COMMENT
UNE COMPRÉHENSION
NOUVELLE DU CERVEAU
CONDUIRA À LA CRÉATION
DE MACHINES
VÉRITABLEMENT
INTELLIGENTES

# INTELLIGENCE

JEFF HAWKINS
avec Sandra Blakeslee

CampusPress
www.pearsoneducation.fr

# INTELLIGENCE

**JEFF HAWKINS**
**AVEC SANDRA BLAKESLEE**



CAMPUSPRESS

L'édition originale de cet ouvrage a été publiée aux États-Unis par Henry Holt® sous le titre *On Intelligence*.
Copyright © 2004 by Jeff Hawkins and Sandra Blakeslee
ISBN original : 0-8050-7456-2

Traduction : Bernard Jolivalt

Mise en pages : MD Graphic

ISBN : 2-7440-1956-9
Copyright © 2005 Pearson Education France
Aucune représentation ou reproduction, même partielle, autre que celles prévues à l'article L. 122-5 2° et 3° a) du code de la propriété intellectuelle ne peut être faite sans l'autorisation expresse de Pearson Education France ou, le cas échéant, sans le respect des modalités prévues à l'article L. 122-10 dudit code.

# SOMMAIRE

*Prologue* 7

**1.** *L'intelligence artificielle* 17
**2.** *Les réseaux neuronaux* 33
**3.** *Le cerveau humain* 53
**4.** *La mémoire* 81
**5.** *Une nouvelle structure de l'intelligence* 103
**6.** *Le fonctionnement du cortex* 127
**7.** *Conscience et créativité* 207
**8.** *L'avenir de l'intelligence* 239

*Épilogue* 273
*Annexe : les prédictions testables* 275
*Bibliographie* 287
*Remerciements* 295
*Index* 299

# PROLOGUE

**Ce livre** et ma vie sont animés par deux passions.
Je m'intéresse depuis un quart de siècle à l'informatique mobile. Dans le domaine des hautes technologies de la Silicon Valley, je suis connu pour avoir fondé deux sociétés, Palm Computing et Handspring, et comme l'architecte de nombreux ordinateurs de poche et téléphones cellulaires comme le PalmPilot et le Treo.

Je nourris cependant une autre passion qui l'emporte sur celle de l'informatique, et qui me semble bien plus importante : le cerveau. Je veux comprendre comment il fonctionne, non seulement d'un point de vue psychique, pas uniquement d'une manière générale, mais d'une manière pratique, physicochimique. Je veux comprendre ce que sont l'intelligence et le fonctionnement du cerveau, mais surtout savoir comment fabriquer des machines fondées sur les mêmes principes. Je veux créer des machines véritablement intelligentes.

L'intelligence est sur terre la dernière grande frontière contre laquelle bute la science. La plupart des problèmes

scientifiques portent sur l'infiniment grand, l'infiniment petit ou ce qui s'est passé il y a des milliards d'années. En revanche, un cerveau, tout le monde en a un. Vous êtes votre cerveau. Pour comprendre ce que vous ressentez, comment vous percevez le monde, pourquoi vous commettez des erreurs, d'où provient votre créativité, pourquoi vous êtes sensible à la musique et aux arts, et en vérité ce qu'il en est d'être un humain, vous devez comprendre le fonctionnement de votre cerveau. Une théorie prouvée de l'intelligence et des fonctions cérébrales produirait des bénéfices sur le plan social qui s'ajouteraient aux applications purement médicales. Nous serions capables de concevoir des machines véritablement intelligentes, bien qu'elles ne ressemblassent sans doute pas aux robots des romans de science-fiction. En fait, les machines intelligentes seront issues d'un ensemble de principes découlant de la nature même de l'intelligence. Elles contribueront à améliorer notre connaissance du monde, à explorer l'univers et rendre notre environnement plus sûr et susciteront l'apparition de nouveaux et vastes secteurs industriels.

Fort heureusement, nous vivons à une époque qui permet d'envisager la résolution des problèmes liés à la compréhension de l'intelligence. Nous avons accès à d'énormes quantités d'informations concernant le cerveau, collectées depuis des centaines d'années et en perpétuel accroissement. A eux seuls, les Etats-Unis comptent des milliers de neurobiologistes. Nous n'avons pas encore élaboré de théorie productive sur ce que sont l'intelligence et le fonctionnement global du cerveau. La plupart des neurobiologistes ne s'attachent guère à une théorie globale car ils sont trop absorbés par leurs expériences et la collecte de données concernant les nombreux sous-systèmes du cerveau. Des foules de programmeurs se sont évertués à rendre les ordinateurs intelligents, en vain. J'estime qu'ils ne sont pas près de réussir tant qu'ils ignoreront les différences entre l'ordinateur et le cerveau.

Qu'est cette intelligence dont peut se prévaloir le cerveau, mais pas les ordinateurs ? Qu'est-ce qui fait qu'un gamin de six

ans peut sauter gracieusement d'un rocher à un autre dans le lit d'un ruisseau alors que nos robots les plus perfectionnés crapahutent maladroitement ? Pourquoi un enfant de trois ans maîtrise-t-il déjà bien le langage alors que les ordinateurs en sont incapables, en dépit du travail acharné des programmeurs depuis plus d'un demi-siècle ? Pourquoi faites-vous instantanément la différence entre un chien et un chat alors qu'un superordinateur en est incapable ? Ce sont là de grands mystères qui attendent une réponse. Nous avons beaucoup d'indices, mais ce qui manque, ce sont quelques aperçus essentiels.

Vous vous demandez sans doute pourquoi c'est un informaticien qui écrit un ouvrage sur le cerveau. Ou dit autrement, si le cerveau m'intéresse tant, pourquoi ne pas avoir fait carrière dans les neurosciences ou l'intelligence artificielle ? J'ai essayé plusieurs fois, mais je me suis toujours refusé à aborder le problème de l'intelligence de la même manière que d'autres le firent avant moi. A mon avis, la meilleure voie pour découvrir la solution consiste à s'appuyer sur la biologie du cerveau, tout en considérant l'intelligence comme un problème de programmation : une position quelque part entre la biologie et l'informatique… Nombre de biologistes tendent à rejeter ou ignorer une approche du cerveau en termes d'ordinateur, et beaucoup d'informaticiens croient volontiers qu'ils n'ont rien à apprendre des biologistes. Le monde de la science est aussi moins enclin à prendre des risques que le monde des affaires. Dans une entreprise technologique, un chercheur qui développe une idée nouvelle selon une approche raisonnée a tout à gagner au niveau de sa carrière, même si cette idée conduit à une impasse. Beaucoup de projets n'ont été couronnés de succès qu'après plusieurs échecs. Mais à l'université, consacrer deux années à une recherche qui n'aboutit pas peut ruiner définitivement une jeune carrière. C'est pourquoi j'ai eu deux passions dans ma vie. Je pensais que la réussite en entreprise me permettrait de réussir dans la compréhension du cerveau. Il me fallait des moyens pour financer mes recherches. Je voulais et je devais apprendre ce

qui change le monde, comment vendre des idées nouvelles. Je comptais pour cela sur mon travail à la Silicon Valley.

En août 2002, j'ai créé le Redwood Neuroscience Institute (RNI), un centre de recherches sur le cerveau. Il existe de nombreux instituts de neurosciences de par le monde, mais aucun n'est spécialisé dans la recherche d'une connaissance théorique générale du néocortex, la partie de l'encéphale où siège l'intelligence. C'est cela que nous étudions au RNI. A bien des égards, cette entreprise est une start-up : nous poursuivons un rêve qui semble hors de portée. Mais nous avons la chance d'être nombreux et nos efforts commencent à porter leurs fruits.

---

Le propos de cet ouvrage est ambitieux : il présente une théorie générale du fonctionnement du cerveau, la nature de l'intelligence et comment le cerveau la produit. La théorie que je propose n'est pas toute nouvelle. Un grand nombre de concepts préexistaient sous une forme ou sous une autre, mais ils n'avaient jamais été mis en perspective d'une manière cohérente. Il est bien connu que les idées nouvelles sont souvent des idées anciennes rhabillées et réinterprétées. C'est sans doute vrai pour les théories exposées ici, mais ce rhabillage et cette réinterprétation font la différence entre une accumulation de notions disparates et une théorie bien construite. Je pense que cela ne vous échappera pas. La réaction que j'entends le plus fréquemment est : « Ça tient la route. Je n'aurais jamais pensé à l'intelligence sous cette forme, mais votre vision semble cohérente. » Du coup, le lecteur porte un regard différent sur lui-même. Il observe son propre comportement, conscient de ce qui se passe dans son cerveau. J'espère bien qu'après avoir lu ce livre vous aurez une conscience nouvelle du pourquoi vous pensez à ce que vous pensez, et pourquoi vous vous comportez de telle ou telle manière. J'espère aussi que des lecteurs mettront à profit les principes esquissés dans ces pages pour se consacrer à l'élaboration de machines intelligentes.

Je fais souvent référence à ces principes ainsi qu'à mon approche de l'étude de l'intelligence au travers des mots « intelligence réelle », afin de la distinguer de l'« intelligence artificielle ». Les spécialistes de l'intelligence artificielle ont tenté de programmer les ordinateurs pour les faire agir comme des êtres humains sans même s'interroger d'abord sur ce qu'est l'intelligence, sur ce qu'elle implique. Pour construire des machines intelligentes, ils ont délaissé l'élément le plus important : l'intelligence. L'« intelligence réelle » stipule qu'avant de vouloir fabriquer des machines intelligentes il faut d'abord comprendre le fonctionnement du cerveau ; il n'y a là rien d'artificiel. Ce n'est qu'ensuite que nous nous demanderons comment concevoir les machines.

L'ouvrage commence par expliquer pourquoi les tentatives passées pour comprendre l'intelligence et inventer des machines intelligentes ont échoué. Je présenterai et développerai ensuite l'idée centrale de ma théorie, ce que j'appelle un cadre de mémoire-prédiction. Au Chapitre 6, j'explique en détail comment l'encéphale implémente le modèle de mémoire-prédiction, autrement dit comment le cerveau fonctionne véritablement. Puis j'aborderai les implications sociales et autres de la théorie qui pour de nombreux lecteurs seront celles qui les inciteront le plus à la réflexion. L'ouvrage s'achève par un exposé sur les machines intelligentes, comment elles seront construites et ce que sera l'avenir. J'espère bien que vous serez captivé. Voici quelques questions qui seront abordées :

*Les ordinateurs peuvent-ils être intelligents ?*
Pendant des décennies, les spécialistes de l'intelligence artificielle ont affirmé que les ordinateurs deviendront intelligents lorsqu'ils seront suffisamment puissants. Ce n'est pas mon avis et je vous expliquerai pourquoi. Le cerveau et l'ordinateur sont en effet fondamentalement différents.

*Les réseaux neuronaux n'étaient-ils pas censés régir les machines intelligentes ?*

Le cerveau est évidemment constitué d'un réseau de neurones, mais sans une connaissance préalable de son fonctionnement, les réseaux neuronaux simples ne réussiront pas à créer des machines intelligentes là où les programmes informatiques ont échoué.

*Pourquoi le fonctionnement du cerveau a-t-il été si difficile à découvrir ?*

Beaucoup de scientifiques soutiennent qu'en raison de sa complexité il nous faudra beaucoup de temps pour comprendre le cerveau. La complexité découle de la confusion, elle n'en est pas la cause. Je maintiens que quelques hypothèses fondées sur l'intuition, mais erronées, nous ont induits en erreur. L'erreur la plus flagrante est de croire que l'intelligence est caractérisée par un comportement intelligent.

*Qu'est l'intelligence, si elle n'est pas caractérisée par le comportement ?*

Le cerveau accapare une grande partie de la mémoire pour élaborer un modèle du monde. Tout ce que vous connaissez et avez appris y est contenu. Le cerveau se sert de ce modèle fondé sur la mémoire pour procéder à d'incessantes prédictions des événements à venir. C'est cette capacité à se projeter dans le futur qui est la clé de voûte de l'intelligence. Je décrirai en profondeur la capacité de prédiction du cerveau ; elle constitue l'idée centrale de ce livre.

*Comment le cerveau fonctionne-t-il ?*

L'intelligence siège dans le néocortex. Même si ses capacités sont nombreuses et sa flexibilité remarquable, le néocortex est étonnamment constant au niveau de ses détails structurels. Qu'elles régissent la vision, l'ouïe, le toucher ou le langage, les

différentes parties du néocortex fonctionnent toutes selon des principes identiques. La connaissance du néocortex passe par la connaissance de ces principes communs, en particulier de leur structure hiérarchique. Nous l'examinerons suffisamment en détail pour découvrir comment sa structure se calque sur celle du monde.

*Quelles sont les implications de cette théorie ?*
Cette théorie du cerveau peut contribuer à expliquer entre autres d'où provient la créativité, pourquoi nous avons conscience de quelque chose, la source de nos préjugés, comment nous apprenons et pourquoi il nous est plus difficile, en avançant en âge, d'apprendre de nouvelles choses. Cette théorie permet de mieux comprendre qui nous sommes et pourquoi nous agissons comme nous le faisons.

*Est-il possible de fabriquer des machines intelligentes et que sauront-elles faire ?*
Oui, c'est possible et nous le ferons. Au cours des prochaines décennies, les capacités de ces machines devraient évoluer rapidement dans des domaines fort intéressants. Des gens craignent que les machines puissent mettre l'humanité en péril, mais je m'oppose fermement à cette idée. Les robots ne nous supplanteront pas. Il est beaucoup plus facile de créer des machines qui nous sont supérieures dans des domaines très évolués comme la physique ou les mathématiques, que d'en fabriquer qui sachent marcher, comme les robots des films de science-fiction. Nous explorerons les incroyables directions qui s'ouvrent à cette technologie.

Mon but est d'expliquer cette nouvelle théorie de l'intelligence et de montrer d'une manière accessible à tous comment le cerveau fonctionne. Une bonne théorie doit être facile à assimiler, débarrassée de tout jargon et de toutes démonstrations alambiquées. Je

commencerai par des généralités avant d'entrer dans les détails. Certaines seront fondées sur la logique, d'autres sur des aspects particuliers de l'ensemble des circuits cérébraux. Certaines de mes hypothèses se révéleront sans doute erronées, mais c'est inévitable, quel que soit le domaine scientifique. Il faut des années pour qu'une théorie arrive à maturité, ce qui n'affecte en rien la force de l'idée centrale.

---

Quand j'ai commencé à m'intéresser au cerveau, il y a des années de cela, j'ai recherché un ouvrage de référence sur son fonctionnement chez mon libraire. Lycéen, j'avais l'habitude d'y trouver tout ce que je désirais sur n'importe quel sujet : la théorie de la relativité, les trous noirs, la magie ou les mathématiques. Bref, tout ce qui m'intéressait à l'époque. Mais je ne découvris aucun titre satisfaisant concernant le cerveau. J'en déduisis que personne ne savait au juste comment il fonctionne. Aucune théorie, fut-elle bonne ou mauvaise, n'était proposée. Rien. C'était inhabituel. A une époque où personne ne savait comment les dinosaures avaient disparu, les théories ne manquaient pas, ni les ouvrages qui leur étaient consacrés. Mais sur le cerveau, rien. Je ne pouvais le croire. Savoir que nous ne savions rien sur cet organe vital me tourmentait. En étudiant ce que nous en connaissions, j'en vins à penser qu'il devait y avoir une explication plus limpide. Le cerveau n'a rien de sorcier et il me semblait que les réponses elles aussi ne devaient pas être bien compliquées. Le mathématicien Paul Erdös était persuadé que les démonstrations les plus simples existent dans quelque livre éthéré et que la tâche du mathématicien est de les découvrir, de « lire le livre ». Dans le même esprit, je pensais que l'explication de l'intelligence « était dans l'air ». Je la sentais. Je voulais « lire le livre ».

Durant toutes ces vingt-cinq dernières années, j'avais en tête ce petit livre limpide consacré au cerveau. C'était la carotte qui me motivait. Il est devenu l'ouvrage que vous tenez à présent entre vos mains. Je n'ai jamais aimé la complexité, que ce soit en sciences ou

dans les technologies. Vous le constatez dans les produits que j'ai conçus, dont la facilité d'emploi a souvent été reconnue. Les objets les plus puissants sont souvent les plus simples. C'est pourquoi ce livre propose une théorie de l'intelligence simple et sans détour. J'espère que vous l'apprécierez.

# 1

# L'INTELLIGENCE ARTIFICIELLE

**Quand** j'ai quitté l'université de Cornell, en juin 1979, diplômé d'ingénierie électrique, je n'avais pas encore prévu ce que je ferais de ma vie. J'ai d'abord été ingénieur sur le nouveau campus Intel à Portland, dans l'Oregon. La micro-informatique était encore à ses débuts et Intel se trouvait au cœur de ce secteur d'avenir. Ma tâche consistait à analyser et corriger les problèmes découverts par d'autres ingénieurs qui travaillaient dans le domaine d'activité principal de la société : les ordinateurs monocarte (placer la totalité d'un ordinateur sur une seule carte n'a été possible que grâce à l'invention, par Intel, du microprocesseur). J'ai publié une lettre d'information, voyagé quelque peu, et j'ai eu la chance de rencontrer des clients. J'étais jeune et la vie était belle, bien que j'aie perdu de vue l'élue de mon cœur, une collègue qui accepta un poste à Cincinnati.

Quelques mois plus tard, un événement changea le cours de ma vie : la lecture d'un numéro du magazine *Scientific American* entièrement consacré au cerveau. Il raviva mon

intérêt d'antan pour la matière grise. J'étais fasciné. Il y était question de l'organisation, du développement et de la chimie du cerveau, des mécanismes neuronaux de la vision, du mouvement et autres activités, et des fondements biologiques des désordres mentaux. C'était le numéro le plus génial de tous les temps. Plusieurs neurobiologistes que je rencontrai par la suite m'avouèrent que sa lecture joua un rôle déterminant dans le choix de leur carrière, comme ce fut le cas pour moi.

L'article final, « Réflexions sur le cerveau », était rédigé par Francis Crick, l'un des codécouvreurs de l'ADN qui avait ensuite appliqué son savoir à l'étude du cerveau. Il reconnaissait qu'en dépit d'une accumulation régulière de nombreuses découvertes parcellaires, le fonctionnement du cerveau était encore un profond mystère. Les scientifiques n'aiment pas reconnaître qu'ils ne savent pas, mais Crick n'en avait cure. Il était comme l'enfant qui montre que le roi est nu. Selon Crick, la neurobiologie était un ensemble de données que ne soutenait aucune théorie. Les termes exacts étaient que « ce qui manque cruellement est un ensemble d'idées structurées ». Pour moi, c'était une manière diplomatique de dire que nous ne savions pas du tout comment tout cela fonctionne. C'était vrai et ça l'est toujours.

Les mots de Crick furent pour moi un cri de ralliement. Mon désir de toujours de comprendre le cerveau et construire des machines intelligentes prenait vie. Bien que j'eusse à peine quitté l'université, je décidais de changer de carrière. J'étudierais le cerveau non seulement pour en comprendre le fonctionnement, mais aussi pour faire de ce savoir le fondement de technologies nouvelles qui permettraient de créer des machines intelligentes. Mettre tout cela en œuvre prendrait du temps.

Au printemps 1980, je rejoignis le bureau d'Intel situé à Boston afin de me rapprocher de ma future épouse qui était encore étudiante. J'étais chargé d'enseigner aux clients et aux techniciens l'art de concevoir des systèmes à base de microprocesseurs. Mais mes pensées étaient ailleurs : je ne cessais de tenter d'élaborer une

théorie du cerveau. L'ingénieur qui est en moi avait compris qu'après avoir découvert le fonctionnement du cerveau nous pourrions le reproduire ; il allait de soi que la construction d'un cerveau artificiel passerait tout naturellement par le silicium. Je travaillais en effet pour la société qui avait inventé le composant de mémoire à base de silicone ainsi que le microprocesseur. Peut-être pourrais-je convaincre Intel de consacrer une partie de mon temps à la conception de composants de mémoire calqués sur le cerveau. J'écrivis au président d'Intel, Gordon Moore, une lettre qui se résumait à cela :

> Cher Docteur Moore,
> Je propose la création d'un groupe de recherche chargé de découvrir le fonctionnement du cerveau. Il peut commencer avec une seule personne, moi, puis s'étoffer. Je suis certain que nous réussirons. Un jour, cela rapportera beaucoup d'argent.
>
> —Jeff Hawkins

Moore me mit en contact avec Ted Hoff, chef du service de recherche et de développement. Je m'envolai pour la Californie afin de le rencontrer et lui exposer mon projet d'étude du cerveau. Hoff était connu pour deux raisons : la première, que je connaissais, était son apport à la conception du premier microprocesseur. La seconde, qui m'était inconnue, était son travail dans l'élaboration des premières théories de réseaux neuronaux. Hoff avait déjà une expérience des neurones artificiels et de ce que l'on pouvait en faire ; je ne m'attendais pas à cela. Après avoir écouté ma proposition, il me dit qu'il ne croyait pas que le fonctionnement du cerveau puisse être dévoilé dans un avenir prévisible, et qu'Intel n'avait aucune raison d'encourager mon projet. Il n'avait pas tort, car vingt-cinq ans plus tard des progrès significatifs dans la compréhension du cerveau commencent tout juste à poindre. Mais à l'époque, j'étais plutôt déçu.

J'ai tendance à rechercher la voie la moins conflictuelle pour atteindre mes objectifs. Travailler sur le cerveau sous l'égide

d'Intel aurait été l'idéal. Cette option ayant échoué, j'en cherchais une autre. Je décidai de m'inscrire au Massachusetts Institute of Technology (MIT), célèbre pour ses travaux sur l'intelligence artificielle et situé pas très loin de chez moi. Cela me convenait parfaitement. J'avais une grande expérience de l'informatique et je désirais concevoir des machines intelligentes. Mais d'abord, il me fallait étudier le cerveau afin d'en découvrir le fonctionnement. Et c'est là que résidait le problème, car pour les chercheurs du laboratoire d'intelligence artificielle du MIT cette démarche était vouée à l'échec.

C'était comme si j'avais foncé dans un mur de briques. Le MIT était le temple de l'intelligence artificielle. Lorsque j'avais posé ma candidature, des dizaines de chercheurs brillants s'y trouvaient déjà, passionnés à l'idée de programmer des ordinateurs afin de produire un comportement intelligent. Pour eux, la vision, le langage, la robotique et les mathématiques se réduisaient à des problèmes de programmation. Les ordinateurs étant capables de faire tout ce que le cerveau sait faire, et même davantage, pourquoi obliger la pensée à s'accommoder de la complexité biologique de cet ordinateur qu'est la nature ? L'étude du cerveau limiterait la réflexion. Ils étaient persuadés qu'il est préférable d'aller aussi loin dans les calculs que le permettraient les calculateurs numériques. Leur Graal était d'écrire des programmes informatiques capables d'égaler puis de dépasser les capacités humaines. Ils avaient adopté une approche du type « la fin justifie les moyens ». Le fonctionnement du cerveau ne les intéressait guère. Certains s'enorgueillissaient même d'ignorer la neurobiologie.

C'était pour moi, de toute évidence, la manière la plus erronée d'aborder le problème. Je pressentais intuitivement que l'approche par l'intelligence artificielle ne parviendrait pas à créer des programmes capables de réaliser ce que savent faire les humains, ni de nous apprendre ce qu'est l'intelligence. Les principes qui régissent l'ordinateur et le cerveau sont complètement différents. L'un est programmé, l'autre est auto-apprenant. L'un

doit réaliser les tâches parfaitement et sans faille, l'autre est naturellement souple et tolérant aux échecs. L'un est équipé d'un microprocesseur central, l'autre est dépourvu de contrôle centralisé. La liste des différences est interminable. La principale raison qui m'incitait à penser que les ordinateurs ne deviendraient jamais intelligents est que j'en connaissais le fonctionnement, jusqu'au niveau de la physique des transistors ; ceci me laissait à penser que le cerveau et l'ordinateur sont fondamentalement différents. Je ne pouvais en apporter la preuve, mais j'en étais persuadé. J'en déduisis que l'intelligence artificielle peut certes favoriser l'invention d'objets utiles, mais qu'elle serait inapte à produire des machines intelligentes.

En revanche, je désirais comprendre les mécanismes de l'intelligence réelle et ceux de la perception, et étudier la physiologie et l'anatomie du cerveau afin de relever le défi de Francis Crick et mettre en évidence la structure du fonctionnement cérébral. Je me suis tout particulièrement attaché au néocortex, la partie du cerveau des mammifères qui s'est développée le plus récemment, et qui est le siège de l'intelligence. Ce n'est qu'après avoir compris son fonctionnement que nous pourrons élaborer des machines intelligentes, pas avant.

Malheureusement, les professeurs et les étudiants que j'avais rencontrés au MIT ne partageaient pas mes points de vue. Ils estimaient qu'il n'était pas nécessaire de connaître le cerveau pour comprendre ce qu'est l'intelligence et fabriquer des machines intelligentes. C'est du moins ce qu'ils me soutinrent. En 1981, l'université rejeta ma candidature.

---

**B**eaucoup de gens pensent que l'intelligence artificielle est à nos portes et qu'elle attend uniquement que les ordinateurs soient devenus plus puissants pour délivrer ses nombreuses promesses. Lorsque la mémoire et la puissance de calcul des ordinateurs seront suffisantes, croient-ils, les programmeurs spécialisés dans

l'intelligence artificielle pourront enfin produire des machines intelligentes. Je ne suis pas d'accord. L'intelligence artificielle souffre d'un défaut fondamental en ce sens qu'elle est incapable de s'adresser adéquatement à ce qu'est l'intelligence ou de révéler le mécanisme de la cognition. Un bref retour sur l'histoire de cette discipline et les principes qui la sous-tendent explique pourquoi elle a dévié de son but.

L'intelligence artificielle est née avec l'informatique. L'un de ses promoteurs fut le mathématicien anglais Alan Turing, l'un des inventeurs de l'ordinateur polyvalent. Son idée de génie fut la démonstration formelle du concept de calcul universel : tous les ordinateurs sont équivalents quels que soient les détails de leur fabrication. Pour soutenir cette argumentation, il conçut une machine imaginaire composée de trois parties essentielles : un système de traitement, une bande de papier et un mécanisme qui lit et écrit des symboles sur le ruban de papier qui va et vient. La bande servait à stocker des informations, à l'instar des 0 et des 1 auxquels se réduit tout programme informatique. Les composants de mémoire et les disques durs n'ayant pas encore été inventés, Turing avait imaginé de stocker les données sur des bandes de papier. Le mécanisme, que nous appelons aujourd'hui CPU (*Central Processing Unit*, unité de traitement centrale, ou microprocesseur), appliquait un ensemble de règles définies pour lire et modifier l'information figurant sur la bande. Turing prouva mathématiquement qu'en choisissant un ensemble adéquat de règles pour le CPU et qu'en lui adjoignant une bande de longueur indéfinie, la machine serait capable d'exécuter n'importe quel ensemble d'opérations défini. Elle serait l'une parmi les nombreuses machines équivalentes désormais appelées Machines de Turing Universelles. Qu'il s'agisse d'extraire une racine carrée, de calculer une trajectoire balistique, de jouer à un jeu, de retoucher des images ou de vérifier des transactions bancaires, tout se ramène à des 0 et à des 1, et toute Machine de Turing peut être programmée afin de se charger de ces opérations. Le traitement

des données est en effet du traitement des données qui est lui-même du traitement des données : tous les calculateurs numériques sont logiquement équivalents.

