on peut en effet désactiver temporairement une petite région du cerveau.

Un domaine d'intérêt particulier est le gyrus fusiforme. Cela ne fait pas partie du cortex. Elle est importante pour la reconnaissance faciale. On le sait car il y a eu des cas de patients présentant des lésions spécifiques de cette zone et on peut en observer les conséquences: ils ont beaucoup de mal à reconnaître les personnes de leur entourage, et même leur propre visage dans une glace. Ce n'est pas une déficience visuelle. Ils voient parfaitement bien. Le nom

de cette condition est prosopagnosie. Citons le philosophe de la biologie Alex Rosenberg, qui nous donne ce qu'on pourrait appeler une perspective cognitive de cette condition [7]:

> [...] il existe en fait un ensemble de neurones [le gyrus fusiforme] dont les synapses ont été câblées ensemble de sorte qu'elle répondent à l'entrée visuelle du visage de votre mère en produisant une sortie neuronale qui conduit à un comportement de reconnaissance de la mère, comme vous en train de dire "Salut, maman" quand vous voyez son visage.

Il s'agit donc d'une expérience simple. Un champ magnétique puissant mais focalisé est utilisé pour désactiver votre gyrus fusiforme. Rosenberg continue:

> Vous reconnaîtrez sa voix, son odeur, son écriture. Mais quand elle entrera dans la pièce, vous ne saurez pas qui elle est. Quand elle parlera, vous accuserez l'inconnu devant vous d'être un imposteur imitant la voix de votre mère. Quand elle vous dit son nom et qu'elle est votre mère, ça n'aidera pas. Vous ne la reconnaîtrez pas. Sinon, vos réponses aux diverses stimulations seront parfaitement normales. Une fois que le neuroscientifique aura éteint le champ magnétique et réactivé ainsi les neurones spécialisés encodant la réponse normale au visage de votre mère, vous n'aurez aucun mal à lui donner un nom.

La grande avantage de ces investigations est que la SMT est réversible!

Notons au passage l'importance considérable des études sur les personnes ayant subi des lésions cérébrales. Les conséquences de telles lésions nous renseignent sur le fonctionnement cérébral. Et ça marche dans les deux sens: en identifiant les fonctions de différents ensembles de neurones, un objectif est de trouver des moyens d'aider les personnes ayant subi des lésions cérébrales. Néanmoins, inutile de dire que les études sur les patients atteints de cette manière ne sont pas un moyen idéal pour comprendre le cerveau car, contrairement à la SMT, ces lésions impliquent généralement plusieurs parties différentes du cerveau, ce qui complique l'analyse.

Il est également possible de cartographier l'activité dans l'ensemble du cerveau. Il existe plusieurs méthodes. La première se concentre sur l'activité électrique ou magnétique et est connue sous le nom d'électroencéphalographie ou de magnétoencéphalographie (voir la figure 12). Cela aussi est non invasif, et comme on peut le voir sur la photo, cette expérience n'a rien de désagréable pour ceux qui la subissent! Le volontaire sera invité à effectuer une tâche spécifique pendant que le neuroscientifique identifie les parties du cerveau utilisées pour l'accomplir. Quel genre de tâches? Cela peut aller du mouvement d'un doigt à la réflexion ou à l'exécution d'un calcul. On étudie ainsi l'organisation fonctionnelle du cerveau, et avec une très bonne résolution temporelle, c'est-à-dire qu'on sait avec une grande précision *quand* une certaine partie du cerveau entre en action. La résolution spatiale n'est pas encore particulièrement précise avec

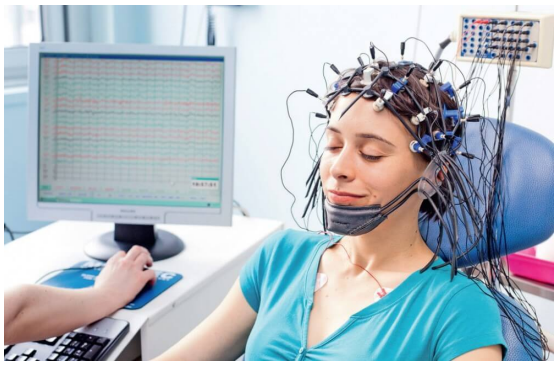

Figure 12: L'électroencéphalographie (EEG). D'après [8]

cette méthode, mais elle est suffisante pour enregistrer des effets se produisant dans de grandes régions du cerveau, comme les ondes cérébrales, par exemple, qui sont en fait des décharges synchronisées de neurones à diverses fréquences assez spécifiques.

Pour une autre méthode d'enquête, considérez ceci. Il est bien connu que le cerveau est l'organe de notre corps qui utilise le plus d'énergie. Cela signifie qu'il a besoin de beaucoup de sang pour lui fournir de l'oxygène, et ceci est exploité par deux méthodes différentes pour cartographier l'activité cérébrale:

- La tomographie par émission de positons, souvent désigné par l'acronyme TEP, est une méthode d'imagerie médicale enregistrant l'activité métabolique ou moléculaire d'un organe en trois dimensions. Un produit radioactif préalablement injecté est marqué par un atome radioactif qui émet des positons—oui, l'antiparticule de l'électron! Chaque positon s'annihile rapidement lors de sa première rencontre avec un électron, pour produire deux photons qui sont émis dos à dos, c'est-à-dire dans des directions opposées à partir du point d'émission. L'emplacement exact du point d'émission peut alors être localisé et la concentration du traceur en chaque point de l'organe peut ensuite être déterminée par un ordinateur. C'est cette information quantitative qui est représentée sous la forme d'une image montrant en couleur les zones de forte concentration du traceur, et donc les zones de débit sanguin élevé. Ces scans ont une bonne résolution spatiale—à quelques millimètres près—mais la résolution temporelle est plutôt médiocre, à environ 30 secondes.

- L'imagerie par résonance magnétique fonctionnelle ou IRM fonctionnelle (IRMf) donne une bonne indication du niveau d'activité cérébrale en détectant les changements associés au flux sanguin, en particulier concernant le niveau d'oxygène dans le sang. Cette technique utilise le fait que le flux sanguin et l'activation des neurones dans le cerveau sont étroitement corrélés. Le mot 'fonctionnel' dans le nom est utilisé pour la même raison que ci-dessus: on peut demander à l'individu d'exécuter une fonction

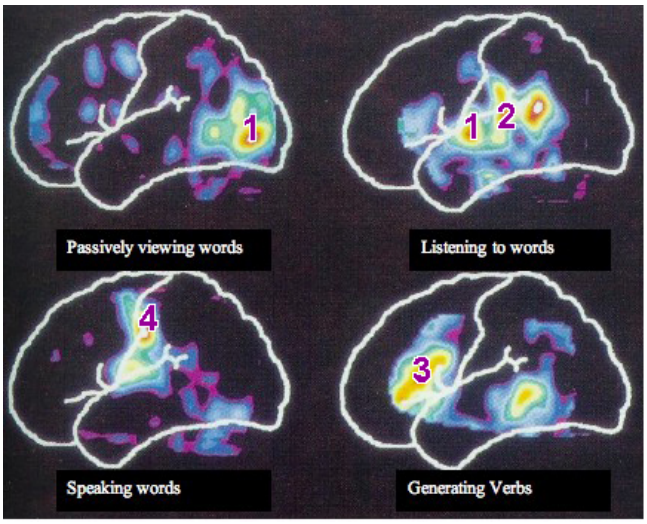

Figure 13: TEP scans d'un individu effectuant différentes tâches linguistiques. Image: Dr Janet Fitzakerley [9]

> spécifique, c'est-à-dire d'effectuer une tâche spécifique, tout en étant allongé dans l'appareil d'IRM, et puis on surveille les parties du cerveau qui 's'allument'. Le grand avantage par rapport à d'autres méthodes est qu'elle n'implique pas de chirurgie, d'injections, d'ingestion de substances ou d'exposition aux rayonnements ionisants. La résolution spatiale est bonne avec cette méthode, mais la résolution temporelle l'est moins: l'activité cérébrale peut être localisée au millimètre près, mais en utilisant les techniques standards, pas mieux que dans une fenêtre de quelques secondes.

La figure 13 montre un exemple d'images produites à l'aide de la méthode TEP. Dans ce cas, l'individu effectuait des tâches linguistiques de différents types:

- en haut à gauche, regardant passivement des mots sur un écran,
- en haut à droite, écoutant des mots passivement,
- en bas à gauche, disant des mots à voix haute,
- en bas à droite, produisant des mots en réponse à une instruction.

Notez que, lorsqu'une tâche implique l'une des autres tâches, par exemple, prononcer un mot implique de le lire à l'écran, on peut soustraire une image d'une autre pour voir quelles *nouvelles* régions du cerveau sont activées par le fait que la personne ne regarde pas seulement le mot, mais le prononce également à haute voix.

Ceci était donc un très bref aperçu de certaines des techniques disponibles pour observer ce qui se passe dans le cerveau lorsqu'un animal entreprend une

tâche quelconque. Inutile de dire qu'il y en a d'autres, et chacun d'eux a des variantes. De plus, toutes ces techniques s'améliorent constamment à tous égards. Mais passons maintenant à une question plus profonde.

Cerveau et ordinateur

Une question naturelle est la suivante: nos cerveaux fonctionnent-ils comme des ordinateurs? Sont-ils des machines qui calculent comme nos ordinateurs? Calculent-ils même quelque chose? L'idée a été lancée par les psychologues cognitifs américains Allen Newell et Herbert Simon dans les années 1970. Fait intéressant, Newell était également informaticien et Simon économiste, montrant à quel point il s'agit d'un domaine véritablement interdisciplinaire. Leur nom pour cette idée était *l'hypothèse du symbole physique*: les règles et les principes sont explicitement représentés sous forme symbolique dans le cerveau. Pour être bref, on pourrait dire quelque chose comme ceci: la syntaxe suit la sémantique. Mais il sera intéressant de développer cette notion.