Les conclusions de Turing étaient indiscutablement vraies et se révélèrent phénoménalement fructueuses. La révolution informatique et tout le secteur industriel qui en est issu en découlent directement. Turing se demanda ensuite comment construire une machine intelligente. Il estimait que les ordinateurs pourraient être intelligents, mais refusa de s'engager dans la discussion de la faisabilité. Estimant qu'il ne pouvait définir formellement l'*intelligence*, il n'essaya même pas. Il proposa à la place de prouver l'existence de l'intelligence au travers du célèbre test de Turing : si un ordinateur parvient à tromper un interrogateur humain auquel il fait croire que lui, l'ordinateur, est une personne, c'est donc que par définition l'ordinateur est intelligent. Et c'est ainsi que, le test de Turing servant de critère de quantification et la Machine de Turing comme support matériel, Alan Turing contribua à lancer le domaine de l'intelligence artificielle. Le dogme principal stipulait que le cerveau n'est qu'une autre sorte d'ordinateur. Qu'importe la façon dont vous élaborez un système d'intelligence artificielle, il lui suffit de reproduire un comportement humain.

Les partisans de l'intelligence artificielle ont mis en parallèle le calcul et la pensée. Ils soutenaient que « les exploits les plus impressionnants de l'intelligence humaine impliquent clairement la manipulation de symboles abstraits. Or, c'est ce que fait l'ordinateur. Que faisons-nous lorsque nous parlons ou écoutons ? Nous manipulons des symboles mentaux appelés mots, organisés selon des règles grammaticales précises. Que faisons-nous lors d'une partie d'échecs ? Nous utilisons des symboles mentaux représentant les propriétés ainsi que l'emplacement des diverses pièces. Que faisons-nous lorsque nous voyons ? Nous utilisons des symboles mentaux qui représentent des objets, leur position, leur nom, etc. Il est certain que tout ceci se produit dans le

cerveau, et non au travers des ordinateurs dont nous disposons actuellement, mais Turing avait démontré que la manière dont les symboles sont implémentés ou manipulés importe peu. Vous pouvez obtenir ces résultats par un assemblage de rouages et d'engrenages, par un système de commutateurs électroniques ou par le réseau de neurones présent dans le cerveau. Qu'importe du moment que le mécanisme est capable de réaliser l'équivalent fonctionnel de la Machine de Turing Universelle. »

Cette hypothèse fut renforcée par une communication scientifique très remarquée, publiée en 1943 par le neurophysiologiste Warren McCulloch et le mathématicien Walter Pitts. Ils décrivaient comment les neurones sont capables d'effectuer des fonctions numériques, c'est-à-dire comment des cellules nerveuses peuvent vraisemblablement reproduire la logique formelle qui s'exerce au cœur des ordinateurs. L'idée était que les neurones sont capables d'agir en tant que portes logiques, comme les appellent les ingénieurs. Une porte logique implémente des opérations logiques simples ET, NON ou OU. Les composants informatiques sont constitués de millions de portes logiques câblées entre elles sous la forme de circuits complexes précis. Un microprocesseur n'est qu'un ensemble de portes logiques.

McCulloch et Pitts montrèrent que des neurones pouvaient aussi être connectés entre eux de manière à effectuer des opérations logiques. Comme ils échangent des signaux et que le traitement des signaux d'entrée décide de l'envoi ou non d'un signal de sortie, les neurones sont sans doute des portes logiques vivantes. Les deux chercheurs en déduisirent que le cerveau pourrait être formé de « portes ET », de « portes OU » et d'autres éléments logiques, tous constitués de neurones, par analogie directe avec le câblage des circuits électroniques numériques. Nous ne savons pas si McCulloch et Pitts pensaient véritablement que le cerveau fonctionne de cette manière. Ils affirmaient seulement que c'est possible. D'un point de vue logique, cette vision des neurones était défendable. Théoriquement, les neurones sont capables

d'implémenter des fonctions numériques. Mais personne ne s'est risqué à soutenir que c'est ainsi que les neurones sont réellement câblés dans le cerveau. Tout le monde considéra cette communication comme la preuve, en dépit du manque d'évidences biologiques, que le cerveau n'est qu'une autre sorte d'ordinateur.

Il est aussi intéressant de remarquer que les idées de l'intelligence artificielle étaient étayées par une tendance dominante de la psychologie au cours de la première moitié du XX^e^ siècle : le comportementalisme. Pour les comportementalistes, il était impossible de savoir ce qui se passe dans le cerveau, qu'ils considéraient comme une impénétrable boîte obscure. Il était cependant possible d'observer et de mesurer l'environnement d'un animal et ses comportements : ce qu'il perçoit et ce qu'il fait, ses signaux en entrée (input) et en sortie (output). Ils découvrirent que le cerveau contient des mécanismes réflexes susceptibles de conditionner un animal afin qu'il adopte de nouveaux comportements selon un système de récompenses et de punitions. Mais il ne leur était pas nécessaire d'étudier le cerveau, notamment les sensations subjectives aussi compliquées que la faim, la crainte ou ce qu'est la cognition. Inutile de dire que ces théories se sont quelque peu défraîchies au cours de la seconde moitié du XX^e^ siècle ; mais l'intelligence artificielle y a adhéré beaucoup plus longuement.

A la fin de la Seconde Guerre mondiale, lorsque des calculateurs numériques furent disponibles pour des applications plus ambitieuses, les pionniers de l'intelligence artificielle retroussèrent leurs manches et se mirent à la programmation. La traduction de langues ? Facile, c'est comme décrypter un code : il suffit de reporter chacun des symboles d'un Système A dans sa contrepartie d'un Système B. La vision ? C'est facile également. Nous connaissons d'ores et déjà les théorèmes géométriques qui régissent la rotation, la mise à l'échelle et le déplacement des objets, et nous savons les programmer sous la forme d'algorithmes informatiques. Dès lors, la moitié du chemin était faite. Tous les

experts de l'intelligence artificielle clament haut et fort que l'intelligence des ordinateurs égalerait rapidement celle des êtres humains et la dépasserait.

Curieusement, le programme informatique qui se rapproche au mieux du test de Turing, un logiciel nommé Eliza, imite un psychanalyste qui reformule les questions posées. Si quelqu'un tape « Mon petit ami et moi, nous ne nous parlons plus », Eliza répondra par exemple : « Dis-m'en davantage sur ton ami » ou encore « Qu'est-ce qui te fait croire que ton ami et toi, vous ne vous parlez plus ? ». Conçu pour divertir, le programme réussit à tromper des gens, bien qu'il soit un peu bête et futile. Des travaux plus sérieux ont porté sur des programmes comme Blocks World (« le monde des blocs »), une simulation de chambre comprenant des volumes de différentes formes et couleurs. Vous pouviez poser à Blocks World des questions comme « Une pyramide verte est-elle placée sur le grand cube rouge ? » ou ordonner de « poser le cube bleu sur le petit cube rouge ». Le programme répondait à votre question ou tentait d'exécuter votre demande. Tout était simulé et fonctionnait. Mais le contexte était limité au monde des blocs éminemment artificiel. Les programmeurs ne purent généraliser ce concept à quoi que ce soit d'utile.

Le public était impressionné par une succession d'apparentes réussites et d'anecdotes sans cesse renouvelées sur les technologies de l'intelligence artificielle. Un programme qui fit sensation était capable de résoudre des théorèmes mathématiques. Jamais depuis Platon les inférences déductives multipas n'avaient été considérées comme le pinacle de l'intelligence humaine, à tel point qu'il sembla d'abord que l'intelligence artificielle avait visé juste. Mais, à l'instar de Blocks World, le programme s'avéra fort limité. Il ne savait résoudre que les théorèmes très simples déjà connus. On parlait aussi beaucoup des « systèmes experts », des bases de données de connaissances capables de répondre aux questions posées par les utilisateurs. Par exemple, un système expert médical savait diagnostiquer une maladie d'après une liste

de symptômes. Mais là encore, leur usage se révéla limité et ils ne manifestaient rien qui puisse se comparer à une intelligence globale. Les ordinateurs sont capables de jouer aux échecs à un niveau égal à celui des grands maîtres, comme Deep Blue d'IBM, célèbre pour avoir battu le champion du monde Gary Kasparov. Mais ces succès étaient vains. Deep Blue n'avait pas gagné en surclassant l'intelligence humaine, mais parce qu'il calcule des millions de fois plus vite que l'homme. Deep Blue était dépourvu d'intuition. Un joueur humain expert regarde les pièces sur l'échiquier et voit immédiatement quelles zones du jeu peuvent être exploitées ou lesquelles sont risquées, alors qu'un ordinateur n'a aucun sens inné de ce qui est important ; il est obligé d'envisager de très nombreuses autres options. De plus, Deep Blue n'avait aucune notion de l'historique du jeu et ne savait rien de son adversaire. Il jouait aux échecs sans rien y comprendre, à la manière d'une calculette qui sait faire des opérations sans rien connaître à l'arithmétique.

Tous les programmes d'intelligence artificielle n'étaient bons que pour la tâche pour laquelle ils avaient été spécifiquement conçus. Leur usage ne pouvait pas être généralisé et ils manquaient de souplesse ; leurs créateurs eux-mêmes reconnaissaient qu'ils ne pensaient pas comme des humains. Des problèmes d'intelligence artificielle, qui semblaient faciles de prime abord, ne trouvèrent jamais de solution. Aujourd'hui encore, aucun ordinateur n'est capable de comprendre un langage aussi bien qu'un enfant de trois ans, ou de voir aussi efficacement qu'une souris.

Après des années d'efforts, de promesses non tenues et de demi-succès, l'intelligence artificielle perdit peu à peu de son lustre. Les chercheurs se tournèrent vers d'autres domaines. Les financements se raréfiant, des start-up firent faillite. Programmer des ordinateurs pour les tâches les plus élémentaires de la perception, du langage et du comportement commença à paraître impossible. Cela n'a pas beaucoup changé aujourd'hui. Comme je l'ai mentionné précédemment, des chercheurs sont encore persuadés que

les problèmes d'intelligence artificielle peuvent être résolus à l'aide d'ordinateurs plus rapides, mais la plupart des scientifiques estiment que cette démarche est totalement erronée.

Ne blâmons pas les pionniers de l'intelligence artificielle à cause de leurs échecs. Alan Turing était un personnage brillant. Tous étaient certains que la Machine de Turing changerait le monde et c'est ce qui se produisit, mais pas par l'intelligence artificielle.

---

Mon scepticisme envers les assertions concernant l'intelligence artificielle s'était manifesté au moment où je posais ma candidature au MIT. A cette époque, John Searle, un professeur de philosophie émérite de l'université de Berkeley, en Californie, affirmait que les ordinateurs n'étaient pas et ne pourraient jamais être intelligents. Pour le prouver, il mit au point, en 1980, une expérience de pensée nommée « la Chambre chinoise ». En voici le principe.

Imaginez une chambre dans laquelle un Français est installé à un bureau. Une ouverture est ménagée dans le mur. Le personnage possède un gros livre rempli d'instructions ainsi qu'une grande quantité de crayons et de papier brouillon. En feuilletant le livre, il découvre que les instructions rédigées en français expliquent comment manier, trier et comparer des idéogrammes chinois. Remarquez que les instructions ne disent rien au sujet de la signification de ces idéogrammes ; elles indiquent seulement comment ils doivent être copiés, effacés, réordonnés, transcrits, etc.

Quelqu'un glisse une feuille de papier par l'ouverture dans le mur. Une histoire est écrite dessus, ainsi que des questions à propos de cette histoire, le tout en chinois. L'homme à l'intérieur de la chambre ne connaît pas cette langue, mais il prend le papier et commence à travailler en s'aidant du livre ; il applique laborieusement les instructions. Parfois, l'une d'elles lui enjoint d'écrire les

idéogrammes sur du papier, parfois, elle lui demande d'en effacer ou d'en déplacer. L'homme applique laborieusement une règle après l'autre, dessine et efface des idéogrammes jusqu'à ce qu'une instruction lui signale que c'est fini. Les réponses aux questions lui sont toutefois toujours inconnues. Le livre lui demande de passer le papier par l'ouverture. Il s'exécute, se demandant à quoi pouvait bien avoir servi cet exercice terriblement fastidieux.

De l'autre côté du mur, une Chinoise lit les pages. Les réponses sont toutes correctes et parfois perspicaces. Si on lui demande si les réponses ont été fournies par un esprit intelligent qui a compris l'histoire, elle répond avec assurance que oui. Mais peut-elle avoir raison ? Qui a compris l'histoire ? Ce n'est sûrement pas le personnage qui se trouvait dans la chambre, car il ne connaît pas le chinois et ne sait rien de l'histoire. Ce n'est évidemment pas le livre, qui n'est jamais rien d'autre qu'un livre, un objet inerte posé au milieu des piles de papier. Alors, d'où provient la compréhension des idéogrammes ? Selon Searle, aucune propriété cognitive ne s'est manifestée. Il ne fut question que de feuilleter un tas de pages dépourvues de sens et de griffonner. Voici où nous voulions en venir : la Chambre chinoise est semblable à un ordinateur. Le Français est le microprocesseur, le livre est le logiciel qui alimente le processeur en instructions, et le tas de papiers est la mémoire. C'est pourquoi, quelle que soit l'habileté avec laquelle un ordinateur a été conçu pour simuler l'intelligence en se comportant comme un humain, il n'a pas de cognition et n'est pas intelligent. (Searle avait clairement dit qu'il ne savait pas ce qu'est l'intelligence et que, quelle qu'en soit sa forme, l'ordinateur n'en avait pas.)

Cet argument fit des remous parmi les théoriciens et les experts de l'intelligence artificielle. Il suscita des centaines d'articles véhéments. Les défenseurs de l'intelligence artificielle usèrent de dizaines de contre-arguments, aussi saugrenus que celui-ci : si aucun élément de la chambre ne comprend le chinois, la globalité de la chambre le comprend. Ou encore que la personne dans la

chambre comprenait le chinois, mais ne le savait pas. A mon avis, Searle avait raison. Quand je pense à l'argument de la Chambre chinoise et au fonctionnement des ordinateurs, je n'y décèle nulle part de la cognition. J'étais convaincu que nous devions d'abord comprendre ce qu'est la cognition, découvrir un moyen qui établirait clairement si un système est intelligent ou pas, s'il comprend le chinois ou s'il ne le comprend pas. Ce n'est pas son comportement qui nous l'apprendrait.

Un humain n'a pas besoin de passer aux actes pour comprendre une histoire. Je peux la lire tranquillement, et bien que je ne manifeste ouvertement aucun comportement, ma compréhension de la narration n'en est pas moins claire, du moins pour moi. Par ailleurs, il vous est impossible de dire, en vous fondant sur mon comportement passif, si je comprends ou non l'histoire, ou même si je comprends la langue dans laquelle elle est écrite. Vous pourriez certes m'interroger par la suite pour le vérifier, mais ma cognition se produisait au cours de la lecture, et pas uniquement au moment précis où je réponds à vos questions. Une des thèses de cet ouvrage est que la cognition ne peut être évaluée selon le comportement extérieur. Comme vous le découvrirez dans les chapitres à venir, il s'agit plutôt d'une quantification interne de la manière dont le cerveau se souvient et se sert de ce qu'il a mémorisé pour effectuer des prédictions. La Chambre chinoise, Deep Blue et la plupart des programmes informatiques n'offrent rien de semblable. Ils ne comprennent pas ce qu'ils font. Le seul moyen de juger si un ordinateur est intelligent ou pas, ce sont ses données en sortie, c'est-à-dire son comportement.

L'argument ultime des tenants de l'intelligence artificielle est que, théoriquement, l'ordinateur pourrait simuler la totalité du cerveau. Il pourrait modéliser tous les neurones et toutes ses connexions, à tel point que plus rien ne permettrait de distinguer l'intelligence cérébrale de celle de la simulation informatique. Bien que ce soit impossible à réaliser en pratique, je suis d'accord avec cet argument. Les chercheurs en intelligence artificielle ne

simulent toutefois pas le cerveau et leurs programmes ne sont pas intelligents. Il est impossible de simuler un cerveau sans comprendre d'abord de quelle manière il fonctionne.

---

Après les refus d'Intel et du MIT, je ne sus que faire. Quand vous ne savez plus comment continuer, la meilleure attitude consiste souvent à ne rien entreprendre tant que les diverses options ne se clarifient pas. J'ai donc poursuivi mon travail dans le domaine de l'informatique. J'aimais bien Boston, mais en 1982, mon épouse désira déménager en Californie, ce que nous fîmes (j'avais de nouveau choisi la voie la moins conflictuelle). Je trouvai un emploi dans la Silicon Valley, dans une start-up nommée Grid Systems. La société avait inventé l'ordinateur portable, une belle machine qui devint le premier ordinateur à entrer dans les collections du Museum of Modern Art de New York. J'ai d'abord travaillé au service du marketing puis comme ingénieur. J'ai enfin créé un langage de programmation de haut niveau appelé GridTask et joué un rôle de plus en plus déterminant dans la réussite de la société. Ma carrière était en bonne voie.

Mais je n'arrivais pas à me défaire de mon intérêt pour le cerveau et les machines intelligentes. Mon désir d'étudier le cerveau restait entier. Je pris donc des cours par correspondance sur la physiologie humaine et appris par moi-même. J'ai ensuite sollicité l'admission à un cours de biologie diplômant, ce qui me permit d'étudier l'intelligence d'un point de vue biologique. Si l'informatique ne voulait pas d'un théoricien du cerveau, la biologie serait peut-être heureuse de voir venir un informaticien. La biologie théorique n'existait pas, encore moins la neurobiologie théorique. La biophysique me sembla être le domaine qui répondrait le mieux à mes aspirations. Je travaillai dur, passai les examens d'entrée requis, rédigeai un curriculum vitæ, sollicitai des lettres de recommandation, après quoi je fus accepté comme étudiant en biophysique à plein temps à l'université de Berkeley.

J'étais ravi. Je pouvais enfin travailler sérieusement et à ma guise sur la théorie du cerveau. Je quittai Grid sans intention de retour dans le secteur de l'informatique. Cela signifiait bien sûr la renonciation à un salaire pour une durée indéterminée. Ma femme pensait que le moment était venu d'acheter une maison et de fonder une famille et voilà que je cessais sans remords de subvenir à nos besoins. Ce n'était pas du tout la voie la moins conflictuelle. Mais c'était la meilleure solution et mon épouse encouragea ma décision.

John Ellenby, le fondateur de Grid, m'entraîna dans son bureau et me dit : « Je sais que tu n'as pas du tout l'intention de revenir chez Grid ou dans l'informatique, mais on ne sait jamais ce qui peut arriver. Au lieu de tout arrêter, pourquoi ne pas prendre une mise en disponibilité ? Comme ça, si dans un an ou deux tu désires revenir, tu retrouveras ton salaire, ta place et tes stock-options. » C'était un geste sympathique. J'acceptai, mais je sentais bien que je quittais définitivement l'informatique.

# 2

# LES RÉSEAUX NEURONAUX

**La première** chose que je fis dès mon entrée à l'université de Berkeley en 1986 fut de réunir une documentation sur l'histoire des théories de l'intelligence et des fonctions cérébrales. Je lus des centaines de communications d'anatomistes, de physiologistes, de philosophes, de psychologues, de linguistes et d'informaticiens. Beaucoup de gens issus de nombreux domaines ont écrit abondamment sur la pensée et l'intelligence. Chaque domaine possède ses publications et utilise sa propre terminologie. J'ai trouvé leurs descriptions inégales et incomplètes. Les linguistes parlent de l'intelligence en termes de « syntaxe » ou de « sémantique ». Pour eux, le cerveau et l'intelligence se réduisent au langage. Les spécialistes de la vision parlent de perception en 2D, en 2,5D et en 3D. Pour eux, le cerveau et l'intelligence se réduisent à la reconnaissance de motifs — ou *patterns* — visuels. Les informaticiens parlent de « schémas » et de « cadres », des termes nouveaux qu'ils ont élaborés pour représenter la connaissance. Aucun de ces spécialistes n'a abordé la structure du cerveau ni la manière dont

leurs théories pourraient s'y intégrer. Par ailleurs, les anatomistes et les physiologistes ont abondamment écrit sur la structure de l'encéphale et le comportement des neurones, mais ils se sont gardés de proposer une théorie globale. Il était ardu et frustrant de tenter de comprendre ces diverses approches ainsi que la masse de données qui les accompagnaient.

A cette même époque surgit une nouvelle approche prometteuse des machines intelligentes. Il était question des réseaux neuronaux dès les années 1960, sous une forme ou sous une autre. Ce concept concurrençait l'intelligence artificielle, tant du point de vue de l'attribution des fonds que de la publicité dont bénéficieraient les organismes prêteurs. L'intelligence artificielle, qui s'est taillé la part du lion, a activement contrarié les recherches sur les réseaux neuronaux. Pendant plusieurs années, les spécialistes de ces réseaux ne trouvèrent plus de subventions. Mais certains persévérèrent néanmoins. Au milieu des années 1980, le vent tourna. Il est difficile de savoir ce qui suscita un intérêt soudain pour les réseaux neuronaux, mais l'un des facteurs déterminants fut indubitablement l'impasse dans laquelle s'était fourvoyée l'intelligence artificielle. Il fallait trouver une autre solution ; les réseaux neuronaux artificiels en proposaient une.

Les réseaux neuronaux marquèrent un progrès par rapport à l'intelligence artificielle, car leur architecture est calquée — certes vaguement — sur celle du véritable système nerveux. Au lieu de programmer des ordinateurs, les spécialistes des réseaux neuronaux, appelés aussi *connexionnistes*, s'efforcent de découvrir quels types de comportements peuvent être obtenus en reliant un ensemble de neurones entre eux. Par conséquent, le cerveau est un réseau neuronal. C'est un fait. Les connexionnistes espèrent que l'étude des interactions entre les neurones révélera les propriétés fugaces de l'intelligence, et que des problèmes que l'intelligence artificielle n'avait pu résoudre le seraient par la reproduction fidèle des connexions entre des ensembles de neurones. Un réseau neuronal n'est pas comparable à un ordinateur

car il ne possède pas de processeur et ne stocke pas l'information dans une mémoire centralisée. Dans un réseau neuronal, les connaissances et les mémorisations sont réparties sur l'ensemble des connexions, exactement comme dans l'encéphale.

De prime abord, les réseaux neuronaux semblaient correspondre exactement à ce qui m'intéressait. Mais cette impression ne dura pas. A cette époque, je m'étais forgé une opinion selon laquelle trois éléments étaient essentiels pour comprendre le cerveau. Le premier était la prise en compte du temps dans les fonctions cérébrales. Le cerveau traite rapidement des flux de données changeants. Rien n'est statique dans les flux entrants et sortants.

Le deuxième critère était l'importance de la rétropropagation, ou *feedback*. Les neuro-anatomistes savent de longue date que le cerveau est plein de connexions qui renvoient une information. Par exemple, dans le circuit qui relie le néocortex à une structure inférieure, le thalamus, le nombre de rétroconnexions (vers les entrées, ou inputs) excède les antéroconnexions (vers l'avant) d'un facteur de dix. Autrement dit, pour chaque fibre qui apporte des informations au néocortex, dix fibres retournent des informations vers les sens. Le feedback régit aussi la plupart des connexions à l'intérieur du néocortex; personne ne connaît son rôle exact, mais selon des publications scientifiques, il est clair qu'il se manifeste partout. Ceci me parut important.

Le troisième critère était que toute théorie ou modèle du cerveau doit prendre en compte l'architecture physique de l'encéphale. Le néocortex n'est pas qu'une simple structure. Comme nous le découvrirons ultérieurement, il est organisé sous la forme d'une hiérarchie répétitive. Tout réseau neuronal qui n'adopterait pas cette structure ne saurait fonctionner à l'instar d'un cerveau.

Lorsque les réseaux neuronaux occupèrent le devant de la scène, c'était le plus souvent sous la forme de modèles simplistes qui ne répondaient à aucun des critères qui viennent d'être énoncés. La plupart ne représentaient qu'un petit nombre de neurones

connectés en trois rangées. Un *pattern* (l'entrée) se trouve sur la première rangée. Ces neurones d'entrée sont connectés à une deuxième rangée de neurones appelés « unités cachées ». Ces unités cachées sont elles-mêmes connectées à la dernière rangée de neurones, les neurones de sortie. L'intensité des connexions entre les neurones varie ; cela signifie que l'activité dans un neurone peut accroître l'activité dans un autre et la réduire dans un troisième selon l'intensité des connexions. En modifiant ces intensités, le réseau apprend à associer les patterns en entrée aux patterns en sortie.

Ces réseaux neuronaux simples ne traitaient que des patterns statiques, étaient dépourvus de feedback et ne ressemblaient en rien à un cerveau. Le type de réseau neuronal le plus courant, « à rétropropagation », apprenait en répercutant une erreur depuis les unités de sortie vers les unités d'entrée. Ceci ressemble à une forme de feedback mais n'en est pas un. La rétropropagation des erreurs ne se produit qu'en phase d'apprentissage. Lorsque le réseau neuronal fonctionne normalement, après avoir été entraîné, l'information ne circule que dans un sens. Il n'existe pas de feedback se propageant des entrées vers les sorties. De plus, les modèles n'avaient pas la notion du temps. Un pattern d'entrée statique était converti en pattern de sortie statique, puis un autre pattern d'entrée était présenté. Aucun historique du réseau n'était enregistré, capable de restituer ce qui s'était déroulé, ne serait-ce que juste auparavant. Enfin, l'architecture de ces réseaux neuronaux était rudimentaire comparée aux structures complexes et hiérarchisées du cerveau.

Je pensais qu'une évolution rapide vers des réseaux plus réalistes verrait le jour, mais il n'en fut rien. Ce qu'effectuaient ces réseaux neuronaux simples étant digne d'intérêt, les chercheurs s'en contentèrent pendant des années. Ils découvrirent de nouveaux outils, et du jour au lendemain, des milliers de scientifiques, d'ingénieurs et d'étudiants obtinrent des subventions et des bourses et rédigèrent nombre de thèses et de livres sur les réseaux neu-

ronaux. Des sociétés furent créées, qui utilisaient des réseaux neuronaux pour prévoir les fluctuations boursières, calculer l'évolution des emprunts, vérifier des signatures et exécuter des centaines d'autres applications fondées sur la classification de patterns. Bien que les intentions de ceux qui avaient fondé ce secteur d'activité soient plus généralistes, le domaine des réseaux neuronaux fut néanmoins dominé par des spécialistes désireux de comprendre le fonctionnement du cerveau et la nature de l'intelligence.