Nous allons faire cela dans ce qui suit, jusqu'à un point, mais fondamentalement, on peut très bien comprendre en se référant à l'ordinateur comme une analogie parfaite: lorsque nous entrons des règles et des principes dans nos ordinateurs via le clavier—règles et principes qui ont évidemment un sens pour nous—ceux-ci sont fidèlement représentés par la *structure* de configurations de matière—disons des atomes pour être simple—à l'intérieur de l'ordinateur. La syntaxe donne les règles structurelles pour organiser les atomes de manière à ce que le résultat corresponde fidèlement à une information bien formée. Et une information bien formée est une information qui a un sens, du moins pour nous, sinon pour l'ordinateur—et ça, c'est de la sémantique. L'espoir est que les choses pourraient peut-être fonctionner ainsi dans le cerveau, si nous remplaçons les atomes par des neurones. Donc, selon les mots de Newell et Simon:

> Un système de symboles physiques a les moyens nécessaires et suffisants pour une action intelligente générale.

Cela a été illustré par le langage de la pensée (en anglais, le *language of thought* ou LOT), également connu sous le nom de 'mentalais', un concept théorique mis en avant par le philosophe américain Jerry Fodor. Le mentalais était censé être réalisé par des configurations de neurones, tout comme l'information est réalisée par des configurations d'atomes dans un ordinateur. Nos pensées, croyances, etc., seraient alors de véritables entités physiques, des phrases en mentalais, écrites physiquement dans des configurations de neurones.

Il existe une anecdote bien connue concernant l'économiste britannique J.M. Keynes, que Bertrand Russell a décrit comme l'homme le plus intelligent qu'il ait jamais rencontré! Lorsqu'on lui a demandé s'il pensait en mots ou en images, Keynes a hésité un instant, puis a répondu: "Je pense en pensées." Naturellement, cela soulève la question: qu'est-ce qu'une pensée? Et c'est une question importante si nous voulons comprendre la cognition. La pensée est un élément

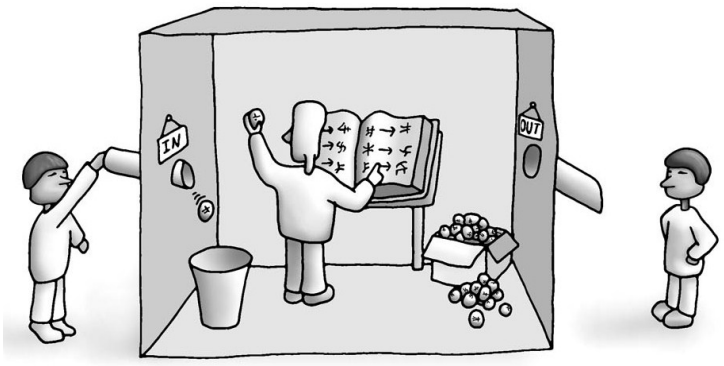

Figure 14: L'expérience de pensée de la chambre chinoise [9]

clé de la cognition dans notre perspective quotidienne normale. Mais qu'est-ce que c'est? Nous reviendrons sur cette question, mais pour l'instant, disons simplement que dans la vision de Fodor une pensée est une configuration de neurones (ou autre matière cérébrale) qui par sa structure même, ou syntaxe, représente fidèlement le sens que nous associons normalement avec la pensée.

Un autre philosophe américain, John Searle, ne voulait rien de tout ça! Sa réaction fut d'inventer une expérience de pensée, désormais célèbre. On l'appelle la chambre chinoise, illustrée sur la figure 14. Un locuteur chinois transmet des messages en chinois à Searle, qui est enfermé dans une pièce. Ce dernier ne comprend pas le chinois, mais il dispose d'un manuel qui lui permet de générer une réponse appropriée de manière mécanique et algorithmique. Cette réponse est ensuite transmise à l'autre locuteur chinois sur la droite. Les deux chinois sont convaincus que la personne présente dans la pièce comprend le chinois.

On voit le parallèle: supposons que les chercheurs en IA réussissent à produire un ordinateur qui semble communiquer en chinois. Cela signifie-t-il que cet ordinateur *comprend* le chinois? C'est évidemment une autre question importante. Comprendre est une activité clé de notre cerveau, mais qu'est-ce que c'est? Searle considère qu'il n'y a pas de différence entre les rôles d'un tel ordinateur et lui-même dans son expérience de pensée. Il introduit ici deux notions importantes, l'IA forte et l'IA faible:

- IA forte: la machine comprend vraiment le chinois.
- IA faible: la machine ne fait que simuler une compréhension du chinois.

Pour Searle, une IA forte est impossible. On pourrait passer le reste de ses jours à lire la littérature sur la chambre chinoise tellement ça a été discuté dans les journaux de philosophie. Mais plus importante est la question: qu'entendons-nous par 'comprendre'? Qu'est-ce qui est impliqué dans la compréhension de quelque chose?

Pour revenir à l'hypothèse du symbole physique, Searle semble avoir montré que l'on peut avoir une manipulation de symboles dans le sens de Newell et Simon sans qu'il n'y ait aucune forme d'intelligence ou de compréhension. Peut-on alors obtenir une compréhension simplement en manipulant de tels symboles physiques?

Un autre paradigme de l'IA

Laissons cette question de côté pour l'instant et passons à une autre approche de l'intelligence artificielle que l'on pourra alors opposer à l'hypothèse du symbole physique. Ce dernier a commencé par regarder ce que font les ordinateurs, et en particulier leur manière de fonctionner, et s'est demandé si nos cerveaux fonctionnaient aussi comme ça. Mais le paradigme que nous allons présenter ici commence par examiner les cerveaux, ou certains de leurs aspects, et se demande si les ordinateurs pourraient être conçus pour capturer ces aspects.

Rappelons d'abord la règle de Hebb, qui dit: chaque fois qu'un neurone en déclenche un autre, cela tend à faciliter la connexion entre eux. En parlant de l'expérience de Kandel, nous avons noté que les forces des connexions synaptiques changent souvent en réponse à une série de stimuli externes, fournissant une base physique pour ce que nous appelons *apprendre*. Il est également vrai qu'un neurone donné ne se déclenchera que lorsqu'il atteint un certain niveau de stimulation. Ces faits ont été notés en 1943 par Warren McCulloch et Walter Pitts, tout au début du développement des ordinateurs. McCulloch était neurophysiologiste et cybernéticien et Pitts logicien. Ils ont eu l'idée d'un réseau de neurones artificiels.

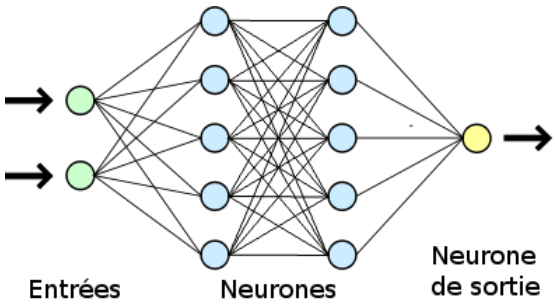

Figure 15: Un réseau de neurones artificiels [10]

Les premiers de ces réseaux, développés par McCulloch et Pitts, étaient des affaires assez simples, capables de calculer certaines des fonctions booléennes si chères aux logiciens. Mais ne nous attardons pas là-dessus. La figure 15 montre un exemple simplifié des réseaux qui sont fabriqués aujourd'hui. Tout d'abord, qu'est-ce que c'est? Qu'est-ce que ça peut faire? C'est un système de traitement de l'information. L'information arrive de la gauche et une réponse émerge de la droite. Une réponse à une sorte de question. Un exemple classique est un

système qui nous aide à détecter les mines terrestres. Lorsque le détecteur balaie le sol, il produit un signal. Ce sera l'information d'entrée. Il faut alors décider s'il a détecté une vraie mine, ou autre chose, comme un rocher par exemple. Les réseaux artificiels peuvent être conçus pour traiter le signal du détecteur et donner la bonne réponse: roche ou mine.

Alors comment fonctionnent-ils? En quoi ressemblent-ils à de vrais réseaux de neurones? Les cercles colorés correspondent à des neurones et les lignes qui les relient sont donc comme des synapses. Les données d'entrée—pour utiliser le terme moderne—arrivent dans les deux neurones verts sur la gauche et les informations sont envoyées à l'une et puis l'autre couche de neurones bleus, et enfin au neurone de sortie jaune. L'information se déplace uniquement de gauche à droite dans ces systèmes. Les couches de neurones bleues sont appelées couches cachées car elles transforment les données d'une manière qui n'est pas visible à l'entrée ou à la sortie. C'est leur existence qui donne le nom de 'deep learning' à ce type de réseau: il est profond dans le sens où il y a au moins une couche cachée. Et en effet, il peut y avoir différents nombres de couches cachées, et différents nombres de neurones dans chaque couche.

Mais ce ne sont pas les seuls paramètres définissant le système. Un seuil est associé à chaque neurone—un nombre compris entre 0 et 1, par exemple. Si le signal total atteignant un neurone donné via les synapses qui l'alimentent est inférieur au seuil, alors ce neurone ne transmettra pas de signal. De plus, un poids est associé à chaque synapse, par exemple un autre nombre compris entre 0 et 1. La connexion entre deux neurones quelconques est ainsi pondérée. Le signal arrivant dans un neurone donné est la somme pondérée de tous les signaux qui lui sont envoyés par les neurones de la couche juste à gauche. Naturellement, c'est loin d'être un modèle de cerveau, ne serait-ce qu'en raison de l'énorme différence dans le nombre de neurones, du moins pour le moment. Mais contrairement à l'hypothèse du symbole physique, qui se demande si les cerveaux fonctionnent comme des ordinateurs, les réseaux de neurones artificiels sont conçus pour voir si les ordinateurs peuvent faire quelque chose comme les cerveaux.

Il reste à voir comment choisir les valeurs des paramètres, et en particulier les poids et les seuils, dont il ne faut pas oublier que ce ne sont que des nombres. Or, il s'avère que ces réseaux peuvent apprendre, c'est-à-dire qu'ils peuvent être configurés pour travailler sur des exemples pratiques et leurs valeurs de paramètres ajustées jusqu'à ce qu'ils obtiennent toujours la bonne réponse. C'est ce qu'on appelle l'apprentissage supervisé. On commence par attribuer des valeurs arbitraires aux poids et aux seuils, et puis on présente au réseau des exemples où la bonne réponse est connue. Si le réseau fait une erreur, c'est-à-dire si la réponse de sortie est incorrecte, ses poids et seuils sont ajustés d'une manière systématique conçue par les mathématiciens. En d'autres termes, on applique un *algorithme d'apprentissage*, qui délivre un nouvel ensemble de poids et de seuils. Il est ensuite présenté avec les mêmes exemples ou d'autres avec des réponses connues et les paramètres sont ajustés si nécessaire—jusqu'à ce qu'il obtienne toujours les bonnes réponses sur les exemples pratiques.