Les médias ne firent pas bien la distinction entre intelligence artificielle et réseaux neuronaux. Pour les journaux, les magazines et les documentaires télévisés, les réseaux neuronaux ressemblaient au cerveau et fonctionnaient selon le même principe. Contrairement à l'intelligence artificielle, où tout devait être programmé, les réseaux neuronaux apprenaient par l'exemple, ce qui semblait apparemment relever de l'intelligence. NetTalk en fut une fameuse démonstration. Ce réseau neuronal apprenait à associer des successions de caractères typographiques à la prononciation de sons. Lorsqu'il fut entraîné à imprimer du texte, il sembla que l'ordinateur lisait les mots à haute voix. Il était tentant d'en déduire que, le temps passant, les réseaux neuronaux parviendraient à converser avec les humains. Un journal national qualifia malencontreusement NetTalk de « machine apprenant à lire ». NetTalk était certes une belle démonstration, mais qui n'allait pas très loin. Il ne lisait pas, il ne comprenait rien. D'un point de vue pratique, sa valeur était insignifiante. Il ne faisait qu'apparier des combinaisons de caractères à des patterns sonores prédéfinis.

Permettez-moi de vous livrer une analogie qui illustre combien un réseau neuronal est loin d'être un cerveau. Supposons qu'au lieu de nous efforcer de comprendre le fonctionnement du cerveau, nous tentions de comprendre le fonctionnement d'un calculateur numérique. Après des années d'études, nous découvririons que tout, dans l'ordinateur, est fait de transistors. Il y en a des centaines de millions, tous interconnectés d'une manière

précise et complexe. Nous ne comprendrions toutefois pas comment l'ordinateur fonctionne, ni pourquoi les transistors sont câblés de telle ou telle manière. Un jour, nous déciderions de ne connecter que quelques transistors entre eux pour voir ce qui se passe. De fil en aiguille, nous découvririons qu'en connectant seulement trois transistors d'une certaine manière, nous obtenons un amplificateur : un signal faible à l'entrée est plus fort à la sortie (c'est ce principe qui est mis en œuvre dans les postes de radio et les téléviseurs). Ce serait là une découverte importante et en un rien de temps, tout un secteur économique se mettrait à fabriquer des postes à transistors, des téléviseurs et autres équipements électroniques utilisant des amplificateurs à transistors. Tout cela est fort bien, mais ne nous apprendrait rien de plus sur le fonctionnement de l'ordinateur. Bien qu'un amplificateur et un ordinateur soient tous deux constitués de transistors, ils n'ont presque rien d'autre en commun. Dans le même esprit, un encéphale et un réseau neuronal à trois rangées sont tous deux constitués de neurones, mais n'ont rien en commun.

A l'été 1987, je fis une expérience qui jeta un plus grand froid encore sur mon enthousiasme déjà tiède pour les réseaux neuronaux. J'assistais à une conférence sur ce sujet, au cours de laquelle la société Nestor fit une présentation. Elle tentait de vendre une application reposant sur un réseau neuronal capable de reconnaître un texte manuscrit écrit sur une tablette tactile. La licence de ce programme était proposée au prix d'un million de dollars, ce qui éveilla mon attention. Bien que Nestor ait mis en avant la sophistication de l'algorithme du réseau neuronal et vantât son innovation, j'eus l'intuition que la reconnaissance de l'écriture manuscrite pouvait s'effectuer d'une manière plus simple, plus traditionnelle. De retour chez moi, je repensai au problème et, en deux jours, je conçus un système de reconnaissance rapide, compact et souple. Ma solution ne reposait pas sur un réseau neuronal et ne fonctionnait pas du tout comme un cerveau. Bien que cette conférence ait déclenché un intérêt pour la conception

d'ordinateurs dont l'interface serait équipée d'un stylet, ce qui mena finalement au PalmPilot dix ans plus tard, elle me convainquit surtout que les réseaux neuronaux n'étaient pas un progrès comparés aux méthodes traditionnelles. Le système de reconnaissance de l'écriture manuscrite que j'avais inventé fut à la base d'un système de saisie de texte nommé Graffiti, utilisé dans la première série des assistants personnels fabriqués par Palm. Il me semble que Nestor a fini par fermer boutique.

Tant d'efforts pour de simples réseaux neuronaux. La plupart de leurs capacités pouvant facilement être prises en charge par d'autres méthodes, le battage médiatique s'estompa. Les spécialistes des réseaux neuronaux s'abstinrent de clamer que leurs modèles étaient intelligents. Après tout, ce n'était que des réseaux extrêmement simples qui en faisaient moins que des programmes d'intelligence artificielle. Loin de moi l'intention de vous laisser croire que la totalité des réseaux neuronaux est une variété simpliste des réseaux à trois couches. Des chercheurs ont continué à travailler et ont conçu des modèles plus sophistiqués. Aujourd'hui, l'expression *réseau neuronal* décrit d'autres modèles dont certains sont biologiquement plus justes et d'autres non. Mais quasiment aucun ne vise à restituer la fonction globale du néocortex ou son architecture.

A mon avis, le problème fondamental de la plupart des réseaux neuronaux réside dans une caractéristique qu'ils partagent avec les programmes d'intelligence artificielle. Tous deux sont inévitablement entravés par leur fixation sur le comportement. Que ce dernier soit appelé « réponse », « pattern » ou « sortie (output) », l'intelligence artificielle et les réseaux neuronaux présument que l'intelligence réside dans le comportement que l'un et l'autre produisent consécutivement à une entrée (input). La propriété la plus importante d'un programme informatique ou d'un réseau neuronal est la validité de la sortie (output). Comme l'avait suggéré Alan Turing, l'intelligence, c'est le comportement.

Mais l'intelligence ne se limite pas à des actions ou des comportements intelligents. Le comportement est une manifestation de l'intelligence, mais ce n'est ni la caractéristique essentielle ni la définition principale du fait d'être intelligent : vous pouvez faire preuve d'intelligence en étant couché dans le noir et en pensant et comprenant. Ignorer ce qui se passe dans la tête et se concentrer sur le comportement a considérablement entravé la compréhension de l'intelligence et la conception de machines intelligentes.

---

Avant d'aborder une nouvelle définition de l'intelligence, je tiens à présenter une autre théorie connexionniste qui se rapproche davantage de la manière dont le cerveau fonctionne. Hélas, peu de chercheurs ont pris conscience de l'importance de ces travaux.

Alors que les réseaux neuronaux étaient sous les feux de la rampe, un petit groupe de théoriciens a élaboré des réseaux qui ne reposent pas sur le comportement. Appelés « mémoires auto-associatives », ils sont eux aussi constitués de simples « neurones » reliés entre eux, qui sont excités lorsqu'un certain seuil est atteint. Mais ils sont interconnectés différemment, usant abondamment de feedbacks. Au lieu de ne transmettre l'information qu'en avant, comme dans un réseau à rétropropagation, les mémoires auto-associatives renvoient l'output de chaque neurone vers son input, un peu comme si vous vous appeliez vous-même au téléphone. Ce feedback en boucle présente d'intéressantes caractéristiques. Lorsqu'un pattern de stimulations est soumis à des neurones artificiels, ces derniers forment la mémoire de ce pattern. Le réseau auto-associatif associe les patterns avec lui-même, d'où les termes de *mémoire auto-associative.*

Le résultat de ce câblage peut, de prime abord, paraître saugrenu. Car pour récupérer un pattern stocké dans une telle mémoire, vous devez fournir le pattern à récupérer. C'est un peu

comme si vous entriez dans une épicerie pour acheter des bananes et, quand l'épicier demande comment vous payez, que vous lui répondiez : « Avec des bananes. » A quoi bon, vous demanderez-vous ? La mémoire auto-associative possède cependant quelques importantes propriétés propres à l'encéphale.

La plus importante de ces propriétés est qu'il n'est pas nécessaire de disposer de l'intégralité du pattern à récupérer. Il suffit d'en posséder un fragment, ou seulement une partie en désordre. La mémoire auto-associative peut en effet récupérer le pattern correct, tel qu'il avait originellement été stocké, même à partir d'une version altérée. C'est un peu comme aller chez l'épicier avec une banane trop mûre, brunie et à moitié grignotée, et obtenir en retour une banane entière. Ou alors, se présenter à la banque avec un billet chiffonné, déchiqueté, à peine lisible, et s'entendre dire par le guichetier : « Je pense que c'est un billet de 100 euros complètement abîmé. Donnez-le moi et je vous en rendrai un autre tout neuf. »

Deuxièmement, contrairement à la plupart des réseaux neuronaux, une mémoire auto-associative peut être conçue pour stocker des séquences de patterns, ou patterns temporels. Cette fonctionnalité est obtenue en ajoutant un retard au feedback. Ce délai permet de présenter à la mémoire auto-associative une succession de patterns, semblables à une mélodie, dont elle se souviendra. Il suffira de proposer les premières notes de la chanson « Quand trois poules vont au champ » pour obtenir, en retour, tout le morceau. Lorsqu'une partie de la séquence lui est présentée, la mémoire se souvient du reste. Comme nous le verrons ultérieurement, c'est ainsi que nous apprenons toutes choses, sous la forme de successions de patterns. J'avance l'idée que dans ce but le cerveau utilise des circuits analogues à ceux d'une mémoire auto-associative.

Les mémoires auto-associatives ont attiré l'attention sur l'importance potentielle des feedbacks et des inputs variant dans le temps. Mais la grande majorité des programmes d'intelligence

artificielle, des réseaux neuronaux et des chercheurs cognitivistes ignorait le temps et le feedback.

Dans leur ensemble, les neurobiologistes n'ont pas mieux fait. Ils savent à présent ce qu'est le feedback — ce sont eux qui l'ont découvert — mais la plupart n'ont aucune théorie à proposer, hormis quelques vagues discussions sur les « phrases » et la « modulation » leur permettant de savoir pourquoi le cerveau en a tant besoin. Quant au temps, il ne tient que peu de place ou ne joue pas un rôle prépondérant dans la plupart de leurs exposés sur le fonctionnement global du cerveau. Ils ont tendance à cartographier l'encéphale en zones où se manifestent des phénomènes, et non selon le moment et la manière dont les patterns neuronaux excités interagissent dans la durée. Un peu de ce parti pris découle des limites de nos actuelles techniques expérimentales. L'une des technologies favorites des années 1990, surnommées « la décade du cerveau », était l'imagerie fonctionnelle, capable de photographier l'activité cérébrale. Elle ne permet toutefois pas de visualiser des changements rapides. Les scientifiques demandent au sujet examiné de se concentrer intensément sur une seule tâche, puis ils lui demandent de rester immobile, le temps de prendre un cliché, c'est-à-dire une photo du cerveau. C'est ainsi que nous disposons de quantités de données révélant *où*, dans le cerveau, se manifeste telle ou telle tâche, mais de bien peu de données révélant la variation dans la durée du flux d'inputs qui parcourt le cerveau. L'imagerie fonctionnelle produit un aperçu de ce qui se passe *quelque part* à un moment donné, mais est incapable de restituer facilement l'évolution de l'activité cérébrale. Les chercheurs aimeraient bien collecter de telles données, mais les techniques éprouvées permettant de le faire sont peu nombreuses. C'est pourquoi beaucoup de neurobiologistes cognitifs traditionnels se fourvoient dans l'erreur des inputs-outputs : une entrée fixe est introduite et vous voyez ce qu'il en sort. Les schémas des connexions corticales se présentent volontiers sous la forme d'organigrammes qui débutent aux aires sensorielles principales, sièges de la vue, de l'ouïe et du

toucher, pour se poursuivre par des aires analytiques, planificatrices et motrices, puis les muscles. Vous ressentez, donc vous agissez.

Je ne veux pas insinuer que tout le monde a ignoré le temps et les feedbacks. C'est un domaine si vaste que presque toutes les idées ont leurs partisans. Ces dernières années, la croyance en l'importance des feedbacks, du temps et de la prédiction avait le vent en poupe. Mais la prédominance de l'intelligence artificielle et des réseaux neuronaux classiques avait fait de l'ombre aux autres approches, d'où une sous-estimation qui perdura des années.

---

Il n'est pas difficile de comprendre pourquoi les gens — les néophytes autant que les experts — crurent que le comportement définit l'intelligence. Pendant au moins deux siècles, le cerveau avait servi à mettre au point des mécanismes d'horlogerie, des pompes et des tuyaux, des machines à vapeur et, plus tard, des ordinateurs. Des décennies de science-fiction ont répandu à foison les idées d'intelligence artificielle, notamment les lois qui régissent les robots des romans d'Isaac Asimov ou les capacités du verbeux robot C3PO de *La Guerre des étoiles*. Le concept de machine intelligence exécutant des tâches est profondément ancré dans notre imagination. Toutes les machines, qu'elles aient été réalisées par l'homme ou imaginées, sont censées faire quelque chose. Nous ne disposons pas de machines qui pensent, seulement de machines qui font. Même quand nous observons nos congénères, nous nous focalisons sur leur comportement, et non sur leurs pensées intimes. C'est pourquoi il paraît intuitivement évident qu'un comportement intelligent devrait être à l'aune de tout système intelligent.

Toutefois, l'histoire des sciences nous apprend que notre intuition a souvent été le plus gros obstacle à notre recherche de la vérité. Les structures scientifiques sont souvent difficiles à découvrir, non en raison de leur complexité, mais parce que des hypothèses intuitives mais erronées nous empêchent de discerner

la réponse juste. Les astronomes qui précédèrent Copernic (1473-1543) présumèrent à tort que la Terre est le centre de l'univers parce qu'elle *semble* immobile et *paraît* occuper le milieu de la voûte céleste. Il était intuitivement évident que les étoiles étaient tapissées sur une gigantesque sphère tournante. Soutenir que la Terre tournoie sur elle-même comme une toupie à la vitesse de 1 674,38 km/h à l'équateur, qu'elle est propulsée à grande vitesse à travers l'espace sidéral — sans même parler des étoiles situées à des milliers de milliards de kilomètres — vous aurait fait passer pour un fou. Mais il s'avéra que l'univers est ainsi fait. Facile à comprendre, mais intuitivement erroné…

Avant Darwin (1809-1882), il semblait évident que les espèces étaient immuables. Les crocodiles ne sont pas apparentés aux colibris ; tous deux sont irrémédiablement différents. L'idée de l'évolution des espèces devait s'imposer non seulement face aux enseignements religieux, mais aussi face au sens commun. L'évolution implique l'existence d'un ancêtre commun à tous les êtres de la planète, y compris les vers de terre et la fleur en pot dans la cuisine. Nous savons à présent que c'est vrai, mais notre intuition suggérait autre chose.

Je mentionne ces exemples célèbres car je pense que la quête de machines intelligentes a été entravée par une hypothèse intuitive qui a empêché les progrès. Lorsque vous vous interrogez sur ce que fait un système intelligent, il va intuitivement de soi d'y penser en termes de comportement. L'intelligence humaine se manifeste par le discours, l'écriture et les actions, n'est-ce pas ? Certes, mais seulement jusqu'à un certain point. L'intelligence se produit dans notre tête. Le comportement est un ingrédient facultatif. Ce n'est intuitivement pas évident, mais n'est pas pour autant difficile à comprendre.

---

Au printemps 1986, alors que je me prélassais à mon bureau après avoir lu des articles scientifiques, élaborant une histoire de

l'intelligence, observant les évolutions de l'intelligence artificielle et des réseaux neuronaux, je me rendis compte que je me perdais dans les détails. Il y avait là une intarissable quantité d'articles à lire et à étudier et je ne parvenais pas du tout à comprendre clairement comment le cerveau fonctionne réellement, ni même ce qu'il fait. C'est parce que le domaine des neurosciences est lui-même inondé de détails. Des milliers de communications scientifiques sont publiées chaque année, mais elles ne font que grossir le tas au lieu d'y apporter de l'ordre. Il n'existe toujours pas de théorie globale ni de structure expliquant ce que fait le cerveau et comment il le fait.

J'ai donc commencé à imaginer ce que pourrait être la solution au problème. Serait-elle extrêmement compliquée parce que le cerveau est complexe ? Remplirait-elle une centaine de pages de denses formulations mathématiques ? Devrais-je élaborer des centaines ou des milliers de circuits distincts avant que l'un d'eux puisse se révéler exploitable ? Je ne le pensais pas. L'histoire montre que les meilleures solutions aux problèmes scientifiques sont simples et élégantes. Bien que les détails puissent être rébarbatifs et la voie vers la théorie finale ardue, l'ultime structure conceptuelle est généralement simple.

Sans une explication fondamentale pour guider leurs recherches, les neurobiologistes n'iront pas loin dans leur tentative d'assembler les détails pour élaborer une représentation cohérente. Le cerveau est d'une incroyable complexité. C'est un vaste et décourageant fouillis de cellules. A première vue, il ressemblerait à un terrain recouvert de spaghettis cuits. On l'a aussi comparé à un cauchemar pour électricien. Mais un attentif examen révèle que le cerveau n'est pas une masse informe. Il fourmille d'organisations et de structures bien trop nombreuses pour que nous puissions l'appréhender dans son ensemble, comme le ferait un archéologue lorsqu'il reconstitue intuitivement un vase brisé à partir de tessons. Ce n'est pas faute de manquer de données ou de ne pas avoir les bonnes ; ce qui manque est une mise en perspective. A condition de

disposer de la structure adéquate, les détails prendront du sens et pourront être exploités. L'analogie qui suit vous en apprendra plus.

Supposons que dans des millénaires le genre humain ait disparu. Des explorateurs venus d'une civilisation extraterrestre avancée arrivent sur Terre. Ils tentent de savoir comment nous vivions. Ils sont particulièrement intrigués par notre réseau routier. A quoi pouvait bien servir cette bizarre structure très élaborée ? Ils commencent par tout cataloguer, à la fois depuis le ciel et au sol. Ce sont de méticuleux archéologues. Ils relèvent l'emplacement de tous les fragments d'asphalte épars, de tous les panneaux de signalisation tombés à terre et entraînés par l'érosion. Ils notent tous les détails qu'ils trouvent. Des réseaux routiers sont différents d'autres. A certains endroits, ils sont sinueux et étroits, presque tracés au hasard. A d'autres, ils sont larges et réguliers. Certains traversent le désert en ligne droite. Les visiteurs collectent une montagne de détails, mais ces détails sont pour eux dépourvus de signification. Ils n'en continuent pas moins à en récolter davantage dans l'espoir de trouver l'information qui expliquerait tout. Ils sont longtemps dans l'expectative.

Jusqu'à ce que l'un d'eux s'exclame dans sa langue : « Eurêka ! Il me semble que… ces créatures étaient incapables de se téléporter comme nous le faisons. Elles devaient se déplacer de lieu en lieu, peut-être sur des plates-formes mobiles astucieusement conçues. » A partir de cette perspicace illumination, beaucoup de détails deviennent limpides. Les petits réseaux routiers sinueux datent d'une époque ancienne, lorsque les moyens de transport étaient lents. Les routes larges servaient à parcourir de grandes distances à des vitesses élevées, ce qui explique les chiffres figurant sur les panneaux de signalisation. Par déduction, les savants extraterrestres distinguent les zones résidentielles des zones industrielles, la manière dont les infrastructures de transport et les nécessités du commerce ont interagi, et ainsi de suite. Bon nombre des détails qu'ils avaient engrangés se révèlent peu pertinents, liés uniquement à des accidents de l'histoire ou à des

exigences de la géographie locale. La quantité de données brutes n'a pas changé, mais elle n'est plus obscure.

Nous pouvons être certains que le même type de révélation nous permettra de comprendre ce qu'il en est de tous les détails épars que nous possédons sur le cerveau.

---

Malheureusement, tout le monde ne croit pas que nous parviendrons un jour à comprendre comment le cerveau fonctionne. Pour un nombre étonnamment élevé de gens, dont quelques neurobiologistes, le cerveau et l'intelligence dépassent l'entendement. Certains sont même persuadés que, même si nous arrivions à comprendre, il serait impossible de créer des machines fonctionnant de la même manière que le cerveau, que l'intelligence exige un corps humain, des neurones, le tout soumis à d'impénétrables lois physiques. Chaque fois que j'entends de tels arguments, je repense à ceux qui, dans le passé, s'opposaient à l'étude du ciel et à la dissection des corps. « A quoi bon étudier tout cela, arguaient-ils, il n'en sortira rien de bon, et même si vous finissiez par comprendre, ce savoir ne nous servirait à rien. » Ce sont de tels arguments qui ont mené à une branche de la philosophie nommée « fonctionnalisme », la dernière étape dans cette brève histoire de ce que nous savons de la pensée.

Selon le fonctionnalisme, être intelligent ou avoir un esprit est une propriété purement organisationnelle qui n'a rien à voir, de par sa nature, avec la manière dont vous êtes matériellement organisé. L'esprit existe dans tout système dont les parties constituantes entretiennent des relations causales adéquates les unes avec les autres. Ces parties peuvent aussi bien être des neurones, des composants électroniques ou tout autre élément. En clair, cette vision est le point de départ standard pour quiconque voudrait construire des machines intelligentes.

Un jeu d'échecs serait-il moins un jeu d'échecs si une salière remplaçait une pièce manquante ? Certainement pas. La salière est

fonctionnellement équivalente au « véritable » cavalier, en raison de la manière dont elle se déplace sur l'échiquier et interagit avec les autres pièces. C'est pourquoi le jeu est encore un véritable jeu d'échecs et non une simulation. Ou alors, cette phrase n'en serait-elle pas moins la même si je supprimais chacun de ses caractères et si je les retapais ? Pour prendre un exemple plus personnel, considérez le fait que tous les tant et tant d'années, la plupart des atomes de votre corps ont été remplacés. En dépit de cela, vous êtes toujours vous-même, dans tous les sens du terme. Un atome en vaut bien un autre du moment qu'il joue le même rôle fonctionnel dans votre composition moléculaire. Il en va de même pour le cerveau : si un savant fou remplaçait chacun de vos neurones par une nano-machine fonctionnellement équivalente, vous ne devriez pas vous sentir plus différent qu'avant cette intervention.

Selon ce principe, un système artificiel qui adopterait la même architecture fonctionnelle qu'un cerveau vivant intelligent devrait lui aussi être intelligent. Et pas seulement théoriquement, mais vraiment, véritablement intelligent.

Les partisans de l'intelligence artificielle, les connexionnistes et moi sommes tous des fonctionnalistes dans la mesure où nous estimons qu'il n'y a rien de naturellement particulier dans le cerveau qui lui permette d'être intelligent. Nous pensons tous qu'il sera possible de créer un jour des machines intelligentes, d'une manière ou d'une autre. Mais il existe différentes interprétations du fonctionnalisme. J'ai déjà exposé ce que je considère comme l'échec principal de l'intelligence artificielle et des paradigmes connexionnistes — le faux raisonnement des inputs-outputs —, et il reste encore beaucoup à dire quant à notre incapacité, jusqu'à présent, à avoir pu créer des machines intelligentes. Tandis que les partisans de l'intelligence artificielle s'en tiennent à ce que je considère comme une impasse, les connexionnistes, à mon avis, ont surtout été trop timides.

Les chercheurs en intelligence artificielle rétorquent : « Pourquoi est-ce que nous, ingénieurs, devrions nous lier à l'évolution

de solutions qui s'avèrent incertaines ? » En principe, ils marquent un point. Les systèmes biologiques comme le cerveau et le génome sont considérés comme notoirement inélégants. La métaphore couramment utilisée aux Etats-Unis est celle de la machine de Rube Goldberg, du nom d'un dessinateur des années 1930 qui avait conçu d'invraisemblables mécaniques comiques hypercompliquées, chargées d'accomplir des tâches insignifiantes. Les programmeurs de logiciels ont un terme pour cela, *kluge* (de l'allemand *klug*, « intelligent, astucieux »), qui désigne un programme écrit sans préparation, d'une lourdeur prussienne, inutilement alambiqué, à un point tel qu'il finit par être incompréhensible même pour celui qui l'a écrit. Les chercheurs en intelligence artificielle craignent que le cerveau soit un fouillis analogue, un *kluge* datant de millions d'années, bourré de scories inefficaces héritées des aléas de l'évolution. Si tel est le cas, se disent-ils, pourquoi ne pas se débarrasser de tout ce désolant fatras et tout reprendre à zéro ?

Bon nombre de philosophes et de psychologues cogniticiens sont bien disposés à cet égard. Ils aiment bien la métaphore présentant l'esprit comme un logiciel exécuté par le cerveau, qui serait l'équivalent biologique de la partie matérielle de l'ordinateur. Dans un ordinateur, le matériel et le logiciel sont nettement séparés. Un même programme informatique peut être développé pour tourner sur n'importe quelle Machine de Turing Universelle. Le traitement de texte WordPerfect peut être utilisé sur un PC, sur un Macintosh ou sur un supercalculateur Cray par exemple, même si la configuration matérielle de chacune de ces machines est très différente. Quant à la partie matérielle, elle ne joue aucun rôle pendant que vous apprenez à utiliser WordPerfect. Par analogie, quand vous pensez, le cerveau n'a rien à vous enseigner au sujet de l'esprit.

Les défenseurs de l'intelligence artificielle aiment aussi attirer l'attention sur des exemples historiques où les solutions technologiques diffèrent radicalement de ce que la nature a trouvé. Par

exemple, comment avons-nous réussi à fabriquer des machines volantes ? En imitant le battement des ailes des oiseaux et des chauves-souris ? Non. Nous y sommes arrivés à l'aide d'ailes fixes et d'hélices, puis de réacteurs. Ce n'est pas la solution que la nature a choisie, mais elle fonctionne, et mieux encore qu'en battant des ailes.

De même, nous avons mis au point des véhicules tout-terrain capables de battre un guépard à la course, non pas en créant une automobile à quatre pattes, comme le félin, mais en inventant la roue. La roue est parfaite pour se déplacer rapidement sur du terrain plat. Le fait que la nature n'ait pas exploité cette propriété ne signifie pas que nous aurions dû rejeter notre solution. Des philosophes se sont entichés de la métaphore de la « roue cognitive » arguant qu'en intelligence artificielle, la solution à certains problèmes, bien que complètement différente de ce que ferait le cerveau, n'en serait pas moins heureuse. En d'autres termes, un programme produisant des outputs identiques aux performances humaines, ou les surpassant, dans un domaine limité mais utile, vaudrait bien la manière dont s'y serait pris le cerveau.