De manière peut-être surprenante, ces systèmes deviennent très forts pour reconnaître les formes et les motifs, et très forts également pour classer les

choses. Par exemple, ils peuvent être utilisés par les agriculteurs pour identifier une maladie des cultures sur la base des caractéristiques observées, et récemment un système a été développé qui peut identifier la maladie conduisant à un certain type de toux—est-ce une bronchite ou le SRAS-CoV-2 virus?

Comparer les deux paradigmes

On peut tirer des conclusions intéressantes en comparant l'hypothèse du symbole physique avec le paradigme du réseau neuronal artificiel. L'hypothèse du symbole physique propose que la structure et la forme—c'est-à-dire la syntaxe—des *informations* représentées dans le cerveau soient directement corrélées avec la structure et la forme des *symboles physiques* qui les représentent, dans ce cas, des configurations de neurones ou d'autres constituants du cerveau. C'est ce qui se passe dans un ordinateur et c'était le point de départ des premières approches à l'intelligence artificielle—en anglais, good old-fashioned artificial intelligence ou GOFAI—qui proposait fondamentalement la même chose comme la meilleure voie vers l'intelligence artificielle. En revanche, les informations sont stockées d'une manière complètement différente dans un réseau de neurones artificiels, dans l'ensemble complet et évolutif de poids et de seuils dont nous avons parlé ci-dessus. De plus, un réseau artificiel peut apprendre. Nous n'avons pas à lui dire au préalable tout ce qu'il 'sait'. C'était l'une des principales critiques, et en fait l'une des principales frustrations, des systèmes GOFAI, qui ne pouvaient faire que ce qu'on leur disait.

Approfondissons un peu ce contraste à l'aide d'un exemple. Renée Baillargeon et Elizabeth Spelke sont des psychologues qui ont étudié la façon dont les très jeunes enfants comprennent le monde. En effet, ils ont identifié quatre principes de physique—on pourrait dire de la physique naïve—que les bébés et les tout-petits utilisent dès l'âge de quelques mois. L'un d'eux est appelé *la permanence de l'objet*. C'est essentiellement l'idée que, lorsque le chat court derrière le canapé, il existe toujours. Il est toujours là, derrière le canapé, même si on ne le voit pas. Baillargeon a mis en place une expérience célèbre pour montrer que les bébés appliquent ce principe avant même de pouvoir parler. Mais comment pourrait-on faire cela, si le bébé ne peut pas s'exprimer? L'idée est simplement de regarder les yeux du bébé lorsqu'on lui présente différents scénarios. Comme nous, les bébés font une 'double prise' quand quelque chose les surprend, et il suffit donc de surveiller leurs yeux de très près. Ce type d'expérience est appelé le paradigme de violation des attentes.

La figure 16 montre ce que le bébé voit sur un écran. D'abord, on lui présente une sorte de pont-levis. On lui montre son mouvement de rotation plusieurs fois—c'est le stimulus d'habituation—jusqu'à ce qu'il s'ennuie et commence à chercher quelque chose de plus intéressant. Les nourrissons sont ensuite présentés avec deux nouveaux stimuli, chacun une variation sur le stimulus d'habituation. L'un des stimuli du test (figure 16c) est un événement possible dans le sens qu'il pourrait se produire physiquement, tandis que l'autre (figure 16d) est impossible, c'est-à-dire qu'il ne pourrait pas se produire physique-

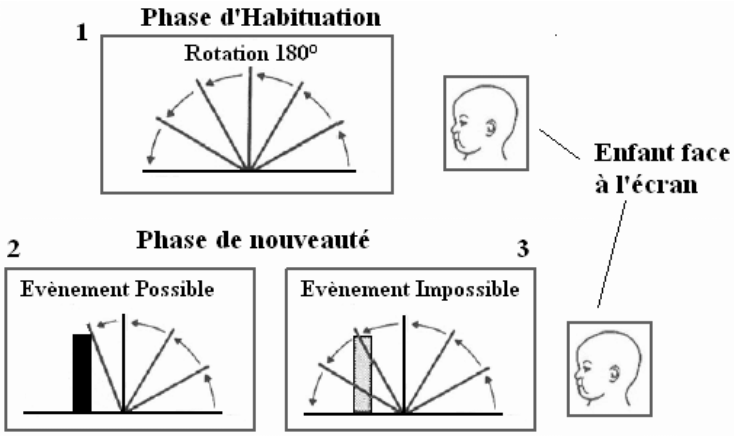

Figure 16: Démontrer que les tout-petits appliquent le principe de permanence de l'objet. Image: Renée Baillargeon [11]

ment de la manière présentée. Le bébé regarde en effet plus longtemps ce dernier cas. Il s'avère donc que les nourrissons savent que la boîte doit toujours être là derrière le pont-levis et qu'un objet solide ne peut pas simplement passer à travers un autre.

Ce qui nous intéresse ici, c'est l'existence de principes physiques dans l'esprit du bébé, principes qui sont fort probablement innés. Ces nourrissons tirent des conclusions sur des choses qu'ils ne peuvent pas voir sur la base de choses qu'ils peuvent voir—comme le font les scientifiques. Or, les principes de base de la physique naïve pourraient être représentés explicitement et appliqués de la manière proposée par l'hypothèse du symbole physique, comme cela se fait dans les ordinateurs, et comme on pourrait essayer de le faire si l'on voulait programmer un robot. C'est l'approche computationnelle de la GOFAI, qui suppose qu'un système cognitif ne peut agir conformément à des principes que si ces principes ont été explicitement représentés dans le système sous une forme symbolique appropriée.

Cependant, il s'avère que l'on peut développer des réseaux de neurones artificiels capables de simuler le comportement de bébés humains dans ces expériences sans coder explicitement aucune règle ou principe. Nous avons donc ici une preuve de concept selon laquelle l'approche computationnelle *n'est pas* la seule option. Quelque chose qui ressemble à un comportement basé sur des règles peut émerger d'un système qui n'exploite aucune règle explicitement représentée. Il peut donc y avoir une forme de traitement de l'information qui n'implique pas le genre de manipulation de symboles régie par des règles avancé par l'hypothèse du symbole physique.

Le problème avec la GOFAI est qu'il s'agit d'une approche intellectuelle du type top-down de l'IA. Toute l'intelligence est effectivement dans la programmation. Et c'est trop lent. Nous avons juste de la chance que les ordinateurs

calculent si vite, car ils doivent tout optimiser et tout planifier à l'avance. Imaginez les calculs nécessaires pour amener un robot GOFAI à se frayer un chemin à travers une forêt, en évitant les nombreux obstacles par le seul biais d'appliquer des principes préprogrammés. Repensez à SHRDLU dans son environnement plutôt simpliste.

Une troisième approche

Une autre approche de la robotique, et en fait des sciences cognitives en général, est connue sous le nom de la *cognition incarnée* (en anglais, embodied cognition). L'un de ses principaux partisans est le roboticien australien Rodney Brooks du MIT. Il a également fondé sa propre société iRobot, qui applique ces idées relativement nouvelles (voir la figure 17). L'objectif est de minimiser le contrôle informatique, en le remplaçant par des connexions directes entre la perception et l'action, c'est-à-dire par diverses formes d'interaction causale directe.

C'est courant en ingénierie. Le thermostat standard en est une bonne illustration. Deux fines bandes de métaux différents sont collées ensemble. Le bilame résultant se pliera si la température augmente, car les deux métaux se dilatent à des degrés différents lorsqu'ils sont chauffés. Cela suffit pour ouvrir ou fermer un circuit électrique et piloter un système de chauffage. Aucun calcul requis. Un exemple plus ingénieux est le régulateur à boules, en fait inventé par Christiaan Huygens puis utilisé bien plus tard, au XIXe siècle, par James Watt pour étendre les applications des machines à vapeur. L'idée était de mieux réguler le débit de vapeur dans ces machines sans aucun calcul [12]. Naturellement, les ingénieurs ont besoin de faire des calculs pour concevoir ces choses, mais la nature peut les concevoir sans calculs, par la méthode essai et erreur, mutation génétique et l'épreuve de la survie. C'est le processus que nous appelons évolution.

Une autre motivation clé de cette approche est l'observation que la pensée n'est pas un processus désincarné qui se déroule presque par magie dans nos têtes. Bien sûr, cela a lieu dans une large mesure à l'intérieur de nos têtes, même si nous ne ressentons rien physiquement dans nos crânes quand nous réfléchissons. (C'est parce qu'il n'y a pas de nerfs sensoriels dans le cerveau—la nature a trouvé un moyen différent de protéger cette partie de notre corps, le crâne lui-même.) Mais il devient clair que la pensée implique bien plus que le cerveau. Rappelez-vous le nombre de neurones dans l'intestin et le cœur, par exemple, et notez comment nos émotions entrent en jeu lorsque nous devons prendre des décisions. Ou pensez à quelqu'un qui parle au téléphone: même si l'autre personne ne peut pas le voir, il gesticule et fait des expressions faciales. En fait, il ne pourrait pas communiquer sans le faire. Tout est lié. Le corps humain est un tout.

Selon Brooks, l'esprit n'est donc pas un processeur d'informations qui passe tout son temps à effectuer des calculs abstraits avec l'aide du cerveau. On peut se demander pourquoi nos premières tentatives d'intelligence artificielle, la bonne vieille intelligence artificielle ou GOFAI, se sont engagées dans cette

Figure 17: Un exemple de robot produit dans l'entreprise de Rodney Brooks. Ce n'est pas une machine à café mais un aspirateur appelé Roomba. Il peut frayer son propre chemin dans votre salon

voie. En fait, nous suivions simplement notre compréhension naturelle de nous-même fournie par l'introspection, c'est-à-dire notre perspective ordinaire de nous-même, plutôt que cognitive ou neuronale. Nous avons pensé que notre cerveau pourrait fonctionner comme nos ordinateurs parce que nos ordinateurs fonctionnent comme nous pensons que notre pensée fonctionne.

Penser

Il est temps de dire quelques mots sur une *perspective cognitive* de la pensée, pour reprendre une tournure de phrase promue par le linguiste Ray Jackendoff dans son excellent livre *A User's Guide to Thought and Meaning* [13]. Rappelons l'anecdote sur Keynes. Si nous ne pensons pas en images et nous ne pensons pas en mots, nous pensons en quoi? Un point de vue largement accepté aujourd'hui, également proposé par Jackendoff, est que toutes nos pensées sont inconscientes. Cela peut sembler fou, mais considérez ceci. Lorsque vous devez communiquer une idée assez subtile à quelqu'un d'autre, ne trouvez-vous pas parfois que vous vous êtes mal exprimés? Vous n'avez pas dit exactement ce que vous voulez dire? Ensuite, vous reformulez, peut-être plusieurs fois, en ajustant ce que vous dites jusqu'à ce qu'à un moment donné, cela sonne juste. Cela correspond enfin à votre idée. Pensez à ce qui se passe lorsque vous devez écrire une lettre plutôt difficile ou un dissertation. Combien de fois finissez-vous par rayer quelque chose ou l'ajuster?

Alors pourquoi disons-nous des choses qui ne signifient pas ce que nous voulons? Jackendoff et d'autres diront que c'est parce que le sens est en fait

caché dans notre inconscient—c'est *l'hypothèse du sens inconscient*. Ce que nous appelons ordinairement la pensée n'est qu'une interprétation linguistique de notre véritable pensée inconsciente, une représentation symbolique dans notre conscience utilisant le symbolisme du langage pour représenter des pensées et des significations qui se trouvent en réalité dans les configurations et la dynamique de nos neurones. La 'pensée consciente' ou la 'réflexion consciente' n'est alors qu'une séquence (logique?) de prises linguistiques sur notre véritable pensée, qui pour sa part est effectuée à notre insu—c'est-à-dire à l'insu de notre introspection—par nos neurones.

Cela nous ramène à la GOFAI, qui est basée sur l'idée que nous pensons comme nous pensons que nous pensons dans notre perspective ordinaire de nous-même. Nous pensons que nous pensons en séquences logiques de phrases—ou de propositions, pour utiliser un terme philosophique fortement connoté—c'est-à-dire des séquences de symboles comme ceux que nous introduisons dans nos ordinateurs et qu'*eux* peuvent manipuler en utilisant des algorithmes. Mais cognitivement, abandonnant pour l'instant les prétentions de notre soi introspectif, penser pourrait bien ne pas être en fait une question d'application de logique à des prémisses et à des principes. Nous savons depuis le développement des réseaux de neurones artificiels qu'il n'y a absolument aucun besoin de prémisses et de principes codés quelque part dans notre cerveau. Dans la perspective ordinaire que nous donne l'introspection, nous avons toujours l'impression que l'information doit être codée quelque part car notre seule prise—ou plutôt la seule prise de notre introspection—sur l'information, et en particulier sa signification, c'est à travers la 'petite voix' que nous entendons presque en permanence dans nos têtes. Mais il se pourrait que ces pensées et significations ne soient codées par le langage qu'à des fins de communication, avec les autres ou avec nous-même. Et en réalité, comme nous le verrons bientôt, nous avons beaucoup de mal à raisonner de manière strictement logique, ce que nos ordinateurs font merveilleusement bien—après tout, ils sont conçus pour.

Les symboles et leurs sens

Les symboles que nous entrons dans nos ordinateurs via le clavier ont tous une signification pour nous. Après tout, un symbole est une marque ou un caractère utilisé comme représentation conventionnelle d'un objet, d'une fonction ou d'un processus. Mais cela a-t-il un sens pour l'ordinateur? Peu de gens comprendraient les choses de cette façon, mais pourquoi pas? Est-ce juste un parti pris contre la machine? C'est une question importante si quelqu'un espère créer une machine intelligente, c'est-à-dire intelligente de la même manière que nous le sommes. Car nous aurons besoin de savoir ce que nous sommes en train de demander de cette machine.

Un problème ici est probablement que nos ordinateurs ne se soucient pas de ce qui leur arrive. Après tout, ils ne se soucient pas de leur propre survie, ou de leur propre intégrité et bien-être. Ils n'ont pas d'objectifs. Ce sont les esclaves dans notre espace de travail, pas des collègues. Ils travaillent simple-

ment pour atteindre *nos* objectifs. Et en effet, nos propres manipulations de représentations symboliques ont un sens pour nous car elles ont une importance pour notre autodétermination et notre survie. Le sens semble donc être quelque chose qui ne concerne que les êtres vivants. Voici ce que le philosophe américain Daniel Dennett dit sur le fait que nos machines sont loin d'avoir le même degré d'autonomie que les êtres vivants [14]:

> Les chercheurs en IA pourraient répondre de manière plausible qu'ils reportent l'examen de phénomènes tels que la capture d'énergie, la reproduction, l'auto-réparation et l'auto-révision permanente, dans l'espoir de récolter les avantages de cette simplification en clarifiant d'abord les phénomènes purement informationnels d'apprentissage et d'autoguidage (une sorte d'autonomie, même incomplète). Personne ne songerait à compliquer la conception d'un ordinateur qui joue aux échecs en exigeant qu'il obtient son énergie de sandwichs et de boissons gazeuses, l'obligeant à surveiller non seulement son temps mais aussi la diminution de ses réserves d'énergie utilisable. Les joueurs d'échecs humains doivent contrôler leur faim et leurs émotions telles que l'humiliation, la peur et l'ennui, mais les ordinateurs peuvent laisser tout cela de côté en toute impunité, n'est-ce pas? Oui, mais à un coût énorme: en se débarrassant de ces préoccupations, les concepteurs de systèmes créent des architectures fragiles (ils ne peuvent pas se réparer eux-mêmes, par exemple) et vulnérables (enfermés dans le seul ensemble de contingences que leurs concepteurs ont anticipé), et totalement dépendant de leurs gestionnaires.

Il est intéressant de constater que nos machines surveillent effectivement leur niveau d'énergie et nous informent de ce besoin. Mais quand ils se tournent vers nous pour obtenir de l'aide, il est clair qu'ils se moquent bien de savoir si nous allons recharger leurs batteries. Il est assez clair que les symboles d'un ordinateur auraient plus de sens pour lui, c'est-à-dire qu'ils deviendraient en fait significatifs pour l'ordinateur, si sa survie dépendait de ce sens, mais jusque-là, ce ne sera qu'une machine, manipulant mécaniquement des structures physiques qui correspondent structurellement à nos chaînes de symboles. Et d'une certaine manière, ces considérations nous ramène à la notion de cognition incarnée.

L'évolution

Il est temps de se tourner vers un thème clé des sciences cognitives, sinon le thème le plus important de tous. Les sciences cognitives s'intéressent principalement aux êtres vivants, même si aucun de nos paradigmes d'IA réussissent à se rapprocher de la façon dont les humains ou d'autres animaux se frayent un chemin dans le monde. Cela signifie qu'il s'agit effectivement d'une branche de la biologie, et comme l'a souligné le généticien américain Theodosius Dobzhansky en 1973, rien en biologie n'a de sens sauf à la lumière de l'évolution. Charles

Figure 18: George Miller, psychologue cognitif américain

Darwin lui-même a affirmé que la psychologie serait basée sur une nouvelle fondation. Tournons-nous donc vers George Miller, considéré comme l'un des fondateurs des sciences cognitives (voir Fig. 18), avec cette citation [15]:

> Le couronnement de l'accomplissement intellectuel du cerveau est le monde réel—tous les aspects fondamentaux du monde réel de notre expérience sont des interprétations adaptatives du monde vraiment réel de la physique.

Il y a ici deux idées clés. La première est que notre cerveau interprète le monde en faisant des représentations. La seconde est que ces représentations sont adaptatives, ce qui signifie qu'elles nous aident dans notre autodétermination et notre survie. C'est ça l'évolution. Il y a donc un monde vraiment réel que nous essayons de comprendre à travers l'étude de la physique, de la chimie et de la biologie, et puis il y a une représentation de cela, faite par notre cerveau. La notion de représentation est importante en sciences cognitives et pas aussi simple qu'il n'y paraît à première vue.

Prenons l'exemple de ce que nous voyons. Les photons du monde vraiment réel pénètrent dans la rétine, qui fait également partie du monde vraiment réel. Notre cerveau représente alors le monde dans notre voisinage—nous appellons cela une représentation visuelle. Cette représentation est construite dans notre crâne mais nous ne la voyons pas là. Nous le projetons tout autour de nous. C'est l'ultime film 3D, car nous pouvons réellement nous déplacer à l'intérieur de cette représentation. Et en effet c'est cela qui est adaptatif: cela nous permet d'aller vers ce qui est bon pour nous dans le monde, et de s'éloigner de ce qui est mauvais. Il a été soigneusement conçu[1] pour servir notre bien-être, si crucial pour nous.

[1]Ce mot est utilisé dans le sens proposé par Dennett pour désigner la conception par mutation génétique et sélection naturelle, où il n'y a pas de concepteur.

Les paragraphes suivants aborderont divers aspects des sciences cognitives du point de vue de l'évolution, ou du moins, en gardant toujours à l'esprit la notion d'évolution.