J'estime que cette interprétation du fonctionnalisme fondée sur la fin justifiant les moyens induit les spécialistes de l'intelligence artificielle en erreur. Ainsi que l'avait démontré Searle avec sa Chambre chinoise, une équivalence comportementale est insuffisante. L'intelligence étant une propriété interne du cerveau, nous devons examiner l'intérieur du cerveau pour comprendre la nature de l'intelligence. Dans notre étude de l'encéphale, notamment du néocortex, nous devrons être très circonspects pour arriver à comprendre quels détails sont superflus, sont des « accidents figés » de notre évolution passée. Sans aucun doute, de nombreux processus de style « Rube Goldberg » sont mêlés à des fonctionnalités importantes. Mais, comme nous le verrons d'ici peu, il existe dans le cerveau une forme d'élégance sous-jacente de grande puissance qui surpasse nos meilleurs ordinateurs et n'attend que d'être extraite de ces circuits neuronaux.

Les connexionnistes avaient intuitivement perçu que le cerveau n'est pas un ordinateur et que son secret réside dans la manière dont les neurones se comportent lorsqu'ils sont interconnectés. C'était un bon point de départ, mais qui a eu du mal à progresser depuis ses premiers succès. Bien que des milliers de spécialistes aient travaillé sur les réseaux neuronaux à trois couches — et y travaillent encore —, les recherches sur des réseaux « corticalement réalistes » étaient et restent rares.

Depuis un demi-siècle, nous avons mis en œuvre la considérable ingéniosité qui caractérise notre espèce pour tenter d'insuffler de l'intelligence aux ordinateurs. Nous avons réussi à programmer des traitements de texte, des bases de données, des jeux vidéo et l'Internet, à fabriquer des téléphones mobiles et modéliser des dinosaures convaincants entièrement en images de synthèse. Mais les machines intelligentes ne sont toujours pas là. Pour cela, nous devrons copier servilement le moteur d'intelligence développé par la nature : le néocortex. Nous devrons extraire l'intelligence des profondeurs du cerveau. Il n'existe pas d'autre voie.

# 3

# LE CERVEAU HUMAIN

**Qu'est-ce** qui fait que le cerveau est si différent de la programmation de l'intelligence artificielle et des réseaux neuronaux ? Qu'est-ce qui est si particulier au cerveau, et en quoi est-ce important ? Comme nous le découvrirons dans les prochains chapitres, l'architecture du cerveau nous en dit long sur la manière dont il fonctionne et en quoi il diffère fondamentalement d'un ordinateur.

Commençons par présenter l'organe dans son entier. Imaginez un cerveau posé sur une table, prêt à être disséqué. Nous remarquerons d'abord que l'aspect extérieur de l'encéphale paraît uniforme. D'un gris rosé, il ressemble à un chou-fleur parcouru de nombreuses sinuosités en relief et en creux : les circonvolutions et les sillons. Il est mou et spongieux au toucher. La partie superficielle est le néocortex, une fine enveloppe de tissus neuronaux qui couvre la plupart des parties archaïques de l'encéphale, le cerveau archaïque. C'est lui qui nous intéresse surtout, car tout ce qui est lié à l'intelligence — la perception, le langage, l'imagination, les mathématiques,

les arts, la musique et la planification — s'y produit. C'est votre néocortex qui lit ce livre.

Je dois reconnaître que je suis un inconditionnel du néocortex. Cette attitude pouvant susciter quelques oppositions, permettez-moi de défendre brièvement mon point de vue. Chaque partie du cerveau est étudiée spécifiquement par des équipes de spécialistes, et l'assertion selon laquelle il est possible de comprendre l'intelligence uniquement par l'étude du néocortex risque de les froisser. Ils objecteront que, sans connaître les propriétés de telle ou telle aire corticale, les chances de comprendre le néocortex sont minces, car tout est étroitement interconnecté, et que les aires en question sont indispensables pour telle ou telle activité. Je ne suis pas opposé à ce point de vue. Certes, le cerveau est constitué de plusieurs parties dont la plupart jouent un rôle prédominant. Curieusement, la seule exception est la partie du cerveau qui compte le plus de cellules, le cervelet. Si un individu est né sans cervelet, ou si ce dernier a été endommagé, il pourra cependant mener une vie presque normale. Ce n'est toutefois pas le cas des autres zones du cerveau. La plupart sont indispensables aux fonctions élémentaires de la vie, notamment la sensorialité.

Mon contre-argument est que je ne suis pas intéressé par la création d'êtres humains. Je désire comprendre l'intelligence et créer des machines intelligentes. L'être humain et l'être intelligent sont deux entités distinctes. Une machine intelligente n'a pas besoin d'envies sexuelles, d'éprouver la faim, des pulsions, des émotions ou de ressentir des contractions musculaires, comme c'est le cas du corps humain. L'être humain est plus qu'une machine intelligente. Nous sommes des créatures biologiques dotées de tout le bagage nécessaire, et parfois superflu, qui nous a été légué par une longue évolution. Pour construire des machines intelligentes qui se comporteraient comme des humains — c'est-à-dire qui réussiraient le test de Turing en toutes circonstances — il faudrait sans doute aussi recréer tous les autres aspects qui font

de nous des êtres humains. Comme nous le verrons plus tard, pour fabriquer des machines indubitablement intelligentes mais pas exactement à la manière des humains, nous nous concentrerons sur la partie de l'encéphale dont dépend strictement l'intelligence.

A ceux que ma singulière fixation sur le néocortex choquerait, je répondrai que je ne conteste pas du tout que les autres structures du cerveau comme le tronc cérébral, les ganglions basaux et les corps amygdaloïdes sont elles aussi importantes pour le fonctionnement du néocortex humain. Mais j'espère vous convaincre du fait que tous les aspects essentiels de l'intelligence se produisent au niveau du néocortex, auquel s'ajoute le rôle crucial joué par deux autres parties du cerveau, le thalamus et l'hippocampe que nous étudierons plus loin dans ce livre. Nous devrons à terme comprendre le rôle fonctionnel de toutes les aires cérébrales. Je pense que ces sujets seront mieux abordés dans le contexte d'une théorie globale de la fonction du néocortex, ou « cortex ».

Prenez six cartes de visite ou six cartes à jouer — peu importe — et empilez-les : vous venez de créer un modèle du cortex. Les six cartes ont environ deux millimètres d'épaisseur, ce qui correspond à celle de l'enveloppe corticale. A l'instar des cartes, le néocortex est en effet épais de deux millimètres et il est constitué de six couches d'égale épaisseur.

Mise à plat, la surface de l'enveloppe néocorticale de l'être humain est celle d'une grande nappe de table. Celle des autres mammifères est plus petite. L'enveloppe corticale du rat est de la taille d'un timbre-poste et celle du singe de la taille d'une enveloppe à lettre. Mais, quelles qu'en soient les dimensions, elle comporte toujours six couches, comme les cartes empilées. L'être humain est plus intelligent parce que, par rapport à son corps, son cortex est plus vaste, et non parce qu'il serait plus épais ou contiendrait des cellules particulières. Sa surface est impressionnante car il enveloppe au plus près le cerveau, jusque dans les sillons des circonvolutions. Pour s'accommoder d'un cerveau

d'aussi grande taille, la nature a été obligée de modifier notre anatomie. Le bassin de la femme s'est élargi afin de ménager de la place à l'enfant à gros cerveau qu'elle porte, une caractéristique qui, selon les anthropologues, aurait évolué en même temps que la capacité à marcher sur deux pieds. Comme cette hypertrophie n'était pas suffisante, l'évolution a plissé le néocortex afin qu'il puisse tenir dans le crâne, à l'instar d'une feuille de papier roulée en boule pour être introduite dans un petit verre à liqueur.

Le néocortex contient des cellules nerveuses : les neurones. Ils sont répartis sous une telle densité que personne ne connaît exactement leur nombre. Il y en aurait cent mille au millimètre carré. Selon des anatomistes, le cortex humain compterait environ trente milliards de neurones, mais ce chiffre est évidemment très approximatif. Il pourrait y en avoir beaucoup plus ou beaucoup moins.

Ces trente milliards de cellules, c'est vous. Elles recèlent quasiment la totalité de votre mémoire, toutes vos connaissances, toutes vos aptitudes et toute l'expérience accumulée au cours de votre vie. Après vingt-cinq années d'étude du cerveau, ceci m'étonne toujours autant. Savoir qu'une mince feuille de cellules voit, ressent et élabore notre vision du monde est tout simplement sidérant. La chaleur d'un jour d'été et nos rêves d'un monde meilleur découlent de l'activité de ces cellules. Bien des années après avoir écrit son article dans *Scientific American*, Francis Crick rédigea un ouvrage sur le cerveau intitulé *L'hypothèse stupéfiante* (édité en France chez Plon). Cette hypothèse est que l'intelligence est créée par les cellules cérébrales. Il n'y a rien d'autre, rien de magique, pas d'ingrédients particuliers : seulement des neurones et une sarabande d'informations. Je présume que vous mesurez combien cette théorie est révolutionnaire. Elle creuse un profond fossé philosophique entre un ensemble de cellules et notre expérience consciente, bien que l'intelligence et le cerveau soient une seule et même chose. En appelant cette théorie « hypothèse », Crick fut politiquement correct. Que les cellules

cérébrales créent l'intelligence est un fait, pas une hypothèse. Nous devons comprendre ce que font ces trente milliards de cellules et comment elles le font. Fort heureusement, le cortex n'est pas qu'un amas de cellules amorphes. Une étude plus approfondie de sa structure nous éclairera sur la manière dont il élève l'intelligence humaine.

---

Revenons à la table de dissection et examinons le cerveau de plus près. A l'œil nu, le néocortex ne présente quasiment aucune délimitation. Il en existe bien sûr, comme le sillon profond qui sépare les deux hémisphères cérébraux et celui qui sépare les parties antérieures et postérieures. Mais, en quelque endroit que vous les examiniez, la surface des circonvolutions semble partout la même. Aucune ligne ou couleur ne différencie les aires spécialisées chacune dans les diverses informations sensorielles ou les divers types de pensée.

Des savants ont cependant pensé qu'il existe des délimitations. Bien avant que les neurobiologistes soient capables de discerner l'ensemble des circuits du cortex, ils savaient que des fonctions mentales étaient localisées dans telle ou telle région du cerveau. Si un traumatisme détruit le lobe pariétal droit d'un individu, ce dernier peut perdre sa capacité à percevoir — ou à concevoir — tout ce qui concerne la partie gauche de son corps, ou l'espace situé à sa gauche. Un traumatisme dans la région frontale de l'encéphale, appelée aire de Broca, compromet la capacité à appliquer les règles de grammaire, bien que le vocabulaire et la compréhension des mots ne soient pas affectés. Un traumatisme dans une aire appelée « gyrus fusiforme » empêche de reconnaître les visages. L'individu ne reconnaît plus sa mère, ses enfants, ni même son propre visage sur une photographie. Ces fascinants désordres ont très tôt incité les neurobiologistes à penser que le cortex est constitué de plusieurs zones, ou régions fonctionnelles.

Nous en avons appris long sur les aires fonctionnelles au cours du dernier siècle, mais il reste encore beaucoup à découvrir. Chacune de ces zones est semi-indépendante et semble spécialisée dans un aspect de la perception ou de la pensée. Elles sont organisées en un patchwork irrégulier qui varie peu d'un individu à un autre. Les fonctions sont rarement délimitées avec précision. D'un point de vue fonctionnel, elles sont réparties hiérarchiquement.

Cette notion de hiérarchie étant cruciale, nous nous y attarderons quelque peu afin de bien la définir. Nous y reviendrons en effet tout au long de cet ouvrage. Dans un système hiérarchique, certains éléments se trouvent, d'une manière abstraite, « au-dessus » ou « sous » d'autres. Dans la hiérarchie d'une entreprise, le sous-directeur est placé au-dessus de l'employé de bureau mais sous le directeur général. Ceci n'a rien à voir avec la notion physique d'au-dessus et d'en dessous. Même si le directeur général travaille à l'étage situé sous celui de l'employé de bureau, il n'en est pas moins hiérarchiquement « au-dessus ». J'insiste sur ce point pour clarifier sans ambiguïté ce que j'entends lorsque je parlerai de régions fonctionnelles au-dessus ou en dessous d'une autre. Ceci n'a rien à voir avec leur disposition physique dans le cerveau. Toutes les aires fonctionnelles du cortex résident dans la même enveloppe corticale convolutée. Ce qui fait qu'une région se trouve « au-dessus » ou « en dessous » d'une autre est la manière dont elles sont connectées. Dans le cortex, les aires inférieures envoient des informations vers les aires supérieures au travers d'un certain pattern de connectivité neuronale, tandis que les aires supérieures renvoient un feedback — un *biofeedback* pour être précis — vers les aires inférieures au travers d'un autre pattern de connexion. Il existe aussi des connexions latérales entre des aires situées dans des branches séparées de la hiérarchie, par analogie avec le sous-directeur qui communiquerait avec son homologue présent dans la filiale d'un autre pays. La cartographie détaillée d'un cortex de singe a été dévoilée par deux

chercheurs, Daniel Felleman et David van Essen. Elle révèle des dizaines de régions interconnectées selon une hiérarchie complexe. Nous pouvons penser que la cartographie du cortex humain est semblable.

La plus basse des régions fonctionnelles, celle des aires sensorielles primaires, est le lieu où les informations sensorielles arrivent en premier au cortex. Elles traitent l'information à leur niveau le plus brut, le plus élémentaire. Par exemple, l'information visuelle pénètre dans le cortex par une aire visuelle primaire appelée V1, en abrégé. Elle est concernée par les fonctionnalités de vision de bas niveau comme celles des détails, la perception d'un mouvement faible, la disparité rétinienne (propre à la vision stéréoscopique), ainsi que la perception des informations chromatiques de base et le contraste. L'aire V1 fournit des informations à d'autres aires comme V2, V4 et IT — nous y reviendrons —, ainsi qu'à de nombreuses autres aires annexes. Chacune est concernée par un aspect plus spécialisé ou abstrait de l'information. Par exemple, les cellules en V4 réagissent à des objets de complexité moyenne comme des formes en étoile de couleurs différenciées. Une autre aire appelée TM est spécialisée dans le mouvement des objets. Les échelons plus élevés du cortex visuel sont des aires qui représentent notre mémoire visuelle de toutes sortes d'objets comme des visages, des animaux, des outils, des parties du corps, etc.

Les autres sens sont dotés de hiérarchies analogues. Le cortex possède une aire auditive primaire appelée A1 ainsi qu'une hiérarchie de régions auditives placées au-dessus. Il est aussi doté d'une aire somatosensorielle (la sensation et l'image du corps) appelée S1, elle-même surmontée d'une hiérarchie de régions somatosensorielles. Enfin, les informations sensorielles excitent des aires associatives, c'est-à-dire des régions du cortex qui reçoivent des informations (inputs) envoyées par plusieurs sens. Par exemple, le cortex possède des aires recevant des informations provenant à la fois de la vue et du toucher. C'est grâce aux régions

associatives que vous prenez conscience du fait que la vision d'une mouche se promenant sur votre bras et le chatouillement que vous éprouvez ont une même cause. La plupart de ces aires reçoivent des informations traitées à un niveau hautement élevé, provenant de plusieurs sens ; les fonctions de ces aires nous sont encore obscures. Nous reviendrons plus longuement sur la hiérarchie corticale plus loin dans cet ouvrage.

Un autre ensemble d'aires, dans les lobes frontaux du cerveau, produit des sorties (outputs) moteurs. Le système moteur du cortex est lui aussi hiérarchiquement agencé. L'aire la plus basse, M1, entretient des connexions avec la moelle épinière et agit directement sur les muscles. Les aires plus élevées transmettent des commandes motrices sophistiquées à M1. La hiérarchie de l'aire motrice et celles des aires sensorielles se ressemblent énormément, comme si elles avaient été constituées de la même manière. Dans la région motrice, l'information descend le long de la hiérarchie vers M1 afin d'agir sur les muscles, tandis que dans les régions sensorielles, l'information remonte le long de la hiérarchie à partir des sens. En réalité, l'information se propage dans les deux sens. Ce qui est considéré comme un biofeedback dans les régions sensorielles est un output de la région motrice, et inversement.

La plupart des descriptions du cerveau sont fondées sur des organigrammes reflétant une vision très simpliste des hiérarchies. Les inputs (vue, ouïe, toucher…) parviennent dans les aires sensorielles primaires et sont traités au fur et à mesure qu'ils s'élèvent dans la hiérarchie. Ils sont ensuite transmis aux aires associatives, puis aux lobes frontaux du cortex, après quoi ils redescendent le long des aires motrices. Il va sans dire que cette vision est totalement erronée. Quand vous lisez à haute voix, l'information visuelle entre en réalité en V1, se propage vers les aires associatives puis vers le cortex moteur frontal, et finit par actionner les muscles de la bouche et du larynx pour former les sons vocaux. Mais ce n'est pas si simple. Dans la vision outrageusement simpliste contre laquelle je viens de mettre en garde, le

processus est généralement traité sous la forme d'une information circulant dans un seul sens, à l'instar des objets fabriqués sur une chaîne de montage. Or, dans le cortex, l'information circule aussi et toujours dans la direction opposée, un nombre plus élevé de projections se propageant tant dans le sens descendant de la hiérarchie que dans le sens montant. Quand vous lisez à haute voix, les régions supérieures du cortex envoient plus d'informations vers le bas du cortex visuel primaire que l'œil en reçoit en parcourant la page imprimée. Nous nous intéresserons au devenir de ces projections dans les chapitres à venir. Pour le moment, je tiens à insister sur ce fait : bien que l'ascension vers le haut de la hiérarchie soit indiscutable, nous devons bien nous garder de croire que l'information ne circule que dans un seul sens.

Revenons encore une fois à la table de dissection et supposons que nous avons installé un puissant microscope, coupé une mince lamelle de l'enveloppe corticale, coloré quelques cellules et collé l'œil à l'oculaire. Si toutes les cellules ont été teintes, nous ne distinguerons qu'une masse noire opaque car les cellules sont trop proches les unes des autres et trop intriquées. En revanche, si nous ne colorons que quelques rares cellules, les six couches précédemment mentionnées deviennent visibles. Elles se manifestent par une variation de densité des corps cellulaires, le type de cellule et leurs connexions.

Tous les neurones possèdent quelques caractéristiques communes. Hormis le corps cellulaire, qui est *grosso modo* arrondi, ils présentent des structures arborescentes, filaires, appelées « axones » et « dendrites ». Lorsque l'axone d'un neurone entre en contact avec la dendrite d'un autre, ils forment une petite connexion appelée « synapse ». C'est au niveau des synapses que l'impulsion nerveuse provenant d'une cellule influence le comportement d'une autre cellule. Le signal neuronal, ou potentiel, parvenant à une synapse incitera plutôt la cellule réceptrice à se potentialiser. Certaines synapses ont un effet opposé, réduisant la potentialisation de la cellule réceptrice. De ce fait, une synapse

peut être inhibitrice ou excitatrice. L'intensité de la synapse peut varier selon le comportement des deux cellules. La forme la plus simple de cette variation synaptique est celle qui se produit lorsque deux neurones génèrent un potentiel quasiment simultané ; l'intensité de la connexion entre les deux neurones est accrue. J'en dirai davantage plus tard sur ce processus appelé « apprentissage hebbien ». Outre la variation de l'intensité d'une synapse, il a été établi que des synapses entièrement nouvelles peuvent se former entre deux neurones. Ceci se produirait en permanence, bien que la preuve scientifique soit encore controversée. Nonobstant les détails sur la manière dont les synapses font varier leur intensité, il est certain que c'est la formation et le renforcement des synapses qui permettent aux mémoires d'être conservées.

Bien que le néocortex contienne de nombreux types de neurones, une seule classe représente huit sur dix d'entre eux. Ce sont les neurones pyramidaux, ainsi nommés à cause de la forme de leur corps cellulaire. Hormis la couche supérieure des six qui composent le cortex, qui possède d'innombrables axones mais très peu de cellules, chaque couche contient des neurones pyramidaux. Chaque neurone pyramidal est connecté à de nombreux autres neurones dans son voisinage immédiat, et chacun étend un long axone latéral vers des zones plus éloignées du cortex, ou vers des structures plus profondément enfouies comme le thalamus.

Une cellule pyramidale possède plusieurs milliers de synapses. Là encore, il est très difficile de savoir exactement combien, en raison de leur extrême densité et de leur petite taille. Le nombre de synapses varie de cellule en cellule, de couche en couche et de région en région. Si nous nous en tenions au fait qu'une cellule pyramidale moyenne comprend un millier de synapses (le chiffre réel serait plus proche de cinq ou dix mille), le néocortex compterait par conséquent trente trillions de synapses environ. C'est une quantité astronomique, qui dépasse notre entendement. Elle est apparemment suffisante pour conserver tout ce qui peut être appris dans une vie.

Albert Einstein aurait reconnu qu'il lui avait été simple, voire facile, de concevoir la théorie de la relativité. Elle découlait tout naturellement d'une seule observation, à savoir que la vitesse de la lumière est constante pour tout observateur, même si certains se déplacent à des vitesses différentes, ce qui est paradoxal. Ceci reviendrait à affirmer que la vitesse d'un ballon est identique quelle que soit la force avec laquelle il a été jeté, ou quelle que soit la vélocité de celui qui l'a jeté ou de celui qui l'observe. Chacun verrait la balle se déplacer à la même vitesse par rapport à lui, en toutes circonstances. Il semble que ceci ne puisse être possible. Mais il fut prouvé que c'est le cas pour la lumière. Einstein s'interrogea intelligemment sur les conséquences de cette singularité. Il réfléchit rationnellement à toutes les implications d'une vitesse constante de la lumière, ce qui l'amena à postuler des théories encore plus surprenantes comme le ralentissement du temps lorsque la vitesse s'accroît, ou le fait que la masse et l'énergie sont fondamentalement de même nature. Des ouvrages entiers ont été consacrés à la relativité, truffés d'exemples mettant en scène des trains, des projectiles et des éclairs. Cette théorie n'est pas ardue, mais elle est sans aucun doute paradoxale.

Une découverte analogue fut faite en neurobiologie, au sujet du cortex, si déroutante que certains chercheurs refusent d'y croire et que la plupart des autres l'ignorent car ils ne sauraient qu'en faire. Mais elle est si importante, qu'en prenant la peine d'en considérer attentivement et méthodiquement les implications, elle révèle ce qui se passe dans le néocortex et comment il fonctionne. Cette surprenante découverte découle en fait de l'anatomie fondamentale du cortex, mais il fallut une intuition rare pour la déceler. C'est à Vernon Mountcastle, un neurobiologiste de l'université John Hopkins de Baltimore, que l'on doit cette découverte. Il publia en 1978 un article intitulé « An Organizing Principle for Cerebral Function » (un principe organisateur de la fonction cérébrale) dans lequel il mettait en évidence la remarquable uniformité du cortex, tant dans son apparence que

dans sa structure. La région dévolue à l'ouïe ressemble à celle du toucher, qui ressemble à celle qui régit les muscles, qui ressemble à la région du langage de l'aire de Broca, qui ressemble à quasiment n'importe quelle autre région du cortex. Mountcastle suggéra que si ces régions se ressemblent toutes, c'est peut-être parce qu'elles exécutent les mêmes opérations de base. Le cortex utiliserait les mêmes outils computationnels pour accomplir toutes ses tâches.

Tous les anatomistes de cette époque, et même des décennies avant Mountcastle, reconnaissaient que le cortex paraît identique partout. C'est indéniable. Mais, au lieu de s'interroger sur la signification de cette similarité, ils recherchèrent les différences entre les aires du cortex, et ils finirent par les trouver. Ils supposèrent que si une région est dévolue au langage et une autre à la vision, il devait forcément y avoir des différences entre elles. En y regardant de près, elles sont discernables. Les régions du cortex varient en épaisseur, en densité de cellules, en proportions relatives de types de cellules, en longueurs de connexions horizontales, en densité synaptique et de bien d'autres manières parfois délicates à révéler. L'une des zones les plus étudiées, qui est une des couches de l'aire visuelle primaire V1, présente effectivement quelques divisions supplémentaires. La situation est la même que celle à laquelle étaient confrontés les biologistes du XVIII^e^ siècle. Ils passaient leur temps à ergoter sur des différences infimes entre les espèces. Leur grande victoire était de démontrer que deux souris qui se ressemblent appartiennent en réalité à des espèces différentes. Pendant des années, Darwin fit de même, étudiant surtout les mollusques. Mais il eut le génie de se demander pourquoi toutes ces espèces pouvaient tant se ressembler. C'est cette similarité qui était surprenante et digne d'intérêt, bien plus que les différences.

Mountcastle fit une observation semblable. Tandis que les anatomistes recherchaient des différences minimes entre les régions corticales, il montra qu'en dépit des différences le néocortex est

remarquablement uniforme. Les mêmes six couches, types de cellules et connexions existent partout. Les différences sont souvent si ténues que même les anatomistes expérimentés ne parviennent pas à se mettre d'accord. Pour Mountcastle, toutes les régions du cortex exécutent la même opération. Ce qui caractérise l'aire de la vision par rapport à l'aire motrice dépend de la manière dont les régions du cortex sont connectées les unes aux autres et aux autres parties du système nerveux central.

En fait, Mountcastle soutint que si une région du cortex diffère légèrement d'une autre, c'est à cause de la nature des connexions, et non parce que les fonctions élémentaires sont différentes. Il en conclut qu'il existe une fonction commune, un algorithme commun exécuté par et dans toutes les régions corticales. La vue ne diffère pas de l'ouïe, qui ne diffère pas d'un influx moteur. Ce sont nos gènes qui spécifient comment les régions du cortex sont connectées, et ce qui est spécifique à la fonction et à l'espèce, mais le tissu cortical lui-même est partout identique.

Réfléchissons-y un instant. Pour moi, la vue, l'ouïe et le toucher semblent très différents. Ces sens ont des caractéristiques fondamentalement différentes. La vue implique la perception de la couleur, de la texture, du modelé, de la profondeur et de la forme. L'ouïe implique la perception de la hauteur d'une note, du rythme et du timbre. Ces sensations sont très différentes. En quoi pourraient-elles être identiques ? Mountcastle n'affirme pas qu'elles sont identiques, mais que la manière dont le cortex traite les signaux auditifs est la même que pour les signaux visuels. Et bien sûr, c'est pareil pour les contrôles moteurs.

Les scientifiques et les ingénieurs ont, dans leur grande majorité, ignoré ou choisi d'ignorer l'hypothèse de Mountcastle. Lorsqu'ils s'efforcent de comprendre la vision ou de fabriquer un ordinateur capable de « voir », ils usent d'un vocabulaire et de notions techniques spécifiques à la vision. Ils parlent d'arêtes, de textures, de représentations tridimensionnelles. S'ils veulent comprendre les mécanismes du langage, ils élaborent des

algorithmes fondés sur les règles de grammaire, la syntaxe et la sémantique. Mais si Mountcastle avait raison, ces approches ne reflètent pas la manière dont le cerveau résout ces problèmes, d'où un risque d'échec. Si Mountcastle avait raison, l'algorithme du cortex doit être exprimé indépendamment de toute fonction ou sens particuliers. Le cerveau aurait recours à un même processus pour voir ou pour entendre. Le cortex exécuterait une action universelle susceptible d'être appliquée à n'importe quel type sensoriel ou système moteur.

La lecture de l'article de Vernon Mountcastle fut pour moi une révélation. C'était la pierre de Rosette de la neurobiologie : un article et une idée uniques unissant les diverses et merveilleuses capacités de l'intelligence humaine en un seul algorithme. Il dévoilait d'un seul coup le fourvoiement de toutes les tentatives passées pour comprendre et élaborer le comportement humain sous la forme de capacités distinctes. Appréciez la radicalité et l'élégance de l'hypothèse de Mountcastle. Les idées scientifiques les plus abouties sont toujours simples, élégantes et inattendues, comme celle-ci. A mon avis, elle fut, est et restera la découverte la plus importante dans le domaine de la neurobiologie. Aussi incroyable que cela puisse paraître, la plupart des scientifiques et des ingénieurs refusèrent d'y accorder crédit, choisirent de l'ignorer ou ne furent pas conscients de la portée de cette découverte.

---

Une partie de cette indifférence découle de la pauvreté des outils permettant d'étudier le flux d'informations à l'intérieur des six couches corticales. Les outils dont nous disposons opèrent à un niveau grossier et visent principalement à localiser où, dans le cortex, se manifestent les diverses capacités humaines, plutôt que quand et comment. Aujourd'hui, par exemple, une grande partie des magazines de vulgarisation favorise implicitement l'idée que le cerveau est un ensemble de modules hautement spécialisés. Les techniques d'imagerie fonctionnelle comme l'IRM (imagerie par

résonance magnétique) et la TEP (tomographie par émission de positrons) se focalisent presque exclusivement sur la cartographie du cerveau et sur les régions fonctionnelles décrites précédemment. Lors des expériences, un sujet volontaire est installé dans le scanner et exécute des tâches mentales ou motrices. Il s'agit pour lui de jouer à un jeu vidéo, de conjuguer des verbes, de lire des phrases, de reconnaître des visages, de nommer des images, d'imaginer quelque chose, de mémoriser des listes, de prendre des décisions financières, etc. Le scanner détecte quelles régions du cerveau sont plus actives que d'ordinaire au cours de ces tâches, et les traduit par des zones de couleur appliquées à la représentation du cerveau du sujet. Ces régions sont censées être celles excitées par la tâche. Des milliers d'expériences d'imagerie fonctionnelle du cerveau ont été faites et des milliers d'autres le seront. Nous traçons ainsi progressivement une image du lieu où se produisent certaines fonctions dans le cerveau humain adulte. Il est très facile de déterminer que « ceci est l'aire de la reconnaissance des visages, celle-là l'aire des mathématiques et celle-ci l'aire des activités musicales », et ainsi de suite. Mais tant que nous ne saurons pas comment le cerveau accomplit ces tâches, il est naturel de supposer qu'il exécute les différentes tâches de différentes manières.

Mais est-ce le cas ? Une masse grandissante d'évidences étayent la supposition de Mountcastle. Des exemples illustrent l'extrême flexibilité du néocortex. Tout cerveau humain correctement nourri et placé dans un environnement favorable peut apprendre n'importe laquelle des milliers de langues vivantes. Ce même cerveau peut aussi apprendre la langue des signes, une langue écrite, le langage musical, le langage mathématique, un langage informatique et l'expression corporelle. Il peut apprendre à vivre dans les contrées glacées du Grand Nord ou dans un désert aride. Il peut apprendre à devenir un grand maître des échecs, un pêcheur émérite, un fermier ou un physicien atomiste. Réfléchissez au fait que votre cerveau dispose d'une aire visuelle

spéciale qui semble tout particulièrement consacrée à la représentation des lettres et des chiffres. Cela signifie-t-il pour autant que vous êtes né avec une aire du langage déjà prête à traiter l'écrit ? C'est peu probable. Le langage écrit est une invention trop récente pour que nos gènes puissent avoir élaboré un mécanisme spécifique pour sa reconnaissance innée. C'est donc que le cortex se divise en aires fonctionnelles spécifiques à des tâches précises lors de l'enfance, en se fondant uniquement sur l'expérience. Le cerveau humain est doté d'une incroyable capacité à apprendre et à s'adapter à d'innombrables environnements qui n'existent que depuis peu. Ceci va dans le sens d'un système extrêmement souple, qui ne serait pas limité à un seul millier de solutions apportées à un seul millier de problèmes.

Les neurobiologistes ont aussi découvert que le câblage du cortex est incroyablement façonnable, c'est-à-dire qu'il est capable de se modifier et se recâbler de lui-même selon le type d'influx qui le parcourt. Par exemple, le cerveau des furets nouveau-nés peut être chirurgicalement recâblé afin que les yeux de l'animal envoient leurs signaux vers l'aire où se développe normalement l'ouïe. Ce qui est surprenant c'est que le furet développe alors un cheminement visuel fonctionnel dans les parties auditives de son cerveau. Autrement dit, il voit avec des tissus cérébraux qui normalement traitent le son. Des expériences analogues ont été faites à partir d'autres sens et d'autres régions. Par exemple, des parties du cortex visuel du rat sont transplantées, juste après la naissance, dans des régions qui sont habituellement celles du sens du toucher. Lorsque le rat grandit, le tissu transplanté traite le toucher plutôt que la vue. Les cellules ne sont donc pas nées spécialisées dans la vue, le toucher ou l'ouïe.

Le néocortex humain est vraiment très façonnable : des adultes muets de naissance traitent l'information visuelle dans des aires qui devraient devenir des régions réservées à l'ouïe. Quant aux adultes congénitalement aveugles, ils se servent de la partie postérieure de leur cortex, habituellement réservée à la vue, pour lire le braille.

L'alphabet braille étant tactile, vous penseriez que sa lecture active principalement les régions du toucher. Or, apparemment, aucune aire du cortex n'est destinée à ne plus rien représenter. Si le cortex visuel ne reçoit pas ou plus d'informations en provenance des yeux, il recherche d'autres patterns d'entrée aux alentours, en l'occurrence dans d'autres régions corticales.

Tout ceci tend à montrer que, lors du développement du cerveau, ses zones élaborent des fonctions spécialisées fondées largement sur le type d'information qui les alimente. Le cortex n'est pas davantage conçu pour exécuter diverses fonctions d'une manière rigide, à l'aide de divers algorithmes, que la surface de la Terre n'a été naturellement prédestinée à être découpée en un patchwork de nations. A l'instar de la géographie politique du globe, notre cortex est susceptible d'évoluer différemment selon les circonstances.

Les gènes imposent l'architecture globale du cortex, y compris la manière dont les régions sont interconnectées. Mais à l'intérieur de cette structure, le système est hautement flexible.

Mountcastle avait raison. Il existe un seul puissant algorithme implémenté dans chaque région du cortex. Lorsque des zones du cortex sont interconnectées selon la hiérarchie appropriée et qu'elles reçoivent un flux d'inputs, elles apprennent ce qu'est leur environnement. C'est pourquoi il n'y a aucune raison pour que les machines intelligentes du futur possèdent les mêmes sens et capacités que nous autres humains. L'algorithme cortical peut être déployé de manière innovante, avec des sens nouveaux et originaux, dans une enveloppe corticale de synthèse afin qu'une intelligence authentique, flexible, émerge par-delà le cerveau biologique.

---

Passons à un sujet se rapportant à la supposition de Mountcastle, et tout aussi surprenant. Les inputs acheminés vers le cortex sont fondamentalement tous identiques. Là encore, vous penserez pro-

bablement que vos sens sont des entités complètement distinctes. Car après tout, le son est transmis dans l'air par une succession d'ondes de compression, l'environnement visuel est transmis sous forme de lumière et le toucher est suscité par une pression sur la peau. Le son semble temporel, la vision principalement picturale et le toucher essentiellement spatial. Qu'y a-t-il de plus différent que le bêlement d'une chèvre comparé à la vue d'une pomme, elle-même comparée aux sensations physiques du volley-ball ?

Examinons tout cela de plus près. L'information visuelle provenant du monde extérieur est acheminée vers le cerveau par les millions de fibres du nerf optique. Après un bref transit par le thalamus, elle parvient au cortex visuel primaire. L'information auditive est transmise par les trente mille fibres du nerf auditif. Elle traverse quelques parties archaïques du cerveau et parvient au cortex auditif primaire. La moelle épinière achemine au cerveau les informations concernant le toucher et les sensations internes par le truchement d'un million de fibres. Elles sont reçues par le cortex somatosensoriel primaire. Voilà comment vous percevez le monde.

Les inputs peuvent être visualisés sous la forme d'un faisceau de fils électriques ou de fibres optiques. Vous avez sans doute déjà vu ces lampes à fibres optiques, avec un point de lumière colorée qui brille à l'extrémité de chaque brin. Les inputs envoyés au cerveau ressemblent à cela. Les fibres sont appelées « axones », le signal neuronal transporté est appelé « potentiel d'action » ou « potentiel » tout court ; il est de nature partiellement électrique et partiellement chimique. Les organes sensoriels qui les fournissent sont différents, mais une fois qu'ils ont lancé un potentiel d'action en direction du cerveau, ils sont tous pareils : ce ne sont que des patterns.

Quand vous regardez un chien, par exemple, un ensemble de patterns se propage le long des fibres du nerf optique jusqu'à la partie visuelle du cortex. Quand le chien aboie, un ensemble de patterns différents parcourt le nerf auditif jusqu'à la partie auditive

du cortex. Quand vous caressez le chien, un ensemble de patterns de sensations tactiles est produit au niveau de la main, se propage le long des fibres de la moelle épinière et atteint la partie du cortex consacrée au toucher. Chaque pattern — la vue du chien, l'audition du chien, le contact avec le chien — est ressenti différemment car chacun est canalisé vers la hiérarchie corticale par des chemins différents. Savoir où vont ces fibres, dans le cerveau, est important. Toutefois, au niveau abstrait des inputs sensoriels, ces derniers sont essentiellement les mêmes, et tous sont pris en charge de la même manière par le cortex à six couches. Vous voyez la lumière, entendez le son et ressentez la pression, mais à l'intérieur de votre cerveau, il n'y a pas de différence fondamentale entre ces types d'information. Un potentiel d'action est un potentiel d'action. Ces potentiels momentanés sont identiques, quelle qu'en soit l'origine. La seule chose que votre cerveau identifie, ce sont les patterns.

Vos perceptions et vos connaissances du monde découlent de ces patterns. Il n'y a pas de lumière dans notre tête : il y fait tout noir. Il n'y a pas davantage de son se propageant dans le cerveau : le silence y règne. En fait, le cerveau est la seule partie de notre corps qui ne possède aucune sensorialité. On pourrait enfoncer un doigt dedans sans que vous éprouviez quoi que ce soit. Toutes les informations qui entrent dans votre esprit y parviennent sous la forme de patterns spatiaux et de patterns temporels propagés par les axones.

Que faut-il entendre exactement par « pattern spatial » et « pattern temporel » ? Examinons tour à tour chacun de nos principaux sens. La vue transporte des informations à la fois spatiales et temporelles. Les *patterns spatiaux* sont des patterns qui coïncident dans le temps ; ils sont produits lorsque de multiples récepteurs d'un même organe sensoriel sont excités simultanément. L'organe sensoriel de la vue est la rétine. L'image entre par la pupille, est inversée par le cristallin, frappe la rétine et produit un pattern spatial, qui est aussitôt acheminé vers le cerveau. Certains s'imaginent qu'une petite image à l'envers apparaît dans les aires

visuelles, mais il n'en est rien. Il n'y a pas de représentation graphique. La notion d'image n'existe même pas. Fondamentalement, il se produit seulement une activité électrique se déclenchant en patterns. Leur qualité graphique s'estompe rapidement tandis que le cortex traite l'information, envoyant les composants du pattern en haut et en bas de différentes aires, les passant au crible et les filtrant.

La vue se fonde aussi sur des *patterns temporels*, qui varient dans le temps. Alors que l'aspect spatial de la vision est intuitivement évident, son aspect temporel l'est moins. Trois fois par seconde environ, les yeux se livrent à de brusques mouvements appelés « saccades ». Ils fixent un point puis soudainement un autre. Chaque fois que l'œil bouge ainsi, l'image sur la rétine change. Il en résulte que le pattern acheminé vers le cerveau a lui aussi complètement changé à chaque saccade. Et c'est là le cas le plus simple, celui d'un observateur immobile regardant une scène fixe. Mais dans notre existence, nous bougeons sans cesse la tête, le corps se déplace dans un environnement changeant, modifiant continuellement le point de vue. Notre impression consciente est celle d'un monde stable empli d'objets et de gens qu'il est facile de suivre visuellement. Cette sensation n'est possible que grâce à la capacité du cerveau à gérer un torrent d'images rétiniennes qui ne répètent jamais le même pattern. La vue, en tant que patterns acheminés vers le cerveau, est un flux aussi changeant qu'un fleuve. La vision est plus comparable à un son qu'à un tableau.

Beaucoup de spécialistes de la vision ignorent la notion de saccades et de patterns rapidement changeants. Travaillant sur des animaux anesthésiés, ils étudient comment la vision se produit lorsqu'un animal inconscient fixe un point. Ce faisant, ils éludent la dimension temporelle. Il n'y a rien d'erroné dans cette démarche ; l'élimination de variables fait partie des méthodes scientifiques. Mais dès lors, ils ne tiennent plus compte d'un composant crucial de la vue, qui y participe totalement. La

notion de temps doit occuper une place centrale dans toute approche scientifique de la vision.

En ce qui concerne l'ouïe, nous avons l'habitude de prendre en compte sa dimension temporelle. Il nous paraît évident que le son, le langage parlé ou la musique varient dans le temps. Il n'est pas possible d'écouter une chanson en l'espace d'un éclair, pas davantage que toute une phrase puisse être dite en un clin d'œil. Un morceau de musique n'existe que dans la durée. C'est pourquoi nous ne pensons usuellement pas au son en termes de pattern spatial. D'une certaine manière, c'est l'inverse de la vue : l'aspect temporel est immédiatement apparent, mais l'aspect spatial l'est moins.

L'ouïe possède également un composant spatial. Le son est converti en potentiels d'action par un organe en colimaçon situé dans chaque oreille interne, la cochlée. Minuscule, opaque, en spirale et incorporée à l'os le plus dur du corps humain, le rôle de la cochlée a été révélé il y a plus d'un demi-siècle par un médecin hongrois, Georg von Bekesy. En élaborant des modèles de l'oreille interne, Bekesy découvrit que chaque composant d'un son fait vibrer une partie spécifique de la cochlée. Les sons à fréquence élevée produisent des vibrations dans la base ferme de l'organe. Les sons en basse fréquence font vibrer la partie la plus externe, qui est souple. Les sons à moyenne fréquence agissent sur les parties intermédiaires. La cochlée est littéralement pavée de neurones qui se déclenchent dès qu'ils sont secoués. Dans la vie courante, les deux cochlées vibrent simultanément à plusieurs fréquences. A chaque moment, un nouveau pattern de stimulations spatial apparaît sur toute la longueur de la cochlée. Et à chaque moment, un nouveau pattern de stimulations spatial se propage le long du nerf auditif. Là encore, nous constatons que cette information sensorielle se réduit à des patterns spatio-temporels.

Les gens ne pensent généralement pas au toucher en termes de phénomène temporel, mais il n'en est pas moins aussi temporel que spatial. Vous pouvez en faire l'expérience : demandez

à quelqu'un de mettre sa main en creux, la paume vers le haut et de fermer les yeux. Placez un petit objet dans sa main — une bague, une gomme… — et demandez-lui de l'identifier sans bouger aucune partie de la main. Il ne disposera, pour seuls indices, que du poids et éventuellement d'une appréciation de la taille. Demandez-lui de garder les yeux fermés mais de refermer ses doigts sur l'objet. Il y a de fortes chances pour qu'alors il l'identifie rapidement. En permettant aux doigts de se mouvoir, la notion de temps a été ajoutée à la perception tactile. Il y a là une analogie directe entre la fovéa située au milieu de la rétine et le bout des doigts : tous deux ont une très grande acuité perceptive. Le toucher aussi est comparable à l'audition. La capacité à faire un usage complexe du toucher, en boutonnant une chemise ou en ouvrant la porte palière dans le noir, dépend des patterns continûment variables produits par la sensation tactile.

Nous apprenons à l'école que l'être humain possède cinq sens : la vue, l'ouïe, le toucher, l'odorat et le goût. En réalité, nous en avons davantage. La vision met en œuvre trois sens : le mouvement, la couleur et la luminance (contraste entre le noir et le blanc). Le toucher est sensible à la pression, à la température, à la douleur et aux vibrations. Tout un ensemble de capteurs nous informe sur la position des parties du corps et les angles qu'elles adoptent : il s'agit du système proprioceptif (de *proprio*, en soi-même, et *ceptif*, contraction de « perceptif »). Nul ne pourrait s'en dispenser car il régit la perception de ce qui se passe à l'intérieur du corps, notamment aux niveaux musculaires et viscéraux. Nous disposons aussi d'un système vestibulaire dans l'oreille interne qui régit le sens de l'équilibre. Certains de ces sens nous paraissent plus riches et plus flagrants que d'autres, mais tous alimentent le cerveau en influx de patterns spatiaux se propageant en permanence le long des axones.

Le cortex ne perçoit ni ne connaît le monde directement. La seule chose qu'il en sait, ce sont les patterns que lui acheminent

les axones. La perception du monde découle de ces patterns, y compris la conscience. En fait, le cerveau ne peut savoir où le corps s'arrête et où commence le monde extérieur. Les neurobiologistes qui ont étudié l'image du corps ont découvert que la perception de son propre corps est beaucoup plus flexible qu'il y paraît. Par exemple, si je vous donne un petit râteau de croupier et si vous l'utilisez pour atteindre des jetons et les ramener vers vous, vous aurez vite l'impression que le râteau fait partie de votre corps. Le cerveau modifiera ses prévisions afin de les adapter aux nouveaux patterns d'entrée tactiles.

---

L'idée que pour le cerveau les patterns produits par différents sens sont équivalents est quelque peu surprenante, et bien que parfaitement comprise, elle n'est pas largement appréciée. Voici d'autres exemples, dont le premier est reproductible chez vous. Il suffit d'un partenaire, d'un panneau en carton capable de tenir à la verticale et d'une fausse main. L'idéal est d'utiliser une de ces mains en caoutchouc vendues dans les boutiques de farces et attrapes, mais l'expérience peut être faite avec une main dessinée sur une feuille de papier. Placez votre véritable main sur le dessus de la table, à quelques centimètres de la fausse main, toutes deux orientées dans le même sens (doigts dans la même direction, chaque paume vers le haut ou vers le bas). Placez ensuite le panneau en carton entre les deux mains de manière que vous ne puissiez voir que la fausse. Pendant que vous fixez la fausse main, la tâche de votre partenaire consistera à toucher simultanément chacune des mains au même endroit. Par exemple, il tapotera chaque petit doigt de sa racine à l'ongle à la même vitesse, puis il tapera trois fois, rapidement, sur la deuxième phalange de chacun des index, tracera quelques cercles sur le dos ou dans la paume de chacune des mains, et ainsi de suite. Au bout d'un petit moment, les aires cérébrales où se rejoignent les patterns visuels et somatosensoriels — les aires associatives mentionnées précédemment dans ce

chapitre — se confondent. Vous ressentirez les sensations appliquées à la fausse main comme si elle vous appartenait.

Un autre exemple fascinant d'« équivalence de pattern » est appelé « substitution sensorielle ». Elle pourrait révolutionner la vie de ceux qui ont perdu la vue dans leur jeune âge, et peut-être apporter une aide précieuse aux aveugles de naissance. Elle pourrait aussi susciter de nouvelles interfaces homme-machine profitables à nous tous.

Réalisant que tout, dans le cerveau, repose sur des patterns, Paul Bach y Rita, professeur d'ingénierie biomédicale à l'université du Wisconsin, a développé une technique permettant d'afficher les patterns visuels sur la langue humaine. A l'aide d'un tel dispositif d'affichage, les aveugles pourraient apprendre à « voir » par le truchement des sensations perçues par la langue.

Voici le principe : une petite caméra est fixée sur le front du sujet et un composant placé sur sa langue. Les images filmées sont converties, pixel par pixel, en points de pression sur la langue. Une scène affichée par des centaines de pixels sur un écran peut être transformée en un pattern de centaines de minuscules points de pression sur la langue. Le cerveau apprend rapidement à interpréter correctement les patterns.

L'une des premières personnes à essayer ce dispositif lingual fut Erik Weihenmayer, un athlète de niveau international devenu aveugle à l'âge de treize ans, et qui avait décidé que son handicap ne mettrait jamais un terme à ses ambitions. En 2002, Weihenmayer escalada le mont Everest, devenant le premier aveugle à entreprendre mais surtout à réussir cet exploit.

En 2003, Weihenmayer testa le dispositif lingual et vit des images pour la première fois depuis son adolescence. Il était capable de distinguer un ballon roulant vers lui sur le plancher, de saisir un verre posé sur la table, de jouer à « Pierre, feuille et ciseaux » et même de pratiquer un art martial issu du « Jan Ken Pon » birman, fondé sur des postures corporelles. Par la suite, il déambula dans un hall, distingua une porte, l'examina ainsi que

son bâti et remarqua qu'elle comportait des signes. Les images d'abord perçues comme sensations linguales furent rapidement perçues comme des images situées dans l'espace.

Ces exemples démontrent une fois de plus l'extrême flexibilité du cortex et le fait que les inputs reçus par le cerveau ne sont que des patterns. Peu importe d'où ils proviennent. S'ils sont corrélés dans le temps d'une manière cohérente, le cerveau saura leur donner du sens.

---

Tout ceci ne devrait pas être trop surprenant si nous admettons que le cerveau ne reconnaît que des patterns. Le cerveau est une machine à patterns. Il n'est pas absurde de soutenir qu'il fonctionne en termes d'écoute de la vision, mais au niveau le plus fondamental, nous en revenons toujours aux patterns. Qu'importe à quel point les activités des différentes aires corticales peuvent être différentes, car c'est toujours le même algorithme cortical qui est à l'œuvre. Le cortex n'a que faire de savoir si les patterns proviennent de la vue, de l'ouïe ou de tout autre sens. Il n'a que faire de savoir si les inputs proviennent d'un seul organe sensoriel ou de quatre. Et il n'a que faire de savoir si vous percevez le monde à l'aide d'un sonar, d'un radar ou par des champs magnétiques, si vous avez des tentacules à la place des mains, ou même si vous vivez dans un monde en quatre dimensions au lieu de trois.

Cela signifie que vous n'avez besoin de connaître aucun de vos sens ni aucune combinaison de sens particulière pour être intelligent. Helen Keller était sourde et aveugle, ce qui ne l'a pas empêchée d'apprendre des langues et devenir un écrivain plus doué que la plupart de ceux qui voient et entendent. C'était une personne très intelligente, privée de deux des principaux sens, mais dont l'incroyable flexibilité du cerveau lui permit de percevoir et de comprendre le monde aussi bien que les individus jouissant de leurs cinq sens.

Cette remarquable flexibilité de l'esprit humain nourrit mes espoirs de voir naître des technologies inspirées par le cerveau. Quand je pense à la création de machines intelligentes, je m'interroge sur la nécessité de s'attacher à tout prix aux sens qui nous sont familiers. Lorsqu'il nous sera possible de décrypter l'algorithme néocortical et développer ainsi une reconnaissance des patterns, nous pourrons l'appliquer à tout système que nous voudrions rendre intelligent. L'une des grandes caractéristiques de l'ensemble de circuits inspiré du néocortex est que nous n'aurons pas besoin d'être particulièrement astucieux pour le programmer. A l'instar du cortex « auditif » d'un furet qui peut devenir « visuel » par un recâblage, à l'instar du cortex « visuel » des aveugles qui s'adapte à un autre usage, un système exécutant l'algorithme néocortical sera intelligent quel que soit le genre de pattern que nous choisirons de lui communiquer. Nous devrons cependant faire preuve d'intelligence lors du réglage des nombreux paramètres du système, que nous devrons aussi entraîner et instruire. Les milliards d'informations neuronales mises en œuvre par le cerveau pour produire des pensées élaborées et créatives iront de soi, comme c'est tout naturellement le cas pour les enfants.

Finalement, l'idée que les patterns sont l'assise fondamentale de l'intelligence soulève quelques intéressantes questions philosophiques. Quand je me trouve dans une pièce avec quelques amis, comment est-ce que je sais qu'ils sont là, voire qu'ils sont réels ? Mon cerveau reçoit un ensemble de patterns qui concordent avec des patterns reçus dans le passé. Ces patterns correspondent aux gens que je connais, à leur visage, leur voix, leur comportement habituel et toutes sortes d'autres faits les concernant. J'ai appris à m'attendre à ce que ces patterns se produisent de manière prévisible. Mais après tout ce n'est qu'un modèle. Toute notre connaissance du monde est un modèle reposant sur des patterns. Sommes-nous certains que le monde est réel ? Ce thème a été abordé par de nombreux philosophes et aussi, accessoirement, par les romans et le cinéma de science-fiction. Il ne s'agit pas ici

de mettre en doute que les gens et les objets sont véritablement là. Ils le sont. Mais notre certitude de l'existence du monde est fondée sur la cohérence de patterns et la manière dont nous les interprétons. Il n'existe rien qui puisse être apparenté à une perception directe. Nous ne sommes pas équipés de « détecteurs de gens ». Rappelez-vous que le cerveau est une boîte obscure inerte ne possédant aucune connaissance autre que la propagation de patterns le long de ses fibres. Votre perception du monde découle de ces patterns et de rien d'autre. L'existence peut certes être objective, mais seuls les patterns spatio-temporels qui parcourent les faisceaux d'axones nous permettent de l'appréhender.

Cette discussion nous conduit à une interrogation sur les relations entre hallucination et réalité. Si vous parvenez à percevoir des sensations provenant d'une main en caoutchouc et à « voir » au travers d'une stimulation de la langue, êtes-vous de la même manière trompé par vos sens lorsque vous percevez le toucher de votre propre main et voyez avec vos propres yeux ? Peut-on se fier au monde tel qu'il apparaît ? Oui. Le monde absolu existe réellement sous une forme très proche de la manière dont nous le percevons. Notre cerveau ne peut cependant l'appréhender directement.

Le cerveau est informé de ce monde absolu grâce à un ensemble de sens qui n'en détectent que des parties. Les sens produisent des patterns acheminés vers le cerveau puis traités par un même algorithme cortical afin de restituer un modèle du monde. A cet égard, le langage parlé et le langage écrit, bien que très différents au niveau sensoriel, sont perçus d'une façon remarquablement semblable. Dans la même veine, le modèle du monde élaboré par Helen Keller était très proche du vôtre et du mien, malgré un équipement sensoriel considérablement diminué. Les patterns construisent dans le cerveau un modèle du monde proche de la réalité et, ce qui est le plus remarquable, le mémorisent. C'est de la mémoire, c'est-à-dire ce qu'il advient des patterns après leur arrivée dans le cortex, qu'il sera question au prochain chapitre.