Une théorie cognitive du raisonnement

Le raisonnement est l'une des choses que les cerveaux sont censés faire le mieux, du moins chez les êtres humains, et c'est donc un sujet important pour les sciences cognitives. Envisageons donc une perspective cognitive sur le raisonnement. Qu'est-ce que c'est? À première vue, cela semble évident: le raisonnement utilise la logique pour résoudre des problèmes et expliquer clairement les choses à soi-même ou aux autres. Et qu'est-ce que la logique? Voici un exemple simple d'un raisonnement:

Si A, alors B
pas B
donc, pas A

Ce genre de petit raisonnement s'appelle un syllogisme. Il a même un nom: *modus tollens*. De telles formes basiques de raisonnement sont connues depuis longtemps, au moins depuis la Grèce antique et probablement avant. A et B sont des propositions, c'est-à-dire des déclarations au sujet de quelque chose dans le monde ou dans notre imagination. Voici un exemple concret:

Si le chat a très faim, alors il mangera du pain sec
mais le chat n'a pas mangé de pain sec
donc le chat n'avait pas très faim

L'important ici est que l'argument est valable quelles que soient les propositions A et B. En effet, cela s'appelle un argument valide. Cela signifie que, chaque fois que les prémisses sont vraies, c'est-à-dire chaque fois qu'il est vrai que "si A, alors B", et aussi que "pas B", la conclusion sera également vraie, c'est-à-dire que A sera faux. Lorsque les prémisses d'un argument valide sont vraies, nous disons que l'argument est concluant, un terme technique pour les logiciens.

Ainsi, la logique s'intéresse à des arguments qui sont vrais en raison de leur forme et de la signification de certains connecteurs et quantificateurs logiques comme 'si', 'alors', 'pas', etc., plutôt qu'à la signification des propositions et des termes comme A et B ci-dessus qui apparaissent dans ces arguments, c'est-à-dire leur contenu sémantique non-logique. Alors, raisonnons-nous logiquement? Enchaînons-nous des syllogismes lorsque nous résolvons des problèmes et expliquons des choses aux autres? Pas si évident. La figure 19 montre le psychologue cognitif britannique Peter Wason, célèbre pour sa manière plutôt intelligente de tester la capacité logique des gens et de découvrir ainsi quelque chose d'absolument fascinant.

Le test de Wason semble très simple (voir la figure 20). On nous présente quatre cartes. Sur chacune il y a une lettre d'un côté et un nombre de l'autre, et on nous dit que les cartes doivent obéir à la règle suivante: s'il y a une

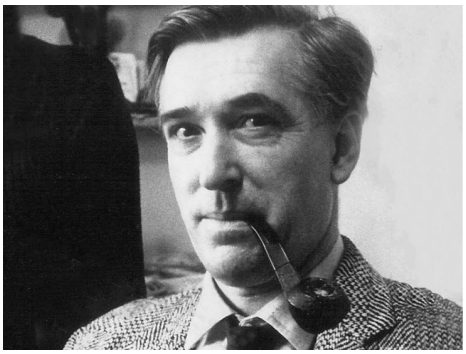

Figure 19: Peter Wason, psychologue cognitif britannique

Figure 20: Quatre des cartes de Wason. D'un côté de chacun il y a une lettre et de l'autre un chiffre. Et il y a une règle: si une carte a une voyelle d'un côté, elle a un nombre pair de l'autre. Quelles cartes doit-on retourner pour voir si la règle est respectée?

voyelle d'un côté, il doit y avoir un nombre pair de l'autre. Alors quelles cartes doit-on retourner pour vérifier que la règle a été respectée? Beaucoup de gens raisonnent ainsi. La première carte montre une voyelle, et il faut donc vérifier qu'il y a bien un nombre pair de l'autre côté. La seconde montre une consonne, et comme la règle ne mentionne pas de consonnes, inutile de vérifier l'autre côté. Le troisième montre un nombre pair, et comme la règle mentionne des nombres pairs, nous devons vérifier qu'il y a une voyelle de l'autre côté. La dernière carte montre un nombre impair, non mentionné par la règle, donc pas besoin de vérifier.

Mais c'est faux. Bien sûr, certaines personnes ne font pas l'erreur avec les deux dernières cartes. Les logiciens, par exemple. On n'a pas besoin de retourner la troisième carte, car quoi qu'il y ait de l'autre côté, consonne ou voyelle, la règle est respectée. Cependant, si la dernière carte avait une voyelle de l'autre côté, la règle ne serait pas respectée, et il faut donc bien vérifier celle-ci.

Cela montre que beaucoup de gens n'appliquent pas correctement ces simples arguments logiques valides. Mais regardez les quatre cartes de Wason de la figure 21. Cette fois, chaque carte indique l'âge de quelqu'un d'un côté et ce qu'il boit

Figure 21: Quatre autres cartes de Wason. D'un côté il y a l'âge de quelqu'un et de l'autre une image de ce qu'il boit. Encore une fois, il y a une règle: il faut avoir au moins dix-huit ans pour être autorisé à boire de l'alcool. Quelles cartes doit-on retourner pour voir si la règle est respectée?

de l'autre. Et cette fois la règle est bien sûr qu'il faut avoir au moins dix-huit ans pour avoir le droit de boire de l'alcool. Quelles cartes doivent être retournées maintenant? C'est très facile et personne, absolument personne ne se trompe ici. La première carte doit être retournée pour vérifier que ce jeune de seize ans ne boit pas d'alcool. On n'a pas besoin de vérifier la seconde, car à vingt-cinq ans, la personne peut boire ce qu'elle veut. La troisième personne boit du soda, donc peu importe son âge, mais la quatrième boit de la bière, et nous devons donc vérifier.

Et alors? Le fait est que, logiquement parlant, le deuxième test est *exactement* le même que le premier. Alors, comment se fait-il que tant de gens fassent des erreurs avec la première version, et personne avec la seconde? La réponse est que le premier test est abstrait tandis que le second traite des conditions déontiques, c'est-à-dire des conditions relatives au devoir et à l'obligation. Et il se trouve que nous sommes extrêmement doués pour tenir les autres responsables de leurs actes, et tout aussi bons pour trouver notre chemin dans un monde dans lequel les autres nous tiendront responsables des nôtres. Pourquoi? Parce que nous vivons dans une société de coopération. En effet, les humains sont les champions du monde de la coopération, bien plus que toute autre espèce. Nous sommes capables de coopérer sur tout et n'importe quoi. D'autres animaux comme les loups, les chimpanzés, les abeilles et les fourmis peuvent vivre en communauté et travailler ensemble de diverses manières, mais les humains inventent continuellement de nouvelles façons de faire les choses ensemble pour obtenir de meilleurs résultats.

Mais cela a un prix, car nous avons besoin de savoir à qui nous pouvons faire confiance et c'est important aussi que tout le monde aient tendance à être dignes de confiance. En d'autres termes, nous avons besoin de normes et d'un système d'éthique. Alors c'est intéressant parce que nous étions à la recherche d'une théorie cognitive du raisonnement et ici nous parlons de moralité. Ce n'est pas un hasard. Il est tout à fait plausible que le raisonnement aurait évolué chez l'humain à des fins moralisatrices, pour nous permettre de vivre dans une société majoritairement coopérative. Les bénéfices pour les gènes conduisant

à un tel comportement sont légion et ces gènes se seraient donc multipliés. Ils sont presque partout aujourd'hui dans la société humaine. On dit parfois qu'il y a eu une sorte d'auto-domestication au cours des dernières dizaines de milliers d'années, ou peut-être plus, c'est-à-dire une coévolution des gènes et des pratiques culturelles qui a fait que certains gènes, ces gènes de coopération, soient devenus dominants.

Deux psychologues cognitifs français, Hugo Mercier et Dan Sperber, ont longuement écrit à ce sujet dans leur livre révolutionnaire *L'Énigma de la raison* [16]. Dans leur récit, les raisons sont des constructions sociales principalement pour la consommation sociale. Pour être plus précis, nous produisons des raisons pour justifier nos pensées et nos actions auprès des autres, mais aussi pour produire des arguments pour convaincre les autres de penser et d'agir comme nous-même. Et nous utilisons la raison pour évaluer les raisons que les autres nous donnent pour se justifier. Le mot clé est encore une fois 'coopération'. Il s'agit d'être sur la même longueur d'onde avec des personnes à qui nous pensons faire confiance.

Mercier et Sperber parlent aussi du paradoxe de la raison: d'une part, les humains ont produit de vastes quantités de connaissances très, très fiables grâce aux efforts de mathématiciens, scientifiques, philosophes, etc., tandis que d'autre part, dans leurs expériences, les psychologues cognitifs montrent régulièrement des exemples de ce qui semble être un comportement irrationnel. Concernant ce dernier, le cas le plus connu est peut-être le biais de confirmation. Il s'agit d'un biais cognitif qui consiste à privilégier les informations qui confirment nos idées reçues tout en accordant moins d'importance aux informations qui semblent les contredire. Alors comment arrivons-nous à produire des connaissances fiables? Et puisque nous faisions la promotion de l'évolution en tant qu'outil explicatif clé dans la compréhension de l'activité cognitive humaine, quel pourrait être l'avantage évolutif du biais de confirmation?

Pour commencer, selon Mercier et Sperber, nous sommes plus groupistes qu'égoïstes et le biais de confirmation est en fait un biais partisan. Ceci est clair dans le débat politique, et peut être observé dans toutes les activités humaines. Mais en réalité, cette tendance peut fournir un moyen efficace de partager la charge de travail cognitive lorsqu'un groupe de personnes a un problème à résoudre. Car il s'avère que nous sommes paresseux lorsqu'il s'agit d'évaluer nos propres raisons et justifications, mais extrêmement efficace pour évaluer celles des autres. Pour comprendre cela, il faut se demander quel est le contexte approprié pour raisonner. D'après ce que nous avons déjà dit sur le rôle social du raisonnement, il semble plausible qu'il puisse s'agir d'une discussion dans laquelle les gens échangent des arguments et des justifications et où il y a un intérêt commun à trouver la vérité, ou du moins à trouver le meilleur compromis. Il suffit de réfléchir un instant à la myriade d'exemples de coopération sociale pour voir que de telles discussions se produisent, même si nous avons tendance à focaliser sur des situations où le débat politique devient non-débat à cause d'une partisanerie trop zélée.

En effet, la raison est bien plus efficace pour évaluer de bons arguments que pour les produire. Il est facile de montrer que les capacités de résolution de

problèmes d'un groupe de personnes sont de loin supérieures en général à celles de n'importe quel individu. Par exemple, les groupes sont bien meilleurs que les individus pour résoudre les tâches de Wason vues ci-dessus. Lorsque les gens reçoivent l'explication correcte de la première tâche de Wason décrite ci-dessus, ils l'acceptent généralement immédiatement, et donc il suffit qu'une personne tombe sur la bonne réponse par hasard.