# 4

# LA MÉMOIRE

**Pendant** que vous lisez ce livre, marchez dans une rue animée, écoutez une symphonie ou réconfortez un enfant qui pleure, votre cerveau est inondé de patterns spatiaux et temporels provenant de tous vos sens. Le monde est un océan de patterns constamment changeants qui viennent déferler dans le cerveau. Comment gérez-vous cet afflux et lui donnez-vous du sens ? Les patterns traversent diverses parties archaïques du cerveau pour finalement parvenir au néocortex. Mais que leur arrive-t-il au moment où ils entrent dans le cortex ?

Dès les débuts de la révolution industrielle, les gens se sont complu à considérer le cerveau comme une sorte de machine. Ils savaient certes qu'il ne contenait ni rouages ni engrenages, mais c'était à leurs yeux la meilleure métaphore : une information entrait dedans, et le cerveau-machine définissait comment le corps devait réagir. A l'ère informatique, le cerveau fut considéré comme un type particulier de machine : l'ordinateur programmable. Comme nous l'avons vu au Chapitre 1, les chercheurs en intelligence artificielle adoptèrent cette

vision, arguant que le peu de progrès en ce domaine n'était dû qu'à la lenteur et au manque de puissance des ordinateurs, comparés au cerveau humain. Selon eux, les performances des ordinateurs d'aujourd'hui sont comparables à celles du cerveau d'un cafard ; quand ces machines seront plus grosses et plus rapides, elles deviendront enfin aussi intelligentes que des humains.

Un problème a été largement ignoré dans cette analogie ordinateur-cerveau : les neurones sont plutôt lents comparés à la réactivité des transistors. Un neurone collecte les inputs des synapses et les combine pour décider quand émettre un potentiel en direction d'autres neurones. Il effectue cette opération et se réinitialise en 5 millisecondes (ms) environ, soit près de deux cents fois par seconde. Ceci semble rapide, mais un ordinateur récent, basé sur des composants en silicium, peut effectuer un milliard d'opérations par seconde. Cela signifie qu'une opération informatique élémentaire est cinq millions de fois plus rapide qu'une opération élémentaire effectuée par le cerveau ! C'est là une différence énorme. Comment se fait-il alors que le cerveau puisse être plus rapide que le plus puissant de nos ordinateurs numériques ? « C'est tout simple, affirment les tenants de l'analogie ordinateur-cerveau : le cerveau est un ordinateur massivement parallèle. Il possède des millions de cellules qui calculent toutes en même temps. Ce parallélisme démultiplie énormément la puissance de traitement de l'encéphale. »

J'ai toujours estimé que cet argument est fallacieux. Une expérience simple, appelée la « règle des cent étapes », le démontrera. Un être humain peut exécuter des tâches significatives en bien moins d'une seconde. Par exemple, je vous montre une photo en vous demandant de déterminer s'il y a un chat dans l'image. Vous devez appuyer sur un bouton s'il y en a un, mais pas si c'est un ours, un phacochère ou un navet. Cette tâche est actuellement difficile voire impossible pour un ordinateur, alors que l'être humain est capable de réagir en une demi-seconde, voire moins. Les neurones étant lents, en une demi-seconde, l'information parvenant au

cerveau ne peut traverser qu'une chaîne longue d'une centaine de neurones. Autrement dit, le cerveau ne « calcule » la solution à un tel problème qu'en cent étapes ou moins, indépendamment du nombre de neurones susceptibles d'être impliqués. Entre le moment où la lumière pénètre dans l'œil et celui où vous appuyez sur le bouton, une chaîne de cent neurones au maximum est activée. Le calculateur numérique qui tenterait de résoudre le même problème passerait par des millions d'étapes. Une centaine d'instructions informatiques sont tout juste suffisantes pour déplacer un unique caractère à l'écran, et ne parlons pas des opérations autrement plus intéressantes.

Mais si plusieurs millions de neurones œuvrent ensemble, n'est-ce pas comparable à un ordinateur parallèle ? Pas vraiment. Le cerveau opère certes en parallèle et les ordinateurs aussi, mais c'est tout ce qu'ils ont en commun. Les ordinateurs parallèles mettent en œuvre plusieurs calculateurs ou processeurs rapides pour résoudre des problèmes complexes, comme la prévision du temps. Pour connaître la météo à venir, il faut procéder à des calculs sur d'innombrables variables correspondant aux conditions physiques en de nombreux points de la planète. Chaque ordinateur peut travailler sur une localisation différente au même moment. Mais, même si des centaines ou des milliers de calculateurs fonctionnent en parallèle, chaque ordinateur particulier doit cependant exécuter des milliards ou des trillions d'étapes pour accomplir la tâche. L'ordinateur massivement parallèle le plus énorme que nous puissions concevoir ne saura rien faire d'utile en une centaine d'étapes, et cela quelles que soient ses performances.

Passons à une analogie. Supposons que je vous demande de transporter à pied une centaine de blocs de pierre à travers le désert. Vous pouvez en porter un seul à la fois et il vous faudra un million de pas pour traverser le désert. Comme vous vous rendez compte qu'achever la tâche sera fort long, vous embauchez une centaine d'ouvriers pour l'effectuer en parallèle. La tâche sera

cent fois plus rapide, mais il faudra quand même que tout le monde effectue un million de pas. L'embauche d'un plus grand nombre d'ouvriers — un millier, disons — n'apportera aucun gain. Quel que soit le nombre de travailleurs auxquels vous ferez appel, le problème ne peut en aucun cas être résolu en moins de temps qu'il n'en faut pour faire un million de pas. Il en va de même pour les ordinateurs parallèles. Passé un certain point, ajouter davantage de processeurs n'apporte plus rien. Quel que soit leur nombre et quelle qu'en soit la vitesse, l'ordinateur ne pourra pas « calculer » la réponse à un problème aussi difficile que celui des cent étapes.

Alors, comment le cerveau s'y prend-t-il pour effectuer en cent étapes une tâche si difficile que l'ordinateur le plus massivement parallèle ne saurait la résoudre en un million ou un milliard d'étapes ? La réponse est que le cerveau ne « calcule » pas les solutions au problème. Il les extrait de la mémoire. Par essence, les solutions ont été stockées dans la mémoire il y a longtemps. Or, il suffit de quelques étapes pour extraire une information. Les lents neurones sont tout juste assez rapides pour procéder de la sorte, mais ils sont aussi la mémoire elle-même. Le cortex tout entier est un système mnémonique. Ce n'est pas du tout un ordinateur.

---

Permettez-moi d'illustrer par un exemple la différence entre le *calcul* d'une solution à un problème et sa résolution *en utilisant la mémoire*. Considérons la tâche consistant à attraper une balle. Quelqu'un vous jette un ballon, vous le voyez arriver et en moins d'une seconde, vous vous élancez dans les airs. Cela ne paraît pas bien compliqué, du moins tant que vous n'essayez pas de programmer un bras robotisé qui en ferait autant. Comme l'ont constaté les nombreux étudiants qui s'y sont essayés, c'est presque impossible. Quand des ingénieurs ou des informaticiens abordent ce problème, ils tentent de calculer d'abord la trajectoire du ballon afin de déterminer l'endroit où il se trouvera au

moment de la prise. Ces calculs exigent la résolution d'un ensemble d'équations assez trapues. Ensuite, toutes les articulations du bras robotisé doivent être ajustées simultanément afin de positionner correctement la main pour la réception. Ceci implique un autre ensemble d'équations encore plus ardues que les premières. Enfin, toute l'opération doit être répétée plusieurs fois afin qu'à chaque nouvelle arrivée du ballon le robot acquière des informations encore meilleures sur l'emplacement et la trajectoire du ballon. Si le robot devait attendre, pour se mettre en mouvement, de savoir par où le ballon va arriver, ce serait trop tard pour l'attraper. Il doit commencer à se mouvoir alors qu'il ne possède encore que des données fort réduites sur l'emplacement du ballon et ajuster continuellement son attitude au fur et à mesure que le ballon se rapproche. Il faudra un million d'étapes à l'ordinateur pour résoudre les nombreuses équations mathématiques permettant de l'attraper. Bien qu'un ordinateur puisse être efficacement programmé pour résoudre ce problème, la règle des cent étapes nous apprend que le cerveau s'y prend autrement : il fait appel à la mémoire.

Comment attraper un ballon en recourant à la mémoire ? Le cerveau conserve en mémoire les commandes aux muscles requises pour attraper un ballon (en plus d'autres comportements acquis). Quand une balle est lancée, il se produit trois événements. D'abord, la mémoire appropriée est spontanément rappelée par la vision du ballon. Ensuite, la mémoire se souvient véritablement d'une succession temporelle d'ordres adressés aux muscles. Enfin, la mémoire extraite est ajustée tandis qu'elle est rappelée, afin qu'elle s'accommode aux particularités du moment, comme la trajectoire réelle du ballon et la position du corps. Aucune mémoire indiquant comment attraper un ballon n'a été programmée dans le cerveau. Elle est le résultat d'années de pratique répétée ; elle est stockée — mais pas calculée — dans les neurones.

Peut-être objecterez-vous que chaque interception de balle est légèrement différente et que vous ajouterez : « Vous venez

d'affirmer que la mémoire ajuste continuellement le geste en fonction de l'endroit où se trouve le ballon qui arrive… Ceci n'exige-t-il pas de résoudre ces mêmes équations que nous cherchions à éviter ? » Cela pourrait être le cas, mais la nature a résolu le problème de la variation d'une manière différente et très astucieuse. Comme nous le verrons plus tard dans ce chapitre, le cortex crée ce que nous appelons des *représentations invariantes*, qui gèrent automatiquement les variations qui se produisent dans le monde environnant. Une bonne analogie consiste à imaginer ce qui se passe lorsque vous vous asseyez au bord d'un matelas d'eau : l'oreiller et les autres personnes qui se trouvent sur le matelas sont aussitôt repoussés en un nouvel agencement. Le matelas ne calcule pas de combien chaque objet doit être élevé ; les propriétés physiques de l'eau et la plasticité du matelas se chargent d'ajuster automatiquement la forme. Au prochain chapitre, nous découvrirons que le cortex à six couches fait presque de même, pour ainsi dire, avec les informations qui s'y propagent.

---

Le néocortex n'est donc pas un ordinateur, parallèle ou non. Au lieu de calculer des réponses à des problèmes, il fait appel à la mémoire stockée pour résoudre les problèmes et engendrer un comportement.

Les ordinateurs aussi ont une mémoire, sous la forme de composants électroniques et de disques durs. Ceci dit, quatre attributs de la mémoire néocorticale sont fondamentalement différents d'une mémoire informatique :

- **Le néocortex stocke des séquences de patterns.**
- **Le néocortex se souvient auto-associativement des patterns.**
- **Le néocortex stocke les patterns sous une forme invariante.**
- **Le néocortex stocke les patterns dans une hiérarchie.**

Les trois premières différences seront étudiées dans ce chapitre. J'ai déjà introduit le concept de hiérarchie du néocortex au

Chapitre 3. Au Chapitre 6, je dévoilerai son importance et comment il fonctionne.

La prochaine fois que vous raconterez une histoire, prenez un peu de recul et observez qu'il est possible de ne relater qu'un seul aspect de la narration à la fois. Il est impossible de dire en une seule fois tout ce qui se passe, quelle que soit la rapidité du débit de votre voix ou celle de mon écoute. Vous devez finir une partie de l'histoire avant de passer à une autre. Ce n'est pas uniquement parce que le langage parlé est sériel. Que la narration soit écrite, orale ou visuelle, elle est toujours sérielle. C'est parce que l'histoire est stockée séquentiellement dans votre tête et que vous ne pouvez vous en souvenir que de la sorte. Il est impossible de se remémorer la totalité de l'histoire en une seule fois. A vrai dire, il est quasiment impossible de penser à quoi que ce soit de complexe autrement qu'en une série d'événements ou de pensées.

Vous aurez peut-être remarqué qu'en racontant une histoire, certaines personnes ne parviennent pas à aller à l'essentiel. Elles se perdent dans des détails superflus et des digressions, ce qui peut être irritant. Vous avez alors envie de lui ordonner : « Allez au fait ! » En réalité, elles racontent l'histoire comme elle leur vient, sans pouvoir l'exprimer d'une autre manière.

Un autre exemple : je vous demande d'imaginer votre foyer. Fermez les yeux et visualisez-le. En imagination, allez à la porte d'entrée. Essayez de vous la représenter. Ouvrez-la et entrez. Regardez à présent à gauche. Que voyez-vous ? Regardez à droite. Qu'est-ce qu'il y a ? Dirigez-vous vers la salle de bains. Qu'y a-t-il à gauche ? A droite ? Sur la commode de droite ? Qu'est-ce qu'on trouve dans la cabine de douche ? Vous connaissez tous ces lieux et aussi des milliers d'autres détails, et vous vous en souvenez sans peine avec précision. Cette mémoire est stockée dans le cortex. Vous seriez tenté de dire que tous ces objets font partie de votre mémoire du foyer, mais vous ne pouvez penser à tous simultanément. Il existe de toute évidence une mémoire thématique, mais tous les détails ne peuvent pas être présents en même temps à l'esprit. Vous avez une mémoire

précise de votre foyer, mais pour vous en souvenir, vous devez procéder séquentiellement, comme vous venez de le faire.

Toutes les mémorisations sont ainsi faites. Vous devez vous astreindre à des séquences temporelles. Un pattern (approcher de la porte) appelle le prochain pattern (franchir le seuil), qui appelle le prochain pattern (s'avancer dans l'entrée ou prendre l'escalier), et ainsi de suite. Chacun correspond à une action que vous aviez faite précédemment. Bien sûr, il est possible par un choix conscient de modifier l'ordre de description de l'habitation. Vous pouvez passer directement du rez-de-chaussée au deuxième étage si vous désirez ne plus respecter la mémorisation séquentielle. Mais si vous décidez de décrire minutieusement chaque chambre ou objet, vous en revenez à la mémorisation séquentielle. De véritables pensées aléatoires n'existent pas. La mémorisation tend toujours à procéder par association d'idées.

Vous connaissez l'alphabet. Essayez de le réciter à rebours : c'est très difficile car vous ne procédez généralement pas ainsi. Si vous voulez savoir ce que représente l'apprentissage de la lecture pour un jeune enfant, forcez-vous à dire l'alphabet à l'envers. C'est plutôt ardu et laborieux. Votre mémoire de l'alphabet est une succession de patterns. Ce n'est pas quelque chose qui est stocké ou dont on peut se rappeler l'ensemble simultanément ou dans un ordre arbitraire. Il en va de même pour les jours de la semaine, les mois de l'année, votre numéro de téléphone et d'innombrables autres choses.

La mémoire des sons est un bel exemple de séquences temporelles mémorisées. Pensez à un air que vous connaissez. Il est impossible de se l'imaginer dans sa globalité, en une seule fois, mais seulement séquentiellement. Vous pouvez commencer à y penser par son début, ou le prendre en cours de route au moment du refrain, une note après l'autre. Mais il est impossible de le chantonner à rebours. Vous ne pouvez vous rappeler de la chanson que dans l'ordre où les notes ont été jouées, dans la durée, et de la même manière que vous l'avez apprise.

Ceci s'applique aussi à la mémoire sensorielle de très bas niveau. Prenons par exemple celle, tactile, des textures. Votre cortex a mémorisé ce que vous ressentez en serrant une poignée de gravier dans la main, en glissant le doigt sur du velours ou en appuyant sur une touche de piano. A l'instar de l'alphabet et des sons, la mémoire est fondée sur des séquences ; ces dernières sont seulement plus brèves, s'étendant sur des fractions de seconde au lieu de secondes ou de minutes. Si vous enfoncez votre main dans un seau de gravier au moment de vous endormir, vous ne saurez pas, au réveil, dans quoi elle se trouve tant que vous n'aurez pas bougé les doigts. La mémoire tactile de la texture du gravier repose sur des séquences de patterns liés aux neurones détectant la pression et les vibrations au niveau de la peau. Ces séquences diffèrent de celles que vous auriez obtenues si la main avait été enfoncée dans du sable, dans des billes de polystyrène expansé ou dans des feuilles mortes. Dès la flexion des doigts, le grattement et le roulement des petits galets produisent la séquence de patterns révélateurs du gravier et déclenchent la mémoire appropriée dans le cortex somatosensoriel.

La prochaine fois que vous sortirez de la douche, observez comment vous vous séchez avec la serviette. Je me suis aperçu que je m'y prends toujours de la même manière. Au cours d'une bien agréable expérience, j'ai remarqué que mon épouse aussi répète des patterns presque identiques après la douche. C'est sans doute pareil pour vous. Essayez de changer votre habitude : vous y parviendrez certainement, mais vous devrez vous concentrer. Vous découvrirez aussi l'effet des patterns en observant vos comportements lors de vos promenades.

La totalité des mémorisations réside dans les connexions synaptiques qui se déroulent entre les neurones. Etant donné l'énorme quantité d'informations stockées dans le cortex, et parce qu'il est à tout moment possible de n'en utiliser qu'une minuscule partie, il va de soi que seul un nombre limité de synapses et de neurones joue un rôle actif dans le processus de

remémoration, à un moment donné. Quand vous vous souvenez de tout ce qu'il y a chez vous, un seul ensemble de neurones est excité, qui tend ensuite à exciter un autre ensemble de neurones, et ainsi de suite. La capacité de mémorisation du néocortex d'un adulte est immense. Mais bien que nous ayons mémorisé une très grande quantité d'informations, nous pouvons ne nous en remémorer qu'une toute petite partie, et seulement par une succession d'associations.

Voici un exercice amusant. Essayez de vous souvenir de détails de votre passé, d'endroits où vous avez vécu, de lieux que vous avez visités et de gens que vous avez connus. Il est toujours possible de redécouvrir des événements, des choses ou des personnes auxquels vous n'avez plus pensé depuis de nombreuses années. Des milliers de petits détails rarement évoqués sont ainsi stockés dans les synapses. A tout moment, nous ne nous souvenons que d'infimes parties de tout ce que nous savons. La plupart des informations attendent passivement qu'une sollicitation appropriée les révèle.

La mémoire informatique ne stocke habituellement pas des successions de patterns. Il est cependant possible de le faire grâce à divers logiciels. C'est le cas lorsque vous stockez de la musique dans l'ordinateur. Mais l'ordinateur ne le fait pas automatiquement. En revanche, le cortex stocke spontanément des séquences ; c'est là un aspect inhérent au système mnémonique cortical.

---

Examinons à présent la deuxième caractéristique importante de notre mémoire : sa nature auto-associative. Comme nous l'avons découvert au Chapitre 2, ce terme signifie tout bonnement que des patterns sont associés à eux-mêmes. Une mémoire auto-associative est capable de se souvenir de la totalité d'un pattern même si un fragment seulement ou une partie déformée lui est présenté. Ceci est vrai pour les patterns spatiaux et pour les patterns temporels. Si vous apercevez les chaussures d'un de vos

enfants qui dépassent de la tenture, vous imaginez aussitôt le personnage en entier. Vous complétez le pattern spatial à partir d'un fragment de ce pattern. Ou alors, imaginez une personne qui attend l'autobus, que vous ne voyez qu'en partie car elle est masquée par un buisson. Votre cerveau n'en est pas pour autant troublé. Vos yeux ne voient qu'une partie du personnage, mais le cerveau le complète, produisant une perception de la globalité de la personne si intense que vous ne vous rendez peut-être pas compte qu'il ne s'agit que d'une déduction.

Les patterns temporels sont eux aussi complétés. Quand vous vous souvenez d'un petit détail de ce qui s'est passé il y a longtemps, la séquence d'événements entière vous revient à l'esprit. Dans le roman de Marcel Proust, *A la recherche du temps perdu*, la saveur d'une madeleine suscite un afflux de souvenirs (« l'édifice immense du souvenir ») que l'auteur développe ensuite sur des centaines de pages. Lors d'une conversation dans un environnement bruyant, il nous est parfois impossible d'entendre tous les mots. Le cerveau complète alors les manques par ce que nous nous attendions à entendre. Il a été établi que nous n'entendons pas vraiment les mots que nous percevons. Certaines personnes complètent à haute voix la phrase dite par quelqu'un d'autre, mais chacun de nous en fait autant mentalement, et pas seulement à la fin des phrases, mais aussi au milieu et au début. Le plus souvent, nous n'avons pas du tout conscience que nous complétons constamment des patterns, mais c'est une caractéristique omniprésente et fondamentale de la mémoire telle qu'elle est stockée dans le cortex. A tout moment, un élément peut activer le tout. C'est l'essence même de la mémoire auto-associative.

Le néocortex est une complexe mémoire auto-associative biologique. A chaque moment, en veille, chaque région fonctionnelle est essentiellement dans l'attente vigilante de l'arrivée de patterns familiers ou de fragments de patterns. Vous pouvez être très absorbé par une tâche, mais au moment où un proche arrive, vos pensées vont aussitôt vers lui. Cette redirection des pensées

n'est pas forcément volontaire. La seule apparence du visiteur oblige votre cerveau à se souvenir des patterns qui lui sont associés. C'est inévitable. Après une interruption, il nous arrive fréquemment de nous demander : « Où en étais-je ? » Une conversation lors d'un dîner se déroule souvent par associations. Elle commencera peut-être par une appréciation des mets. La salade évoquera celle qu'avait faite votre mère à votre mariage, ce qui fera penser au mariage de quelqu'un d'autre, puis à l'endroit où ils ont passé leur lune de miel et les actuels problèmes politiques dans cette partie du monde, et ainsi de suite. Pensées et mémoire sont associativement liées, et là encore, les pensées ne sont jamais vraiment aléatoires. Les inputs reçus par le cerveau sont associativement liés à eux-mêmes, accaparant le présent, et associativement liés à ce qui suivra normalement. C'est cet enchaînement mnémonique que nous appelons « pensée », et bien que son cheminement ne soit pas déterministe, nous ne le contrôlons pas totalement.

---

Nous pouvons maintenant examiner la troisième caractéristique majeure de la mémoire néocorticale, à savoir comment elle forme ce que nous appelons des représentations invariantes. Les bases de cette notion seront évoquées dans ce chapitre, et au Chapitre 6, nous verrons en détail comment le cortex les génère.

Une mémoire informatique est conçue pour stocker des données exactement comme elles se présentent. Lorsque vous copiez un logiciel d'un CD-ROM vers le disque dur, chaque octet (un ensemble de huit valeurs binaires 0 ou 1 appelé « bit ») est recopié avec une absolue fidélité. Une seule erreur ou divergence entre l'original et sa copie risque d'entraîner un dysfonctionnement du logiciel. La mémoire qui réside dans notre néocortex est différente. Le cerveau ne mémorise pas avec exactitude ce que nous voyons, entendons ou ressentons. Nous ne nous rappelons ni ne nous souvenons avec une totale fidélité, non parce que le cortex et

ses neurones manquent de rigueur ou sont enclins aux erreurs, mais parce que le cerveau se souvient des relations importantes qui s'établissent, et non des détails. Quelques exemples illustreront ce point.

Comme nous l'avons vu au Chapitre 2, des modèles de mémoire auto-associative simples existent depuis des décennies, et ainsi que je l'ai décrit précédemment, le cerveau se souvient d'une manière auto-associative. Il existe cependant une grande différence entre la mémoire auto-associative élaborée par les spécialistes des réseaux neuronaux et celle présente dans le cortex. Les mémoires auto-associatives artificielles ne font pas appel à des représentations invariantes ; c'est pourquoi elles échouent en certains domaines très élémentaires. Imaginez l'image d'un visage composée d'un ensemble de points noirs et blancs. C'est un pattern, et si je possède une mémoire auto-associative artificielle, je peux y stocker de nombreuses images de visages du même genre. Cette mémoire auto-associative artificielle est sophistiquée dans la mesure où, en ne lui présentant que la moitié d'un visage ou une paire d'yeux, elle les reconnaît et complète correctement le portrait. Cette expérience a été faite plusieurs fois. Mais si je décale chaque point de l'image de quelques pixels vers la gauche, la mémoire est incapable de reconnaître le visage. Pour la mémoire auto-associative artificielle, c'est à présent un pattern complètement nouveau car aucun des pixels entre le pattern précédemment stocké et le nouveau ne concorde. Mais vous et moi, nous n'aurons aucune peine à reconnaître le même visage dans le pattern décalé. Peut-être même ne remarquerons-nous pas la modification. Une mémoire auto-associative artificielle est incapable de reconnaître des patterns s'ils ont été déplacés, pivotés, mis à une autre échelle ou transformés de quelque manière que ce soit, alors que notre cerveau s'accommode très facilement de ces variations. Comment pouvons-nous percevoir qu'une chose est la même ou qu'elle est constante alors que les patterns d'entrée qui la représentent sont nouveaux ou ont changé ? Prenons un autre exemple.

Vous tenez probablement un livre entre vos mains en ce moment. Quand vous le déplacez, modifiez l'éclairage ou vous repositionnez dans votre fauteuil, ou quand vous dirigez votre regard sur différentes parties de la page, le pattern lumineux qui se forme sur votre rétine change complètement. L'input visuel varie sans cesse et ne se répète jamais. Vous auriez beau regarder la page pendant une centaine d'années, le pattern s'inscrivant sur votre rétine n'aurait aucune chance d'être de nouveau exactement le même. Mais pas un instant vous n'avez le moindre doute sur le fait que vous tenez un livre et que ce livre est le même. Le pattern interne, dans votre cerveau, qui représente « ce livre » ne varie pas en dépit du flux constant des stimuli qui vous informent. C'est pourquoi nous appliquons les termes de *représentation invariante* à la représentation interne inscrite dans le cerveau.

Comme autre exemple, pensez au visage d'une amie. Vous la reconnaissez chaque fois que vous la rencontrez, et ceci en moins d'une seconde. Qu'importe la distance où elle se trouve. Lorsqu'elle est près de vous, son visage occupe une grande partie de la rétine. Si elle est loin, il ne s'inscrit que sur une très petite partie. Elle peut vous faire face, se tourner légèrement de côté ou se montrer de profil. Elle peut être souriante, faire la moue ou bâiller. Elle peut être vivement éclairée, se tenir dans l'ombre ou se trémousser dans les faisceaux de lumière d'une boîte de nuit. Son visage peut apparaître sous d'innombrables attitudes et variantes. Pour chacune, le pattern lumineux formé sur votre rétine est unique, bien que dans chaque cas vous sachiez instantanément que c'est elle que vous voyez.