Penser seul n'est pas un excellent moyen d'obtenir de bonnes solutions. Je cite à nouveau Keynes:

> C'est étonnant les bêtises qu'on peut temporairement croire si l'on réfléchit trop longtemps seul.[2]

On pourrait ajouter, si l'on réfléchit trop longtemps dans un petit groupe de personnes ayant les mêmes idées.

La moralité

L'Américain Jonathan Haidt est un psychologue cognitif qui a appliqué le pouvoir de l'explication évolutionniste à tous les aspects de la question de la moralité. Il en décrit les conclusions dans son livre fascinant *The Righteous Mind. Why Good People Are Divided by Politics and Religion* [17]. Il résume ses conclusions dans ce qu'il appelle son premier principe de psychologie morale, en disant que, lorsqu'il s'agit de jugement et de raisonnement moral, les intuitions viennent en premier et le raisonnement stratégique seulement après.

Concernant l'intuition, Haidt montre à quel point nos jugements moraux sont fortement dépendants de choses dont nous ne sommes tout simplement pas conscients. Par exemple, il décrit une expérience simple pour montrer comment la familiarité peut influencer nos préférences [18]. On demande aux gens d'évaluer les pictogrammes japonais sur une échelle d'un à dix—comme vous le diront les inventeurs de Tripadvisor, les gens sont prêts à évaluer n'importe quoi sur une échelle d'un à dix. Cependant, lors de la lecture des instructions à l'écran, certains pictogrammes apparaissent pendant un temps très court, trop court pour que l'individu en ait conscience. Des notes plus élevées sont généralement attribuées à ces pictogrammes affichés de manière subliminale. D'un point de vue évolutif, il est facile de comprendre que ce serait une bonne stratégie de privilégier les choses que l'on connaît.

Un autre facteur influençant notre jugement moral est une mauvaise odeur dans notre environnement local [19]. On demande aux gens de juger des situations moralement douteuses lorsqu'une mauvaise odeur est introduite dans leur voisinage. Ils ne se rendent pas compte que l'odeur a été introduite délibérément. Leurs jugements deviennent plus sévères. De même si on leur demande de se laver soigneusement les mains avant de commencer [20]. Et les durées des peines prononcées par des juges dans des études de cas peuvent être influencées en les présentant de diverses manières avec des chiffres plus ou moins grands et totalement arbitraires avant qu'elles ne commencent, encore une fois d'une manière

[2] Et il conclut ici "particulièrement en économie".

qu'ils ignorent tout lien entre les chiffres présentés et le but de l'expérience. Une discussion détaillée de cette importante découverte peut être trouvé dans [21]. Ces effets sont connus des psychologues cognitifs sous le nom d'amorçage ou d'ancrage.

La seconde partie du premier principe de psychologie morale de Haidt affirme que le raisonnement moral a une fonction stratégique. Dans tout ce que nous faisons et disons, nous sommes tous obsédés par ce que les autres pensent de nous. Notre pensée morale ressemble plus à un avocat défendant son client qu'à un scientifique essayant d'établir la vérité. Du point de vue évolutif, cela nous rappelle que ce qui était le plus important pour la survie de nos ancêtres n'était pas la vérité mais leur réputation. On peut citer Benjamin Franklin, scientifique, inventeur et l'un des pères fondateurs de la constitution américaine, bien qu'il ne soit jamais devenu président:

> Il est si pratique d'être une créature raisonnable, car cela permet de trouver ou de fabriquer une raison pour tout ce que l'on a envie de faire.

On peut alors se demander comment les humains ont réussi à créer un tel corps de connaissances fiables, pour se rapprocher si bien de la vérité sur le monde, comme nous pouvons le voir dans notre maîtrise du monde physique à travers la technologie.

La réponse vient de ce que nous avons déjà dit plus haut: la connaissance scientifique *n'est pas* le produit de l'esprit humain, ou du moins n'est pas le produit *d'un seul* esprit humain travaillant dans son coin. C'est le résultat d'un énorme effort collectif de personnes qui partagent un intérêt commun à trouver la vérité. Le mythe du génie scientifique et du grand philosophe est bien un mythe, au service peut-être d'un autre besoin humain: le besoin de dirigeants. Aucun de ceux qui sont si souvent acclamés comme étant la source de toutes les grandes découvertes ne travaillait seul, et si la science avance à un rythme aussi fulgurant aujourd'hui, ce n'est certainement pas parce qu'il y a plus de génies à l'œuvre. C'est simplement parce que tant d'humains s'acharnent ensemble à comprendre le monde.

Avant de quitter le thème de la moralité, il est intéressant de décrire les six fondements moraux de Haidt, produits de l'auto-domestication humaine. Dans chaque cas, une réaction morale a évolué pour permettre à nos ancêtres—nos ancêtres humains ici—de faire face à un défi adaptatif spécifique, et elle a coévolué avec une émotion qui caractérise cette réaction:

- Soin/protection contre le mal. Le défi adaptatif était de protéger et de prendre soin de nos enfants, et l'émotion clé qui a évolué en tandem était la compassion.
- Équité/tricherie. Le défi d'adaptation était de récolter les fruits des partenariats bilatéraux, et les émotions associées étaient la gratitude, la colère et la culpabilité. C'est facile à comprendre. Dans une large mesure, la moralité a évolué comme une solution au problème du passager clandestin, toujours une menace dans une société de coopération. Il est important

pour une telle société de dissuader les profiteurs, et important, par exemple, que les profiteurs se sentent coupables lorsqu'ils trichent, ou craignent notre colère, etc. Les resquilleurs doivent être détectés et punis, sinon leurs gènes profiteurs prolifèreront.

- Loyauté/trahison. Le défi adaptatif était de former des coalitions cohésives, et les émotions caractéristiques sont la fierté de groupe et la rage contre les traîtres.

- Autorité/subversion. Le défi adaptatif était de créer des relations bénéfiques au sein des hiérarchies, et les émotions caractéristiques sont le respect et la peur.

- Sainteté/dégradation. Celui-ci peut sembler un peu étrange, mais le défi adaptatif était d'éviter les contaminants. La clé pour comprendre cela est le *dilemme de l'omnivore*. Haidt explique:

  > Être omnivore a l'énorme avantage de donner de la flexibilité: vous pouvez vous promener dans un nouveau continent et être tout à fait sûr de trouver quelque chose à manger. Mais il présente également l'inconvénient que les nouveaux aliments peuvent être toxiques, infectés de microbes ou criblés de vers parasites. Le dilemme de l'omnivore [...] est que les omnivores doivent rechercher et explorer de nouveaux aliments potentiels tout en se méfiant d'eux jusqu'à ce qu'ils soient prouvés sûrs.

  Les deux motivations concurrentes d'un omnivore sont la néophilie—une attirance pour de nouvelles choses—et la néophobie—une peur des nouvelles choses. Différentes personnes présentent ces motivations à des degrés divers. L'émotion caractéristique ici est le dégoût. Le rôle de cette fondation dans les mouvements religieux est clair, mais c'est une autre histoire.

- Le dernier fondement moral, qui, selon Haidt, est toujours en débat, est le fondement de liberté/oppression. Le défi adaptatif serait d'éviter la tyrannie au sein des groupes, et l'émotion caractéristique le sentiment d'injustice. Comme il le souligne, il existe une certaine tension entre ce fondement et le fondement d'autorité/subversion. L'autorité peut bien sûr être légitime, mais nous restons méfiants vis-à-vis de nos dirigeants.

La modularité

Une idée importante en psychologie évolutionniste, c'est-à-dire dans la manière évolutionniste de comprendre l'activité cognitive, est la théorie modulaire de l'esprit. En un mot, c'est l'idée que l'esprit n'est pas unitaire et basé sur des principes, mais modulaire et opportuniste, un assemblage de modules que l'évolution a rassemblés pour faire face aux problèmes spécifiques rencontrés par nos ancêtres—pas seulement nos ancêtres humains, mais tous nos ancêtres,

comme nous le verrons. Naturellement, il y a ici une manière de parler qui peut être trompeuse. Dame nature n'a fait aucun calcul ici. Ces conceptions se sont produits par hasard dans divers contextes où ils se sont avérés favorables aux gènes qui les ont fait naître, dans le sens où ces gènes pouvaient alors proliférer.

L'idée d'un esprit modulaire ressemble un peu à l'idée largement discréditée de la phrénologie: l'étude de la taille et de la forme du crâne était censée nous renseigner sur le caractère et les capacités mentales d'une personne. Le concept a été réhabilitée dans un schéma très différent par Fodor dans les années 1970 [22], abandonnant l'idée que les modules doivent nécessairement être localisés, c'est-à-dire correspondre à des régions spécifiques du cerveau. Sa théorie était que certaines activités cognitives spécifiques pourraient être gérées par un module spécialement conçu[3]—chaque module travaillerait de manière plutôt autonome sur le problème spécifique qu'il avait évolué pour résoudre, par exemple, la perception des couleurs, l'analyse des formes, l'analyse des relations dans l'espace, le guidage visuel des mouvements du corps, la reconnaissance faciale, l'analyse grammaticale du langage parlé, la reconnaissance vocale et bien d'autres. Mais selon Fodor, les modules ne traiteraient que des problèmes spécifiques, chacun n'impliquant qu'un certain type d'informations. C'est ce qu'on appelle l'encapsulation de l'information.

La version de la modularité proposée par Fodor était plutôt stricte [23]. Pour Fodor, un module devait être spécifique au domaine, encapsulé du point de vue de l'information, automatique et rapide. De nombreux types de traitement cognitif de l'information ne pourraient donc pas être modulaires—il est facile de trouver des activités cognitives qui ne semblent pas satisfaire ces conditions. Une chose que fait le cerveau et qui a des domaines d'application très généraux est le raisonnement. Après tout, quand on raisonne, on peut parcourir tous les sujets, et faire des liens entre des domaines très différents. Le raisonnement prende en compte notre vision du monde et il est sensible au contexte de plusieurs manières différentes. Il semble donc qu'il ne puisse y avoir d'encapsulation d'informations, et pour Fodor un simple module ne pourrait donc pas le gérer.

Fodor distingue ainsi le traitement de l'information modulaire et le traitement de l'information non-modulaire, qu'il appelle le traitement central. Et ce dernier lui posait un problème particulier. En fait, il est arrivé à une conclusion assez radicale. Selon lui, les processus globaux comme le raisonnement doivent rester à jamais inaccessibles aux sciences cognitives, une vision pessimiste qu'il appelait la première loi de Fodor. Mais comment arriver à cette conclusion? La réponse nous renvoie sur sa théorie du langage de la pensée (le mentalais). Cette hypothèse exige que les transitions entre les phrases inscrites dans nos cerveaux soient purement fonction des propriétés syntaxiques de ces phrases. Mais pour les processus cognitifs comme le raisonnement, ces transitions sont contextuelles, et elles ne peuvent donc pas être déterminées uniquement par les propriétés syntaxiques intrinsèques des représentations mentales impliquées. En bref, il a estimé que nous ne pourrions jamais comprendre le traitement de

[3]Comme précédemment, ce mot est utilisé dans le sens proposé par Dennett pour désigner la conception par mutation génétique et sélection naturelle, où il n'y a pas de concepteur.

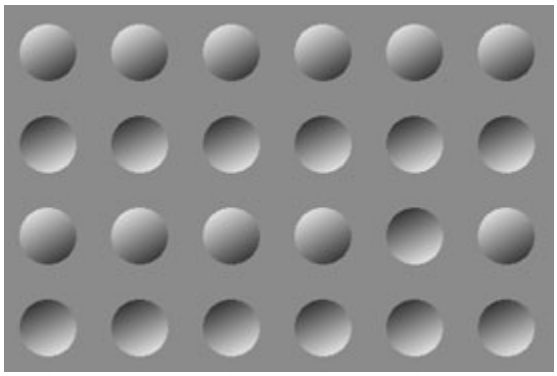

Figure 22: Que voit-on? La rangée du haut semble être composée de bosses alors que la prochaine serait composée de creux, mais c'est parce que votre esprit a fait une supposition sur la source de lumière—sans vous consulter

l'information contextuelle en termes informatiques.

Bermùdez suggère deux solutions [23]:

- La première consiste à se demander s'il existe vraiment des formes de traitement de l'information complètement générales par rapport au domaine d'application, donc capable potentiellement de s'appuyer sur n'importe quel type d'information—une question raisonnable malgré les remarques ci-dessus sur le raisonnement, comme nous le verrons.

- La seconde est de suggérer qu'il n'y a pas de réelle différence de nature entre le traitement modulaire et le traitement central, et en fait rien de tel qu'un traitement central au sens de Fodor—tout traitement peut en fait être modulaire. C'est ce qu'on appelle l'hypothèse de la modularité massive.

En tout cas, peu de psychologues cognitifs adhèrent à la première loi de Fodor, qui peut être considérée plus comme un problème pour le langage de la pensée—et l'hypothèse du symbole physique—que pour la science cognitive dans son ensemble.

Toutes ces considérations concernent la question de l'architecture de l'esprit: quels différents types de traitement de l'information s'y déroulent, dans quels différents types de structure, et avec quelles connexions entre ces différents types de structure? Mais une chose est claire: il est généralement admis que les modules jouent un rôle clé à cet égard. Et les modules en question sont darwiniens au sens décrit ci-dessus. L'idée est apparue en 1987 dans les travaux de l'anthropologue américain John Tooby et de la psychologue américaine Leda Cosmides [24]. Malheureusement, nous ne pouvons que frôler la surface de ce sujet fascinant.

Mais prenons quand même un peu de temps pour l'examiner plus attentivement. Selon l'hypothèse de la modularité massive, l'esprit humain est un en-

semble de modules spécialisés, dont chacun a évolué pour résoudre un problème spécifique rencontré par nos ancêtres, humains ou non. Chaque module s'occupe d'une application spécifique, ne traitant que les informations pertinentes à cette application. Il est autonome, rapide et inconscient. La vitesse est bien sûr essentielle dans de nombreuses situations pour l'efficacité de la réaction, et la survie elle-même peut être en jeu. Mais qu'en est-il de cette idée que nous ne sommes même pas conscients de ce que font nos modules? C'est l'une des grandes découvertes de la psychologie cognitive: nous sommes en fait conscients de peu des activités dans laquelles notre esprit est engagé. On considère généralement aujourd'hui que nous—le nous, ou le soi, de l'introspection—ne sommes qu'une construction à des fins de communication, et sommes à peine impliqués dans la plupart des activités cognitives, malgré l'impression que notre propre esprit veut nous donner.

Nous avons donc des modules qui font des inférences dans des domaines nombreux et variés, chacun profitant d'une certaine régularité dans la nature, c'est-à-dire quelque chose qui se présente toujours d'une certaine manière, ou quelque chose qui arrive toujours quand quelque chose d'autre vient de se passer. Et tout cela se passe sans même que nous nous en rendions compte. Regardez la figure 22, qui montre ce qui semble être une dalle de matériau avec des rangées de bosses et de creux. La plupart des gens voient la rangée du haut comme une rangée de bosses et la seconde comme une rangée de creux. C'est parce qu'un module dans notre esprit a supposé, sans nous consulter, que la dalle est éclairée d'en haut. Mais imaginez un instant—cela demande un effort—que la dalle soit éclairée par le bas. Maintenant, la rangée du haut devient une rangée de creux et la seconde une rangée de bosses. En fait, on peut passer d'une perception à l'autre, avec un effort suffisant. Naturellement, tous nos ancêtres, humains ou autres, auraient été habitués à ce que les choses soient éclairées d'en haut, par le soleil—une régularité dans notre monde.

Voici donc un point important. Un module fait une inférence, c'est-à-dire qu'il produit de nouvelles informations sur la base d'autres informations, mais il n'a besoin d'aucune raison pour justifier son inférence et il n'en fournit pas une. C'est en ce sens qu'il fonctionne automatiquement. Il raisonne sans raisonner. Ceci est à nouveau illustré dans le célèbre damier d'Adelson illustré à la figure 23. Nous n'avons aucune raison *consciente* de voir le carré A plus sombre que le carré B, mais nous voyons des choses comme ça. Les deux carrés sont identiques dans l'image—le niveau de gris est le même dans chacun (voir la figure 24)—mais notre cerveau est intervenu en apportant sa 'connaissance' des ombres.

Une autre expérience, ou plutôt type d'expérience, illustre magnifiquement comment notre cerveau est le juge final des informations qu'il reçoit de nos yeux, et sans aucun effort conscient de notre part. Ces expériences exploitent le 'paradigme de contingence du regard' [26]: pendant que quelqu'un regarde un écran d'ordinateur, il est possible d'informer l'ordinateur en temps réel de l'emplacement exact du regard de la personne, puis de brouiller tout le reste de l'écran sans que la personne ne soit jamais au courant. Ces techniques de suivi oculaire sont largement utilisées en psychologie pour étudier la façon dont nous voyons le monde. Par exemple, quelqu'un lit une histoire courte à l'écran

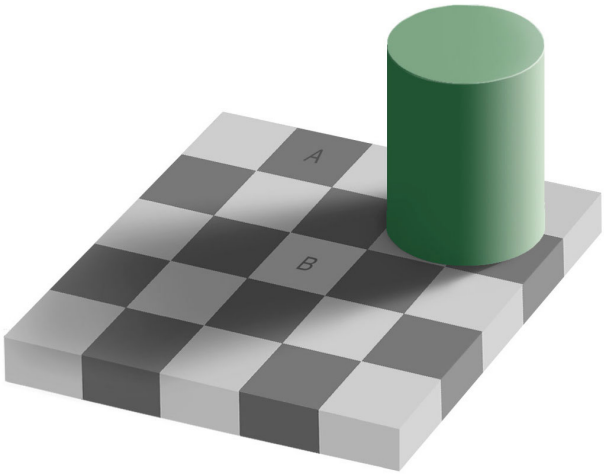

Figure 23: C'est peut-être difficile à croire, mais bien que A soit évidemment un carré sombre et B un blanc, le niveau de gris dans les deux carrés est le même. Jetez un œil à la figure 24 pour voir la démonstration [25]. Image: Edward H. Adelson, vectorisée par Pbroks13, CC BY-SA 4.0

et reste complètement inconscient du fait qu'à chaque instant presque tout le texte est brouillé. Quelqu'un qui se tient derrière le lecteur trouvera étonnant qu'il parvienne à en tirer le moindre sens.

Alors qu'est-ce que cela a à voir avec les modules? En fait, un module construit notre image interne de ce qui est visualisé. Dans cet exemple, l'image interne ressemble peu à ce qui est réellement là. C'est parce que le module en question 'suppose' que ce que nous étions en train de regarder ailleurs il y a un instant sera toujours là, une hypothèse naturelle dans le monde ordinaire de la nature. Il exploite une régularité dans le monde: les choses ne changent pas si vite.