Soulevons la calotte crânienne et voyons ce qui, dans le cerveau, régit cette remarquable disposition. Nous savons par des expériences qu'en surveillant l'activité des neurones dans l'aire des inputs visuels du cortex nommée V1, le pattern d'activité diffère pour chaque vision du visage regardé. Chaque fois qu'il bouge ou que nos yeux se fixent de nouveau dessus, le pattern d'activité en V1 change, un peu comme le pattern changeant sur

la rétine. Mais si nous surveillons l'activité des cellules dans l'aire de la reconnaissance des visages — une région fonctionnelle située quelques niveaux plus haut que V1, dans la hiérarchie corticale —, nous constatons qu'elle est stable. Des ensembles de cellules, dans l'aire de reconnaissance des visages, restent actifs aussi longtemps que le visage de votre amie se trouve quelque part dans votre champ de vision, ou même seulement présent à votre esprit, et cela indépendamment de sa taille, position, orientation, échelle et expression. Cette stabilité des cellules excitées est une représentation invariante.

Introspectivement, cette tâche paraît si facile que nous osons à peine parler de problème. Elle est aussi naturelle que la respiration. Elle semble insignifiante car nous n'en sommes pas conscients. Et dans un certain sens, elle l'est car notre cerveau peut la résoudre très rapidement (rappelez-vous la règle des cent étapes). Toutefois, le problème de la compréhension du mécanisme mis en œuvre par le cortex pour former des représentations invariantes reste l'un des plus grands mystères de la science. Jusqu'à quel niveau de difficulté ? Il est si compliqué que personne, même en recourant aux ordinateurs les plus puissants, n'a réussi à trouver la solution. Et ce n'est pas faute d'avoir essayé !

Les spéculations concernant ce problème remontent à longtemps, jusqu'à Platon, il y a vingt-deux siècles. Il s'était interrogé sur la pensée et la perception du monde, et avait remarqué que la réalité des choses et des idées est toujours imparfaite et toujours différente. Par exemple, bien qu'un cercle parfait puisse être conçu, vous n'en avez en réalité jamais vu un. Tous les cercles tracés sont imparfaits. Même si vous utilisez un compas de précision, le prétendu cercle est tracé d'un trait noir : où est alors la circonférence, puisque celle d'un cercle parfait est sans épaisseur ? Ou, comme autre exemple, pensez au concept de « chien ». Tous les chiens que vous avez rencontrés sont différents les uns des autres, et chaque fois que vous apercevez le même individu, vous le voyez d'une manière différente. Tous les chiens sont différents

et l'un d'eux ne peut jamais être vu deux fois d'une manière absolument identique. Mais la totalité de votre expérience des chiens s'inscrit dans le concept mental de « chien », qui est stable et identique pour toute la gente canine. Ceci rendit Platon perplexe. Comment nous était-il possible d'acquérir ce concept et l'appliquer dans ce monde d'une infinie variété de formes et de sensations sans cesse changeantes ?

La solution de Platon se trouve dans sa célèbre *théorie des Formes*. Il conclut que nos pensées les plus élevées doivent être rattachées à quelque plan de supra-réalité transcendant où les idées immuables, stables (les Formes) existent dans leur éternelle perfection. Notre âme proviendrait de ce lieu mythique avant notre naissance ; c'est de là que nous viendrait notre connaissance des Formes. Après la naissance, nous en conserverions une connaissance latente. L'apprentissage et la cognition seraient possibles parce que les formes du monde réel nous rappelleraient les Formes correspondantes. Vous savez ce que sont le cercle et les chiens car ces notions déclencheraient la mémoire d'âme du Cercle et du Chien.

Tout ceci paraît aujourd'hui fantaisiste. Mais si vous prenez vos distances avec cette métaphysique ampoulée, vous constaterez que Platon parlait en réalité d'invariance. Ses explications étaient largement à côté de la plaque, mais pas son intuition, qui suscita l'une des questions les plus importantes que nous pouvons nous poser sur notre propre nature.

---

Pour que vous ne pensiez surtout pas que l'invariance ne se rapporte qu'à la vision, prenons un exemple impliquant d'autres sens, le toucher en l'occurrence. Quand vous introduisez la main dans la boîte à gants de votre voiture à la recherche des lunettes de soleil, vos doigts doivent les tâter pour s'assurer que vous les avez trouvées. Peu importe la partie de la main qui entre en contact ; il peut s'agir du pouce, de n'importe quelle partie d'un doigt ou de la paume. Le contact peut se produire avec n'importe quel élément

des lunettes : un verre, les plaques nasales, une charnière ou une partie de la monture. Mouvoir n'importe quelle partie de la main pendant une seconde sur n'importe quelle partie des lunettes est suffisant pour permettre au cerveau de les identifier. Dans chaque cas de tâtonnement, le flux de patterns spatiaux et temporels provenant des récepteurs tactiles est totalement différent — partie différente de l'objet, partie différente de la peau… — ce qui permet de saisir les lunettes sans même y penser.

Etudions la tâche sensori-motrice qui consiste à introduire la clé dans le Neiman de la voiture. La position du siège, du corps, du bras et de la main est chaque fois toujours légèrement différente. Pour vous, l'action est répétitive, mais c'est parce que vous en avez une représentation invariante dans votre cerveau. Si vous deviez fabriquer un bras robotisé devant entrer dans la voiture et introduire la clé, vous constateriez rapidement que c'est presque impossible à réaliser, à moins de faire en sorte que le robot ne soit toujours exactement dans la même position, et tienne chaque fois la clé de la même façon. Et même si vous y parveniez, le robot devrait être reprogrammé en fonction de chaque type de voiture. A l'instar des mémoires auto-associatives artificielles, les robots et les programmes informatiques sont nuls lorsqu'il s'agit de gérer des variations.

La signature est un autre exemple intéressant. Quelque part dans votre cortex moteur, dans le lobe frontal, se trouve une représentation invariante de votre autographe. Chaque fois que vous signez de votre nom, vous utilisez la même succession de tracés, d'angles et de rythmes. Ceci est vrai, que votre signature soit toute ratatinée parce que la place manque ou qu'elle s'étale voluptueusement sur tout le papier, que vous signiez assis à une table ou à la volée sur une feuille maintenue contre un mur. La signature sera chaque fois un peu différente, notamment lorsque vous la faites dans des conditions inconfortables. Mais quels que soient son échelle, la façon d'écrire ou le mouvement du bras, vous exécutez toujours le même « programme moteur » abstrait pour la tracer.

La signature montre que la représentation invariante dans le cortex moteur est, à certains égards, le reflet de la représentation invariante siégeant dans le cortex. Du côté sensoriel, une grande variété de patterns d'entrée peut activer un assemblage de cellules stables représentant un pattern abstrait (le visage de l'amie, les lunettes de soleil…). Du côté moteur, un assemblage de cellules stables représentant une commande motrice abstraite (attraper un ballon, signer…) est capable de s'exprimer par une grande variété de groupes musculaires, en respectant une grande variété d'autres contraintes. C'est à cette symétrie entre la perception et l'action que nous devons nous attendre si, comme le conjecturait Mountcastle, le cortex exécute un seul et même algorithme élémentaire dans toutes les aires.

Pour notre dernier exemple, revenons au cortex sensoriel et à la musique, car elle permet d'évoquer facilement tous les problèmes que le néocortex doit résoudre. En musique, la représentation invariante est illustrée par notre capacité à reconnaître une mélodie quelle qu'en soit la clé. La clé est la position d'une note sur l'échelle musicale ; une même mélodie jouée sur différentes clés commence par une note différente. En choisissant une clé pour l'interprétation, vous déterminez le restant des notes du morceau. Toute mélodie peut être jouée dans n'importe quelle clé. Cela signifie que chaque interprétation de la « même » mélodie dans une autre clé est en réalité une séquence de notes totalement différente. Chaque interprétation excite une partie différente de la cochlée, produisant l'envoi, vers le cortex auditif, d'un ensemble de patterns spatio-temporels entièrement différent, et ceci bien que dans chaque cas vous perceviez la même mélodie.

Prenons la chanson « Somewhere over the Rainbow » que chante Judy Garland dans le film *Le Magicien d'Oz*, et qu'elle interprète en la bémol. Si je me mets au piano et que je commence à la jouer avec une clé que vous n'avez jamais entendue — disons, de ré —, vous entendrez la même chanson. Vous ne

remarquerez pas que toutes les notes sont différentes de la version qui vous est familière. Cela signifie que votre mémoire de la mélodie doit être dans une forme qui ignore la hauteur de ton. La mémoire doit stocker d'importantes relations à l'intérieur de la chanson, et non les véritables notes. Dans ce cas, ces relations importantes sont les hauteurs relatives des notes, ou intervalles. « Somewhere over the Rainbow » commence à l'octave d'au-dessus, suivi d'un demi-ton en dessous, puis d'une tierce majeure en dessous, et ainsi de suite. Les intervalles entre les notes restent les mêmes quelle que soit la clé utilisée pour l'interprétation. Votre capacité à reconnaître facilement la mélodie quelle que soit la clé indique que votre cerveau a stocké la chanson dans une forme à hauteurs de notes invariantes.

De même, la mémoire du visage de votre amie doit elle aussi être stockée sous une forme indépendante de toute vision particulière. Ce qui rend le visage reconnaissable sont ses mensurations relatives, ses couleurs relatives, ses proportions relatives et non son apparence à un instant donné mardi dernier au déjeuner. Il existe des « intervalles spatiaux » entre les caractéristiques de son visage tout comme il existe des « intervalles de hauteur de notes » entre les notes d'une chanson. Les traits de son visage sont largement définis par rapport aux yeux. Son nez est court comparé à l'écartement de ses yeux. La couleur de ses cheveux et celle de ses yeux entretiennent une relation relative identique qui demeure constante, même si sous différentes conditions d'éclairage les couleurs absolues varient significativement. Quand vous mémorisez un visage, vous mémorisez ses attributs relatifs.

Je pense qu'une telle abstraction de forme se produit sur tout le cortex, dans toutes les régions. C'est une propriété générale du néocortex. Les mémoires sont stockées sous une forme qui s'approprie l'essence des relations, et non les détails du moment. Quand vous voyez, entendez ou ressentez, le cortex prend les inputs détaillés, hautement spécifiques et les convertit en leur forme invariante. C'est cette dernière qui est stockée en mémoire,

et c'est à la forme invariante de chaque pattern d'entrée qu'elle est comparée. Le stockage, le rappel et la reconnaissance de la mémoire se produisent tous au niveau des formes invariantes. Il n'existe aucun concept équivalent en informatique.

---

Ceci soulève un intéressant problème. Au prochain chapitre, je soutiens qu'une importante fonction du néocortex est le recours à sa mémoire pour faire des prédictions. Mais étant donné que le cortex stocke des formes invariantes, comment peut-il se lancer dans des prédictions spécifiques ? Voici quelques exemples qui illustrent le problème et sa solution.

Imaginez-vous en 1890, dans une ville-frontière de l'Ouest américain. L'élue de votre cœur a pris le train sur la côte Est afin de vous rejoindre dans votre nouveau foyer frontalier. Bien sûr, vous tenez à l'accueillir à la gare. Quelques semaines avant son arrivée, vous vous renseignez sur la circulation des trains. Il n'y a pas d'horaire précis et, pour autant que vous sachiez, les trains ne sont jamais arrivés et partis aux mêmes heures, d'un jour à un autre. Vous commencez à vous rendre compte que vous ne pouvez pas prédire quand le train arrivera. Mais vous remarquez une structure dans la circulation. Le train en provenance de la côte Est arrive quatre heures après le départ de celui qui retourne vers l'est. Cet intervalle de quatre heures est régulier jour après jour, bien que l'heure elle-même varie considérablement. A la date de l'arrivée de l'élue, vous surveillez le train en partance vers l'est et, dès qu'il s'en va, vous réglez votre montre. Quatre heures plus tard, vous êtes à la gare au moment même où le train tant attendu arrive à quai. Cette histoire illustre à la fois le problème posé au néocortex et la solution qu'il trouve.

Le monde que perçoivent nos sens n'est jamais le même. A l'instar des heures de départ et d'arrivée des trains de notre histoire, il est toujours différent. La manière dont vous appréhendez le monde repose sur la découverte de structures invariantes dans

le flux constamment changeant des inputs. Cependant, la structure invariante n'est pas à elle seule suffisante pour constituer une base solide pour des prédictions spécifiques. Savoir seulement qu'un train arrive quatre heures après un départ n'est pas suffisant pour vous permettre d'être sur le quai à temps pour accueillir votre élue. Pour établir une prédiction spécifique, le cerveau doit associer la connaissance de la structure invariante aux détails les plus récents. La prédiction de l'arrivée du train exige la connaissance de la structure des quatre heures des mouvements de trains, associée à la connaissance plus détaillée de l'heure de départ du train vers l'est.

Quand vous écoutez une musique familière jouée au piano, votre cortex prédit la prochaine note avant qu'elle ne soit jouée. Mais la mémoire du morceau, comme nous l'avons vu, est une forme à hauteur de note invariante. Votre mémoire vous apprend ce qu'est le prochain intervalle mais ne révèle rien, par et d'elle-même, sur la note réelle. La prédiction de l'exacte prochaine note exige d'associer le prochain intervalle avec la dernière note spécifique. Si le prochain intervalle est une tierce majeure et si la dernière note entendue était un do, vous pouvez prédire la note suivante : un mi. Vous entendez un mi dans votre esprit, et non « une tierce majeure ». Et, à moins que vous n'ayez mal entendu ou que le pianiste ait fait une fausse note, votre prédiction sera correcte.

Quand vous apercevez le visage de votre amie, votre cortex le complète et prédit la myriade de détails de son unique image à cet instant précis. Il vérifie si les yeux concordent, si le nez, les lèvres et les cheveux sont exactement comme ils doivent être. Le cortex établit ces prédictions avec une grande spécificité. Il est capable de prédire des détails faciaux de niveau inférieur, même si la personne n'a jamais été vue sous cet angle ou cet environnement particuliers auparavant. Si vous savez exactement comment les yeux et le nez de votre amie sont agencés, et si vous connaissez la structure de son visage, alors vous pouvez prédire avec précision

où doivent se trouver ses lèvres. Si vous savez que sa peau doit être teintée en orange par la lumière du soir, vous savez par conséquent de quelle couleur doivent être ses cheveux. Une fois de plus, le cerveau accomplit tout cela en associant la mémoire de la structure invariante de son visage aux particularités de votre expérience immédiate.

L'exemple des horaires de trains n'est qu'une analogie révélant ce qui se déroule dans le cortex, mais ce n'est pas le cas des exemples de la musique et du visage. L'association d'une représentation invariante aux inputs courants afin de produire des prédictions détaillées est bel et bien ce qui se produit. C'est un processus omniprésent qui a lieu dans toutes les régions du cortex. C'est ainsi que vous procédez à des prédictions spécifiques portant sur la pièce où vous vous trouvez actuellement. C'est ainsi que vous pouvez prédire non seulement ce que diront les autres, mais aussi le ton sur lequel ils s'exprimeront, leur accent et d'où, dans la chambre, devrait provenir la voix. C'est ainsi que vous savez précisément quand votre pied se posera sur le sol, et ce qu'il éprouvera en montant les escaliers. C'est ainsi que vous pourrez signer de votre nom, même avec votre pied, ou attraper le ballon qui vous a été lancé.

Les trois propriétés de la mémoire corticale exposées dans ce chapitre — le stockage de séquences, le souvenir auto-associatif et les représentations invariantes — sont les éléments indispensables pour prédire d'après la mémoire du passé. Au prochain chapitre, je développe l'idée que procéder à des prédictions est l'essence de l'intelligence.

# 5

# UNE NOUVELLE STRUCTURE DE L'INTELLIGENCE

**Un jour** d'avril 1986, je réfléchissais sur ce que signifie « comprendre » quelque chose. Pendant des mois, j'avais été taraudé par la question fondamentale de savoir ce que fait le cerveau s'il ne produit pas du comportement. Que fait-il lors de l'écoute passive d'un discours ? Que fait-il maintenant pendant que je lis ? L'information entre dans le cerveau mais n'en sort pas : que devient-elle ? Votre comportement à ce moment précis est sans doute élémentaire — respiration et mouvements oculaires — mais, comme vous en êtes conscient, le cerveau en fait certainement plus tandis que vous lisez et comprenez ces lignes. Comprendre doit être le résultat d'une activité neuronale. Mais qu'en est-il ? Que font les neurones lorsque nous comprenons ?

Je contemplais mon bureau, ce jour-là, et vis des chaises, des affiches, la fenêtre, des plantes, des crayons, etc. Des centaines de choses et d'objets m'entouraient. Mes yeux les parcouraient, mais me contenter de les regarder ne me faisait

effectuer aucune action. Aucun comportement n'était ni suscité ni requis, bien que je « comprisse » la chambre et son contenu. Je faisais ce qui ne pouvait être fait dans la Chambre chinoise de John Searle et je n'avais rien à glisser par une ouverture. Je comprenais, mais aucune action n'en apportait la preuve. Que pouvait bien signifier « comprendre » ?

C'est pendant que je méditais sur ce dilemme que j'eus une illumination, l'un de ces moments rares où soudainement tout ce qui n'était que confusion devient tout à coup clair et limpide. Ce qui la déclencha fut de me demander ce qui se passerait si un nouvel objet que je n'avais jamais vu auparavant — une tasse à café bleue, par exemple — apparaissait dans la pièce.

La réponse semblait simple : je remarquerais le nouvel objet en tant qu'objet ne m'appartenant pas. Il attirerait mon attention de par sa nouveauté. Je n'aurais pas à me demander consciemment si la tasse est nouvelle. Il s'avérerait seulement qu'elle n'est pas à moi. Sous cette réponse apparemment anodine se cache un puissant concept. Pour remarquer que quelque chose est différent, certains neurones de mon cerveau qui n'étaient jusque-là pas actifs commencent à le devenir. Comment ces neurones savent-ils que la tasse est nouvelle et que les centaines d'autres objets du bureau ne le sont pas ? La réponse à cette question me surprend encore : notre cerveau fait appel aux mémorisations qu'il a stockées pour procéder à des prédictions continuelles concernant tout ce que nous voyons, entendons et ressentons. Quand je contemple mon bureau, mon cerveau fait appel à des mémoires pour élaborer des prédictions sur ce à quoi je m'attends. La grande majorité de ces prédictions est inconsciente. C'est comme si des parties de mon cerveau se demandaient : « Cet ordinateur est bien au milieu du bureau ? Oui, il l'est. Est-il noir ? Oui. La lampe est bien dans le coin droit ? Elle l'est. Le dictionnaire est là où je l'avais laissé ? Oui. La fenêtre est rectangulaire et les murs verticaux ? En effet. La lumière du soleil parvient-elle de la bonne direction selon le jour et l'heure ? Oui. » Mais si un pattern visuel surgit, qui n'a pas été

mémorisé dans ce contexte, la prédiction est erronée. Mon attention se porte alors sur cette erreur.

Bien sûr, le cerveau ne se parle pas à lui-même quand il fait des prédictions, et il ne les fait pas d'une manière sérielle. Il ne fait pas non plus des prédictions uniquement au sujet des objets distincts comme la tasse à café. Il procède en fait sans relâche à des prédictions sur tout ce qui constitue le monde qui nous entoure, en parallèle. Il détectera tout aussi bien une texture curieuse, un nez déformé ou un mouvement inhabituel. Il n'est pas évident de se rendre compte combien ces prédictions pour la plupart apparentes sont pénétrantes, et c'est sans doute pourquoi leur importance nous a si longtemps échappé. Elles adviennent si spontanément, si facilement, que ce qui se produit dans notre crâne nous échappe. J'espère que vous mesurez la puissance de cette idée. La prédiction est si fine qu'il s'avère que ce que nous percevons — c'est-à-dire comment le monde nous apparaît — ne provient pas de nos seuls sens. Ce que nous percevons est une combinaison de ce que nous ressentons et des prédictions faites par le cerveau en se fondant sur la mémoire.

---

Je conçus aussitôt une expérience sur la pensée permettant de comprendre ce que je venais de découvrir. Je l'ai appelée « la porte faussée ». La voici.

Quand vous rentrez chez vous, il ne vous faut généralement que quelques instants pour franchir la porte d'entrée, quel qu'en soit le modèle. Vous l'atteignez, vous actionnez la poignée, entrez et fermez la porte derrière vous. C'est une solide habitude, une action que vous effectuez tout le temps et à laquelle vous n'accordez qu'une minime attention. Supposons qu'en votre absence je m'introduise chez vous et modifie la porte. Par exemple, je remonterais le bouton de deux ou trois centimètres, ou alors je remplacerais le bouton de la poignée par une clenche ou inversement, ou je remplacerais la serrure en laiton par une autre

chromée. Je pourrais aussi modifier le poids de la porte, remplacer le panneau en chêne par du verre cathédrale ou inversement. Je pourrais encore rendre les charnières dures et grinçantes ou les lubrifier, ou encore élargir ou réduire la largeur de la porte et de son bâti. Je pourrais changer sa couleur, remplacer le judas optique par un heurtoir ou découper une petite ouverture grillagée. Je pourrais imaginer des milliers de modifications susceptibles de vous rendre la porte complètement étrangère. A votre retour, en l'ouvrant, vous remarqueriez aussitôt que quelque chose cloche. Il vous faudra peut-être un petit moment pour réaliser exactement ce qui ne va pas, mais la modification, vous la percevrez aussitôt. Lorsque votre main se portera vers le bouton, vous vous rendrez compte qu'il n'est plus au même endroit. Ou si vous apercevez la petite ouverture grillagée, cela vous paraîtra bizarre. Si le poids de la porte a été changé, vous serez surpris par la nature du retour d'effort, lorsque vous la pousserez. Bref, quelle que soit la modification, vous ne serez pas long à la remarquer.

Il n'y a qu'un seul moyen d'interpréter votre réaction à la porte faussée : votre cerveau procède à des prédictions sensorielles de bas niveau sur ce qu'il s'attend à voir, entendre et ressentir à n'importe quel moment, et il fait ces prédictions en parallèle. Toutes les régions du néocortex tentent simultanément de prédire ce que sera la prochaine expérience. Les aires visuelles font des prédictions sur des droites, des formes, des objets, leur emplacement et leurs mouvements. Les aires auditives font des prédictions sur la tonalité, la provenance du son et les patterns sonores. Les aires somatosensorielles font des prédictions sur le toucher, la texture, les contours et la température.

Le terme « prédiction » signifie que les neurones impliqués dans la perception de votre porte deviennent actifs à l'avance, anticipant le moment où ils reçoivent véritablement les inputs sensoriels. Quand ces derniers arrivent, ils sont comparés à ce qui est attendu. Quand vous vous approchez de la porte, votre cortex produit un ensemble de prédictions fondées sur l'expérience

passée. Quand vous l'atteignez, il prédit ce que ressentiront vos doigts, et sous quel angle ils toucheront la poignée. Quand vous vous apprêtez à pousser la porte, le cortex prédit la résistance qu'elle offrira et le bruit qu'elle fera. Lorsque toutes les prédictions se sont réalisées, vous franchissez le seuil sans même vous rendre compte de ces supputations. Mais si l'une d'elles n'est pas vérifiée, l'erreur produite attirera votre attention. Les prédictions correctes produisent de l'entendement : la porte est normale. Des prédictions erronées produisent de la confusion et vous incitent à faire attention : le verrou n'est pas là où il est censé être, la porte pivote trop facilement, la porte est décentrée, la texture de la poignée a changé… Nous procédons sans cesse à des prédictions de bas niveau, en parallèle sur tous nos sens.

Mais ce n'est pas tout. A mon avis, la prédiction n'est pas qu'une activité du cerveau parmi d'autres. C'est *la fonction principale* du néocortex, le fondement de l'intelligence. Le cortex est un organe voué à la prédiction. Pour comprendre ce que sont l'intelligence et la créativité, pour comprendre comment fonctionne le cerveau et comment construire des machines intelligentes, nous devons comprendre la nature de ces prédictions et aussi comment le cortex les génère. Même les comportements sont mieux appréhendés si nous les considérons comme des sous-produits des prédictions.

---

J'ignore qui fut le premier à suggérer que la prédiction est la clé qui permet de comprendre l'intelligence. Dans les domaines scientifique et industriel, personne n'invente quoi que ce soit de complètement nouveau. Les chercheurs et inventeurs se fondent en réalité sur l'existant pour découvrir de nouvelles structures. Les éléments d'une nouvelle idée sont généralement déjà présents dans le discours des milieux scientifiques avant qu'elle ne soit découverte. Ce qui est généralement nouveau parmi ces éléments est leur réunion en un tout cohésif. De même, l'idée que la

fonction principale du cortex est de procéder à des prédictions n'est pas entièrement nouvelle. Elle était depuis longtemps dans l'air sous diverses formes. Mais elle n'avait pas encore été placée correctement au cœur d'une théorie du cerveau et d'une définition de l'intelligence.

Des pionniers de l'intelligence artificielle avaient envisagé des ordinateurs élaborant un modèle du monde qui serait utilisé pour faire des prédictions. En 1956, D. M. Mackay affirma que les machines intelligentes devraient être dotées d'un « mécanisme de réponse interne » conçu pour « s'accorder à ce qui serait reçu ». Il n'avait pas utilisé les mots « mémoire » et « prédiction », mais ce à quoi il pensait était du même acabit.

Dès le milieu des années 1990, des termes comme *inférence*, *modèles génératifs* et *prédiction* s'étaient introduits dans le langage scientifique. Tous se rapportent à des idées en relation les unes avec les autres. Par exemple, dans un ouvrage publié en 2001, *i of the vortex*, Rodolfo Llinás, neurophysiologiste à l'Ecole de médecine de New York, écrivait : « La capacité de prédire le résultat d'événements futurs — importante pour la réussite d'un mouvement — est très probablement l'ultime et la plus commune de toutes les fonctions globales du cerveau. » Des savants comme David Mumford, de l'université Brown, Rajesh Rao, de l'université de Washington, Stephen Grossberg, de l'université de Boston, et beaucoup d'autres ont écrit et proposé des théories sur le rôle des biofeedbacks et de la prédiction. Un sous-domaine entier des mathématiques est consacré aux réseaux bayésiens (du nom de Thomas Bayes, un ministre anglais né en 1702 qui fut un pionnier des statistiques. Les réseaux bayésiens utilisent la théorie des probabilités pour faire des prédictions.)

Ce qui manquait était une structuration théorique cohérente de tous ces éléments disparates. J'affirme que cela n'a jamais été fait auparavant, et c'est le but du présent ouvrage.

---

Avant d'examiner de près comment le cortex procède à des prédictions, nous examinerons quelques autres exemples. Plus vous réfléchissez à cette idée et plus vous vous rendez compte que la notion de prédiction est pénétrante et qu'elle est à la base de la compréhension du monde que nous percevons.