Un autre exemple fascinant de module, qui montre que certains modules peuvent avoir évolué il y a très longtemps, ou peut-être avoir évolué plusieurs fois dans différentes espèces, concerne la façon dont les créatures vivantes se guident dans des espaces étroits sans toucher aux limites de ces espaces. Lorsque nous traversons un tunnel, les parois semblent reculer à une certaine vitesse. Si le mur de droite est plus proche que le mur de gauche, il semblera reculer plus rapidement. Cette expérience, connue sous le nom de flux optique, peut nous indiquer quel mur est le plus proche, et si nous souhaitons passer par le milieu du tunnel, il suffit d'ajuster notre trajectoire pour que le flux optique soit le même de chaque côté. Ceci est bien décrit par Sloman et Fernbach dans leur excellent livre *The Knowledge Illusion* [27, 28]:

> [Un] endroit où vous rencontrez un flux optique est sur l'autoroute. Les lignes sur la route peintes par le ministère des transports sont là pour vous garder sur le droit chemin. Tant que le flux de lignes

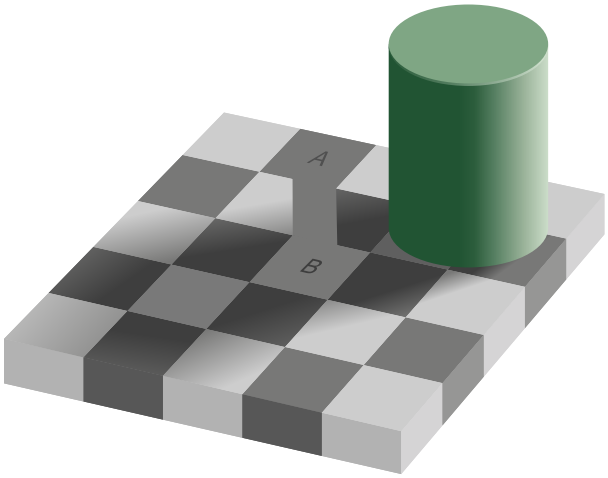

Figure 24: Le niveau de gris est le même dans les carrés A et B [25]. Image: Edward H. Adelson, vectorisée par Pbroks13, CC BY-SA 4.0

> d'un côté de vous semble se déplacer à la même vitesse que le flux de l'autre côté, vous resterez dans votre voie. Nous le savons grâce à des expériences faites dans des simulateurs de conduite. Si vous placez quelqu'un dans un simulateur avec un écran d'ordinateur et que les lignes vont plus vite d'un côté que de l'autre, la personne dérivera vers le côté avec les lignes les plus lentes.

Imaginez un robot GOFAI faisant tous les calculs nécessaires pour voler entre deux poteaux sans les toucher. Ce n'est pas un problème, bien sûr, si vous pouvez calculer assez rapidement. Mais le flux optique est tellement plus facile, à condition que vous disposiez d'un système pour convertir les estimations sensorielles du flux optique directement en effets moteurs.

Ce qui est fascinant, c'est que les abeilles utilisent aussi ce système [29]. On peut faire en sorte que les abeilles d'une ruche doivent passer par un tunnel pour se rendre dans les prairies, puis projeter une image animée du mur d'un côté. Si la paroi de droite semble reculer plus rapidement, les abeilles ne passent plus par le centre du tunnel, mais plutôt vers la gauche. Peut-être que ce module a évolué dans un ancêtre commun des abeilles et des humains, ou peut-être qu'il a évolué séparément plusieurs fois. C'est une astuce tellement utile, ce ne serait pas si surprenant.

Sloman et Fernbach encore [27, p. 100]:

> Toutes les études montrent que les gens (et les insectes) ne sont pas en permanence en train de faire des tas de calculs qui leur permettent de temps en temps d'agir selon un modèle compréhensif du monde. En réalité, les gens utilisent des faits constatés directement—comme l'optique des surfaces—pour simplifier ce qu'ils font.

Les modules évoluent ou se développent lorsqu'il existe une régularité dans le monde qui peut être exploitée pour tirer des inférences, et lorsqu'il est adaptatif de le faire. Ces inférences ne sont pas le résultat d'un raisonnement au sens où nous l'entendons normalement, comme vous pouvez le voir dans les exemples. Ce sont des conséquences directes d'une partie du cerveau sur les autres. Le cerveau n'applique aucune théorie. Cela ne nécessite aucune implication du soi ou l'accord de la pensée consciente par le soi.

Nous revenons donc à la question du raisonnement. Qu'est-ce que le raisonnement? Peut-il y avoir un module pour produire des raisons? Mercier et Sperber le proposent [16], malgré l'inquiétude de Fodor concernant le problème de généralité de domaine que nous avons vu ci-dessus. Ils prétendent bien que nous avons un module qui produit des raisons, souvent dans le contexte que nous avons décrit de justifier nos actions *a posteriori*, ou de persuader les autres d'aller dans notre sens, mais en tout cas généralement pour la consommation sociale. Dans leur conception, une raison est une représentation produite inconsciemment par le module de raisonnement. Elle apparait comme une intuition dans la conscience pour expliquer ou justifier quelque chose après coup. Le module de raisonnement est donc un mécanisme qui tire des inférences intuitives. Le raisonnement n'est pas une alternative à l'inférence intuitive—le raisonnement est une utilisation d'inférences intuitives sur les raisons. Chaque raison prend racine dans l'intuition. La régularité exploitée par ce module est la régularité des régularités dans le monde, le fait qu'il y ait toujours des régularités.

Et qu'en est-il du problème de généralité de domaine de Fodor? D'après Mercier et Sperber [16], le véritable domaine du module de raisonnement—les raisons—est en fait très spécifique. Mais ces raisons elles-mêmes peuvent concerner absolument toutes les combinaisons de choses dans le monde. Le raisonnement peut ainsi être spécialisé dans ses opérations et pourtant généralisé dans sa portée. On le voit dans les syllogismes évoqués plus haut: on peut insérer toutes les propositions qui nous passent par la tête.

Une conclusion?

La découverte clé des sciences cognitives jusqu'à présent est peut-être celle-ci: alors que nous ne nous connaissons qu'à travers ce dont nous sommes conscients via l'introspection, et nous supposons naturellement que les choses se passent vraiment selon le schéma que l'introspection et la conscience nous présentent, en réalité nous restons largement inconscients, non seulement de la plupart de ce qui se passe dans notre cerveau, mais aussi de la vraie nature des peu de choses dont nous sommes réellement conscients. Cela peut sembler plutôt négatif comme constat, mais il n'en est rien. Cela ouvre la voie pour contourner des schémas que nous appelons le soi, la conscience et le libre arbitre, que la nature a trouvés pour nous guider si bien dans nos vies de tous les jours. Comme ça nous pourrons peut-être voir ce qui est au-delà, la manière dont les choses se passent vraiment dans une perspective cognitive ou neurale, et enfin, comprendre pourquoi ces choses nous semblent comme elles nous semblent.

References

[1] L. Gamwell, M. Solms, *From Neurology to Psychoanalysis. Sigmund Freud's Neurological Drawings and Diagrams of the Mind*, Binghamton University Art Museum in cooperation with the Arnold Pfeffer Center for Neuro-Psychoanalysis at the New York Psychoanalytic Institute and the American Psychoanalytic Association (2006). See also the Guardian article https://www.theguardian.com/science/neurophilosophy/2014/mar/10/neuroscience-history-science

[2] A.M. Turing, On computable numbers, with an application to the entscheidungsproblem, Proceedings of the London Mathematical Society **s2-42**, 230–265 (1937)

[3] J.-Y. Girard, La machine de Turing: de la calculabilité à la complexité, dans *La machine de Turing*, Seuil, 1995

[4] https://sites.google.com/site/inteligenciaartificialist/exemplos-de-ia/shrdlu

[5] http://physique.unice.fr/sem6/2011-2012/PagesWeb/PT/Modelisation/neruone.html

[6] https://laskerfoundation.org/eric-kandel-learning-about-the-human-brain-from-sea-slugs/

[7] A. Rosenberg, *The Atheist's Guide to Reality. Enjoying Life without Illusions*, W.W. Norton, New York, London (2012)

[8] https://ganpatidiagnostics.com/index.php/electroencephalography/

[9] https://www.d.umn.edu/~jfitzake/Lectures/DMED/SpeechLanguage/CorticalS_LAreas/CorticalLanguageAreas.html

[10] http://tpe-ia.lescigales.org/maths.php

[11] R. Baillargeon, Object permanence in 3 1/2 and 4 1/2 month-old infants. Developmental Psychology **23** (5), 655 (1987)

[12] https://en.wikipedia.org/wiki/Centrifugal_governor

[13] R. Jackendoff, *A User's Guide to Thought and Meaning*, Oxford University Press (2015)

[14] D. Dennett, *From Bacteria to Bach and Back. The Evolution of Minds*, W.W. Norton, New York, London (2017)

[15] G.A. Miller, Trends and debates in cognitive psychology, Cognition **10**, 215–225 (1981)

[16] H. Mercier, D. Sperber, *L'Énigme de la raison*, Odile Jacob (2021)

[17] J. Haidt, *The Righteous Mind. Why Good People Are Divided by Politics and Religion*, Penguin (2013)

[18] R.B. Zajonc, Attitudinal effects of mere exposure, Journal of Personality and Social Psychology **9**, 1–27 (1968)

[19] S. Schnall, J. Haidt, G.L. Clore, A.H. Jordan, Disgust as embodied moral judgment, Personality and Social Psychology Bulletin **34**, 1096–1109 (2008)

[20] C.B. Zhong, B. Strejcek, N. Sivanathan, A clean self can render harsh moral judgment, Journal of Experimental Social Psychology **46**, 859–62 (2010)

[21] B. Englich, Blind or biased? Justitia's susceptibility to anchoring effects in the courtroom based on given numerical representations, Law and Policy **28**, no. 4, 497–514 (October 2006)

[22] J.A. Fodor, *The Modularity of Mind: An Essay on Faculty Psychology*, MIT Press, Cambridge, Massachusetts (1983)

[23] My remarks here follow closely the excellent account given in the introductory text J.L. Bermúdez, *Cognitive Science. An Introduction to the Science of the Mind*, 2nd edn, Cambridge University Press (2014)

[24] See, for example, L. Cosmides, J. Tooby, From evolution to behavior: Evolutionary psychology as the missing link. In J. Dupré (Ed.), *The Latest on the Best: Essays on Evolution and Optimality* (pp. 276–306), The MIT Press (1987)

[25] https://en.wikipedia.org/wiki/Checker_shadow_illusion

[26] https://en.wikipedia.org/wiki/Gaze-contingency_paradigm

[27] S. Sloman, P. Fernbach, *The Knowledge Illusion. The Myth of Individual Thought and the Power of Collective Wisdom*, Pan Books, London (2018)

[28] A.P. Duchon, W.H. Warren Jr, A visual equalization strategy for locomotor control. Of honeybees, robots, and humans, Psychological Science **13** (3), 272–278 (2002)

[29] M.V. Srinivasan, M. Lehrer, W.H. Kirchner, S.W. Zhang, Range perception through apparent image speed in freely flying honeybees, Visual Neuroscience **6** (5), 519–535 (1991)