Ce matin, j'ai fait des crêpes. A un moment, je me suis baissé pour ouvrir la porte du buffet. Je savais intuitivement, sans rien voir, ce que je ressentirais — en l'occurrence, le contact avec le bouton de la porte — et à quel moment. Je dévissai le bouchon du pack de lait, m'attendant à ce qu'il pivote et libère l'orifice. J'allumai la plaque de fonte, m'attendant à pousser un peu fort sur le bouton et le tourner en rencontrant une légère résistance. Je m'attendais à percevoir, une seconde plus tard, le bref souffle du gaz qui vient de s'allumer. A chaque minute, je faisais dans la cuisine des dizaines voire des centaines de gestes et de mouvements, chacun impliquant de nombreuses prédictions. Je le savais car si le moindre de ces gestes habituels avait produit un résultat différent de celui escompté je l'aurais remarqué.

Chaque fois que vous mettez un pied en avant lorsque vous marchez, votre cerveau prédit le moment où il se posera ainsi que la perception du revêtement, juste dessous. Quand vous manquez une marche d'un escalier, vous vous apercevez immédiatement que quelque chose ne va pas : le pied s'abaisse, et quand il semble passer à travers la marche manquée, cela vous trouble. Le pied ne perçoit rien, mais votre cerveau avait fait une prédiction, et cette prédiction ne s'est pas vérifiée. Un robot informatisé trébucherait en beauté sans même réaliser que quelque chose ne va pas, alors que vous, vous prenez conscience de l'incident même si le pied ne parcourt qu'un centimètre de trop sous le plan où le cerveau avait estimé qu'il devait se poser.

Quand vous écoutez une mélodie familière, vous entendez la prochaine note dans votre tête avant même qu'elle ne retentisse. Quand vous écoutez votre album favori, vous connaissez la prochaine chanson une ou deux secondes avant qu'elle ne commence.

Que se passe-t-il ? Les neurones qui devraient être excités au moment où vous entendez la prochaine note le sont en réalité à l'avance ; c'est pourquoi vous entendez la note à venir dans votre tête. Les neurones réagissent en réponse à la mémoire. Cette mémoire peut s'avérer étonnamment durable. Il n'est pas rare qu'en réécoutant un album des années après l'avoir oublié dans un meuble, vous entendiez déjà le morceau à venir au moment où la chanson se termine. La fonction d'écoute aléatoire d'un CD suscite une plaisante sensation de douce incertitude, car vous savez que la prédiction du prochain morceau sera fausse.

Quand vous écoutez des gens qui parlent, vous savez souvent ce qu'ils vont dire avant qu'ils aient fini, ou du moins vous pensez le savoir. Parfois nous n'écoutons pas même ce qui est véritablement dit et entendons ce que nous nous attendons à entendre. Ceci m'arriva si souvent dans mon enfance que cela inquiéta ma mère qui me fit examiner deux fois par un oto-rhino-laryngologiste. Ce phénomène d'anticipation se produit parce que les gens ont tendance à utiliser des formules toutes faites dans leur conversation. Par exemple, si je dis « quand il reviendra le temps des », votre cerveau active les neurones représentant le mot « cerises » avant même que j'aie prononcé le mot (mais cet exemple n'est valide que pour des gens connaissant la chanson de Jean-Baptiste Clément). Bien sûr, nous ne savons pas tout le temps ce que les autres vont dire. La prédiction n'est jamais infaillible. A vrai dire, notre intelligence procède à des prédictions probabilistes. Parfois, nous savons exactement ce qui va se produire, parfois nos attentes envisagent plusieurs possibilités. Si au cours d'un repas, je dis « Passe-moi… », votre cerveau ne sera pas surpris si je dis « le sel », « le poivre » ou « la moutarde ». D'une certaine manière, il prédit simultanément toutes ces possibilités. Mais si je dis « Passe-moi le trottoir », vous vous rendrez bien compte que quelque chose sonne faux.

Revenons à la musique. Les prédictions probabilistes s'y manifestent aussi. Quand vous écoutez une chanson que vous

n'avez jamais entendue, vous pouvez cependant avoir de fortes attentes. Si c'est de la musique occidentale, je m'attends à un rythme, à des phrases rimées s'étendant sur un même nombre de mesures, et à ce que la musique se termine par une note tonique, ou finale. Même si vous ne connaissez pas ces subtilités techniques — mais que vous avez entendu de la musique du même genre —, votre cerveau prédit spontanément le battement, les rythmes répétés, la fin des phrases et la fin du morceau. Si un morceau enfreint ces principes, vous savez immédiatement que quelque chose sonne faux. Réfléchissez-y : vous entendez une chanson nouvelle, votre cerveau prend connaissance d'un pattern qui ne lui a jamais été soumis auparavant, et il vous est néanmoins possible de faire des prédictions et repérer ce qui est faux. La base de ces prédictions pour la plupart inconscientes est l'ensemble d'éléments mémorisés stockés dans notre cortex. Votre cerveau n'est pas capable de dire exactement ce qui suivra, mais il parvient toutefois à prédire quels patterns de notes sont censés se produire, et lesquels ne le sont pas.

Nous avons tous un jour remarqué soudainement qu'un bruit constant — un marteau-piqueur au loin, une douce musique d'ambiance… — vient juste de cesser, bien que nous ne l'ayons pas du tout noté pendant qu'il se manifestait. Nos aires auditives prédisaient la continuation, moment après moment, et aussi longtemps que le bruit ne variait pas, nous n'y prêtions aucune attention. Dès que le bruit cesse, la prédiction est contredite et notre attention mise en éveil. Voici un exemple historique : quand la municipalité de New York avait interrompu le métro aérien, des gens appelèrent la police en pleine nuit parce que quelque chose les avait réveillés. Or, ils appelaient surtout à l'heure où les rames passaient habituellement près de leur appartement.

Un adage dit que « voir, c'est croire ». Pourtant, nous voyons ce que nous nous attendons à voir autant que nous voyons ce qui est véritablement. L'un des exemples les plus fascinants est ce que les chercheurs appellent le « remplissage ». Vous devez pour cela

savoir qu'une tache aveugle, la papille, se trouve à la racine du nerf optique de chaque œil. Cette zone étant dépourvue de photorécepteurs, ce qui s'y inscrit ne peut être perçu. Deux raisons font que vous ne remarquez généralement pas cette zone aveugle. L'une est terre à terre, l'autre instructive. La raison terre à terre est que les deux taches ne se superposent pas lors de la vision ; un œil compense donc l'autre.

Mais, ce qui est plus intéressant est que vous ne décelez pas la tache aveugle même si l'un des yeux est fermé. Votre système visuel « remplit » la zone dépourvue d'informations. Quand vous fermez un œil et regardez les arabesques d'un tapis turc ou les veines sinueuses d'une table en merisier, vous ne remarquez aucun trou ou tache dans votre vision. Tous les nœuds du tapis et tous les nœuds du bois sont visibles, sans aucune solution de continuité. Votre cortex visuel reconstitue de mémoire des patterns identiques, suscitant ainsi un flux continu de prédictions qui remplit toute zone manquante.

Le remplissage se produit partout dans l'image perçue, et pas seulement dans la tache aveugle. Prenons une photo d'un rivage montrant du bois d'épave échoué dans des rochers. La limite entre le bois et le roc est claire et évidente. Mais si nous observons l'image à la loupe, nous constatons que là où ils se touchent, la texture et la couleur des rochers et les rondins sont identiques. Impossible de distinguer le roc du bois. Mais si nous regardons la scène dans son entier, le bord du bois est évident ; en réalité, nous avons inféré, c'est-à-dire déduit, le bord à partir du restant de l'image. Quand nous regardons autour de nous, les objets sont nettement délimités, mais les données brutes qui parviennent à nos yeux sont souvent brouillées et ambiguës. Notre cortex remplit les parties manquantes ou brouillées par ce qu'il estime devoir se trouver à ces endroits. Nous percevons ainsi une image sans ambiguïté.

La prédiction au niveau de la vision est une fonction liée à la manière dont nos yeux regardent le monde. Au Chapitre 3, j'ai

mentionné le phénomène de saccades. Environ trois fois par seconde, vos yeux se fixent à un point puis soudainement à un autre. Ces mouvements sont généralement inconscients et ne peuvent être contrôlés. Chaque fois que votre regard se fixe sur un nouveau point, le pattern que vos yeux envoient au cerveau est complètement différent du précédent. Trois fois par seconde, votre cerveau « voit » une autre image. Les saccades ne sont pas aléatoires. Quand vous regardez un visage, votre regard se fixe d'abord sur un œil, puis sur l'autre, passe de l'un à l'autre, et se porte parfois sur le nez, la bouche, les oreilles ou tout autre endroit. Vous percevez certes un visage, mais les yeux voient successivement un œil, un autre œil, le nez, la bouche, un œil, et ainsi de suite. J'admets que ce n'est pas ainsi que vous ressentez les choses. Vous avez conscience d'une vision continue du monde. Mais les données acheminées au cerveau n'en sont pas moins aussi tressautantes que l'image issue d'un Caméscope manié d'une main hésitante.

Imaginons que vous rencontriez quelqu'un affublé d'un nez supplémentaire à la place d'un œil. Votre regard fixe un œil, puis l'autre, mais au lieu de découvrir un œil, c'est un nez. Vous vous rendez assurément compte que quelque chose est faux. Pour en arriver à cette conclusion, le cerveau a dû supposer ce qu'il aurait dû voir. Lorsqu'un œil est prédit et qu'un nez apparaît, la prédiction est contredite. Plusieurs fois par seconde, au gré des saccades, le cerveau fait des prédictions sur ce qu'il va voir. Si l'une d'elles se révèle fausse, votre attention est immédiatement mise en éveil. C'est pourquoi nous avons des difficultés à nous empêcher de regarder des gens affectés d'une difformité ; si vous rencontriez un individu à deux nez, ne seriez-vous pas tenté de le dévisager ? Mais si vous viviez avec cette personne, après une période d'accoutumance, ses deux nez n'auraient plus rien de surprenant et vous ne remarqueriez même plus cette particularité.

Passons maintenant à vous-même. Quelles prédictions faites-vous ? Quand vous tournez les pages de ce livre, vous vous attendez

à ce qu'elles s'incurvent d'une manière prévisible et soient plus souples que la couverture. Si vous êtes assis, vous prédisez que la sensation de calage du corps dans le fauteuil persistera. Mais si ce dernier commence à se mouiller, part en arrière ou subit un autre changement inattendu, vous cesserez de prêter attention au livre afin de déterminer ce qui se passe. En vous observant vous-même, vous pouvez vous rendre compte que votre perception du monde est intimement liée aux prédictions. Votre cerveau a élaboré un modèle du monde qu'il vérifie constamment en le comparant à la réalité. Vous savez où vous êtes et ce que vous faites grâce à des validations de ce modèle.

Les prédictions ne se limitent pas qu'aux patterns d'informations sensorielles comme la vue et l'ouïe. Jusqu'à présent, nous avons limité la discussion à ces exemples car ce sont les plus faciles pour aborder la structure permettant de comprendre ce qu'est l'intelligence. Toutefois, selon le principe de Mountcastle, ce qui est vrai pour les aires sensorielles de bas niveau doit aussi l'être pour toutes les aires corticales. Le cerveau humain est plus intelligent que celui des animaux car il est capable de procéder à des prédictions sur des types de patterns plus abstraits, et parce que ses séquences de patterns temporels sont plus longues. Pour prédire ce que dira ma femme quand elle me verra, je dois savoir ce qu'elle a dit dans le passé, savoir que c'est aujourd'hui vendredi, que c'est le vendredi soir qu'il faut sortir la poubelle, que je ne l'avais pas fait à temps la semaine dernière et que le visage de ma femme a telle ou telle apparence. Quand elle commence à parler, j'ai une assez bonne idée de ce qu'elle va dire. En l'occurrence, je ne sais pas quels seront les termes exacts, mais je sais qu'elle me rappellera de ne pas oublier la poubelle. Le point important ici est que le processus qui régit une intelligence plus élevée n'est pas différent du processus d'une intelligence perceptuelle. Il repose fondamentalement sur la même mémoire néocorticale et le même algorithme de prédiction.

Remarquez que les tests d'intelligence sont par essence des tests de prédiction. De l'école maternelle à l'université, les tests de

QI sont fondés sur la création de prédictions. Une série de chiffres est donnée et il faut trouver le suivant. Différentes vues d'un objet complexe sont présentées et il faut choisir la suivante. Un mot A est à un mot B ce que le mot C est au mot à trouver.

La science est par elle-même un exercice de prédictions. Notre connaissance du monde progresse grâce à une méthodologie fondée sur des observations, des hypothèses et des essais conduisant à la confirmation ou à l'infirmation. Cet ouvrage est essentiellement une prédiction de ce qu'est l'intelligence et comment le cerveau fonctionne. Même l'industrie est fondamentalement prédictive. Qu'il s'agisse de vêtements ou de téléphones mobiles, les concepteurs et les ingénieurs tentent de prédire — ou plus exactement, prévoir — les demandes de la clientèle, la réaction de la concurrence, combien coûtera un nouveau modèle et ce qui sera à la mode.

L'intelligence s'évalue par la capacité à se souvenir des patterns décrivant le monde et les prédire, y compris dans les domaines du langage, des mathématiques, des propriétés physiques d'un objet ou de la situation sociale. Notre cerveau reçoit des patterns en provenance du monde extérieur, les mémorise puis procède à des prédictions en combinant ce qui a été vu auparavant avec ce qui se passe actuellement.

---

Peut-être penserez-vous : « J'accepte l'idée que mon cerveau fasse des prédictions et l'idée que je peux faire preuve d'intelligence même couché dans le noir. Comme vous l'avez souligné, il n'est pas nécessaire d'agir pour comprendre ou être intelligent. Mais ces situations ne sont-elles pas exceptionnelles ? Insinuez-vous véritablement que la cognition intelligente et le comportement sont complètement distincts ? En fin de compte, n'est-ce pas le comportement, et non la prédiction, qui nous rend intelligents ? Car après tout, le comportement n'est-il pas l'ultime garant de notre survie ? »

C'est une bonne question et bien sûr, c'est finalement du comportement que dépend avant tout la survie d'un animal. La prédiction et le comportement ne sont pas complètement séparés, mais leur relation est subtile. Premièrement, le néocortex est apparu dans l'évolution après que les animaux eurent élaboré des comportements sophistiqués. C'est pourquoi la valeur du cortex pour la survie doit d'abord être comprise en termes d'améliorations successives plaquées par-dessus les comportements existants de l'animal. Le comportement est apparu d'abord, puis l'intelligence. Deuxièmement, la plus grande partie de ce que perçoivent nos sens dépend grandement de ce que nous faisons et comment nous nous mouvons dans le monde. Par conséquent, prédictions et comportements sont étroitement liés. Examinons ces points.

Les mammifères ont développé un cortex important car il leur procurait un avantage pour la survie. Or, cet avantage doit finalement être enraciné dans le comportement. Mais au début, le cortex servait à exploiter plus efficacement les comportements existants, et non à en créer de nouveaux. Pour tirer cela au clair, nous devons étudier comment notre cerveau a évolué.

Des systèmes nerveux simples apparurent peu de temps après que des créatures multicellulaires eurent commencé à se répandre sur la terre, il y a des centaines de millions d'années, mais l'histoire de l'intelligence commence plus récemment avec nos ancêtres les reptiles. Les grands sauriens parvinrent à conquérir les terres. Ils se répandirent sur tous les continents et se diversifièrent en de nombreuses espèces. Ils possédaient des sens affûtés et un cerveau bien développé qui les avaient dotés de comportements complexes. Leurs descendants directs, les reptiles d'aujourd'hui, les ont encore. Les sens d'un alligator, par exemple, sont aussi perfectionnés que ceux de l'homme. Ses yeux, ses oreilles, son museau, sa gueule et sa peau sont bien développés. Il manifeste des comportements complexes, notamment des capacités pour nager, se déplacer rapidement à terre, se cacher, chasser, attendre sa proie, s'exposer au soleil, faire un nid et s'accoupler.

Quelle est la différence entre le cerveau humain et celui d'un reptile ? Elle est à la fois grande et petite. Petite car grossièrement, tout ce qui existe dans le cerveau du reptile existe aussi dans le cerveau de l'homme. Grande parce que le cerveau humain possède un élément véritablement important qui fait défaut au reptile : un vaste cortex. Nous avons déjà mentionné la partie du cerveau archaïque, et vous avez sans doute entendu parler de cerveau « ancien », « limbique » ou « primitif ». Chacun de nous possède dans son cerveau cette structure remontant à la nuit des temps, semblable à celle des sauriens, d'où son autre nom : le « cerveau reptilien ». Il régule la pression sanguine, contrôle la sensation de faim, les pulsions sexuelles, les émotions et beaucoup d'aspects de la gestuelle et du mouvement. Se tenir debout, maintenir l'équilibre et marcher, par exemple, dépend grandement du cerveau archaïque. Quand vous entendez un bruit ou un cri épouvantable, que vous paniquez et commencez à courir, c'est surtout parce que le cerveau archaïque vient de prendre les commandes. Un cerveau reptilien est plus que suffisant pour effectuer quantité de tâches intéressantes et utiles. A quoi sert alors le cortex s'il n'est pas strictement limité à voir, entendre et mettre en mouvement ?

Les mammifères sont plus intelligents que les reptiles à cause de leur néocortex. L'étymologie du mot « néocortex » est « nouvelle écorce », « nouvelle croûte », car il recouvre littéralement le cerveau ancien. Le néocortex est apparu il y a quelques dizaines de millions d'années et seuls les mammifères en possèdent un. Ce qui rend les humains plus intelligents que les autres mammifères est principalement la vaste surface de leur néocortex, qui s'est considérablement accrue il y a deux millions d'années. Rappelez-vous que le cortex est formé par la répétition d'un élément commun. L'épaisseur de l'enveloppe corticale de l'être humain est la même que celle du cortex des autres mammifères et sa structure est quasiment identique. Quand l'évolution fait très rapidement un progrès décisif, comme ce fut le cas du cortex humain, elle le

fait en copiant une structure déjà existante. Nous sommes devenus plus intelligents grâce à l'ajout de beaucoup d'éléments à un algorithme cortical commun. Une idée fausse très répandue veut que le cerveau humain soit l'aboutissement de milliards d'années d'évolution. C'est peut-être vrai en ce qui concerne le système nerveux dans sa totalité. En revanche, le néocortex humain est une structure relativement récente qui existe depuis trop peu de temps pour avoir pu bénéficier des améliorations apportées par l'évolution des espèces.

Nous en arrivons au cœur de mon argumentation visant à comprendre le néocortex et savoir pourquoi la mémoire et la prédiction sont les clés du mystère de l'intelligence. Commençons par le cerveau des reptiles, qui est dépourvu de cortex. L'évolution a découvert que si elle ajoute un système de mémorisation (le néocortex) à la fibre sensorielle d'un cerveau primitif, l'animal y gagne la capacité de prédire le futur. Imaginez le vieux cerveau du reptile qui vit sa vie comme d'habitude, sauf que des patterns sensoriels sont à présent simultanément introduits dans le néocortex. Ce dernier les stocke dans sa mémoire. Par la suite, lorsque l'animal rencontrera une situation identique ou similaire, la mémoire reconnaîtra que les inputs sont semblables et se rappellera de ce qui s'était produit dans le passé. La mémoire rappelée est comparée au flux des inputs sensoriels. Tout à la fois, elle s'imprègne des inputs courants et prédit ce qui va être vu. En comparant les inputs sensoriels visuels avec la mémoire rappelée, l'animal sait non seulement où il est, mais il peut aussi se projeter dans le futur.

Supposons maintenant que le cortex se souvienne non seulement de ce que l'animal a vu, mais aussi du comportement du cerveau archaïque dans une situation identique. Il n'est pas même nécessaire de présumer que le cortex sait faire la différence entre les sensations et le comportement ; pour lui, ce ne sont que des patterns. Quand notre animal se retrouve dans une situation identique ou similaire, il ne fait pas que prévoir (dans le sens

étymologique du terme : voir par avance), mais se rappelle aussi quels sont les comportements qui tendent vers cette vision du futur. De ce fait, la mémoire et la prédiction permettent à l'animal d'utiliser plus intelligemment ses comportements existants (cerveau archaïque).

Par exemple, supposons que vous êtes un rat qui apprend à s'orienter dans un labyrinthe pour la première fois. Excité par l'incertitude ou la faim, vous exploitez les aptitudes inhérentes à votre cerveau archaïque pour explorer votre nouvel environnement. Vous écoutez, regardez, reniflez et courez le long des parois. Toutes ces informations sensorielles sont utilisées par votre cerveau archaïque, mais aussi transmises au néocortex, où elles sont stockées. Quelque temps plus tard, vous vous retrouvez dans le même labyrinthe. Votre néocortex reconnaît l'input courant comme l'un de ceux qu'il a déjà rencontrés et se remémore des patterns stockés correspondant à ce qui s'était passé autrefois. Il permet essentiellement de vous projeter un tout petit peu dans le futur. Le rat que vous êtes se dirait : « Ah oui, je reconnais ce labyrinthe et je me souviens de ce coin. » Tandis que votre néocortex se rappelle de ce qui s'était produit dans le passé, vous envisagez de retrouver le morceau de fromage que vous aviez découvert la dernière fois dans le labyrinthe, et comment vous y étiez parvenu : « Si je tourne à droite ici, je sais ce qu'il y a ensuite. Le morceau de fromage se trouve au bout du couloir. Je le vois en pensée. » Tandis que vous courez frénétiquement dans le labyrinthe, vous vous fiez à des structures primitives, plus anciennes, pour entreprendre des mouvements comme vous dresser sur vos pattes et agiter vos moustaches. Grâce à votre (relativement) grand néocortex, vous vous souvenez des endroits où vous avez été, vous les reconnaissez par la suite, et vous faites des prédictions sur ce qui se passera. Un lézard dépourvu de néocortex a bien moins de capacités pour se souvenir du passé, et chaque fois qu'il est placé dans le labyrinthe, il devra procéder comme s'il n'y avait jamais été. Vous, le rat, comprenez le monde et le futur

immédiat grâce à votre mémoire corticale. Vous avez une représentation claire des récompenses et des dangers découlant de chacune de vos décisions, et vous vous déplacez efficacement dans votre univers. Vous pouvez littéralement voir le futur.

Remarquez toutefois que vous n'exécutez aucun nouveau comportement particulièrement complexe ou fondamental. Vous ne fabriquez pas un deltaplane pour voler vers le fromage qui attend au bout du couloir. Votre néocortex élabore des prédictions au sujet des patterns sensoriels qui permettent de prévoir, mais votre palette de comportements n'est quasiment pas affectée. Votre capacité à courir à petits pas, grimper et explorer est très comparable à celle du lézard.

Sa surface augmentant au gré de l'évolution, le cortex se révéla capable de retenir de plus en plus de choses. Il lui fut possible d'accroître sa mémoire et procéder à davantage de prédictions. La complexité de la mémoire et des prédictions s'accrut aussi. Mais quelque chose d'autre se produisit, qui conduisit à la capacité propre à l'homme de se comporter intelligemment.

Le comportement humain transcende le vieux répertoire des comportements de base consistant à se mouvoir de-ci de-là avec les aptitudes d'un rat. Nous avons porté l'évolution néocorticale à un nouveau niveau. Seuls les humains ont créé un langage parlé et écrit. Seuls les humains cuisent leur nourriture, cousent des vêtements, pilotent des avions et construisent des immeubles. Nos capacités motrices et organisationnelles dépassent de loin celles des animaux qui nous sont physiologiquement les plus proches. Comment le cortex, qui a été conçu pour établir des prédictions sensorielles, peut-il générer ce comportement incroyablement sophistiqué, unique au genre humain ? Comment le comportement supérieur a-t-il pu évoluer si soudainement ? Il y a deux réponses à ces questions. La première est que l'algorithme néocortical est si puissant et si souple qu'au prix d'un petit recâblage propre aux humains, il est capable de produire de nouveaux comportements très sophistiqués. La seconde réponse est que

comportements et prédictions sont les deux aspects d'un même concept. Bien que le cortex puisse envisager le futur, il ne peut procéder à des prédictions sensorielles exactes que s'il sait quels comportements ont été effectués.

Dans l'exemple simple du rat cherchant le fromage, l'animal se souvient du labyrinthe et se sert de sa mémoire pour prédire qu'il verra la nourriture après avoir tourné au coin. Mais le rat pourrait aller à droite ou à gauche. C'est seulement parce qu'il se souvient à la fois du fromage et du comportement correct, « tourner à droite à l'embranchement », qu'il peut faire en sorte que sa prédiction concernant le fromage soit avérée. Bien que cet exemple soit élémentaire, il illustre combien la prédiction sensorielle et le comportement sont intimement liés. Tous les comportements agissent sur ce que nous voyons, entendons et ressentons. La plus grande part de ce que nous éprouvons à quelque moment que ce soit dépend grandement de nos propres actions. Approchez votre bras du visage. Pour prédire que vous verrez le bras, votre cortex doit savoir qu'il a commandé de le bouger. Si le cortex prenait connaissance du mouvement du bras sans qu'il ait émis les ordres moteurs correspondants, vous seriez quelque peu surpris. La manière la plus simple d'interpréter cet exemple serait de supposer que le cerveau fait d'abord bouger le bras puis prédit que vous le verrez. Je pense que c'est faux. Je crois plutôt que le cortex prédit la vision du bras, et c'est cette prédiction qui fait que les commandes motrices agissent pour que cela se réalise. Vous pensez d'abord, ce qui vous fait agir pour que vos pensées se réalisent.

Examinons à présent les changements qui ont conduit les êtres humains à disposer d'un répertoire comportemental extrêmement étendu. Existe-t-il des différences physiques entre le cortex d'un singe et celui d'un humain qui expliqueraient pourquoi seuls les humains ont un langage et des comportements complexes ? Le cerveau humain est environ trois fois plus gros que celui d'un chimpanzé. Mais ce volume supérieur n'explique pas tout.

Une des clés permettant de comprendre le bond en avant que représente le comportement humain se trouve dans le câblage qui relie les régions du cortex aux parties archaïques du cerveau. Bref, nos cerveaux sont connectés différemment de ceux des singes.

Examinons cela de près. La notion d'hémisphère cérébral droit et gauche vous est sans doute familière. Il existe toutefois une autre subdivision, moins connue, et c'est là que nous devons rechercher les différences qui distinguent l'être humain de l'animal. Dans tous les cerveaux, notamment ceux de grande taille, le cortex est divisé en une moitié antérieure et une moitié postérieure séparées par un large sillon, la scissure centrale. C'est dans la partie arrière, occipitale, du cortex que parviennent les inputs visuels, auditifs et tactiles. C'est surtout là que se produit la perception sensorielle. La partie avant, frontale, contient les régions du cortex impliquées dans les tâches de niveau élevé : la planification et la pensée. C'est là aussi que se trouve le cortex moteur, c'est-à-dire la partie du cerveau qui actionne les muscles, et par conséquent produit le comportement.

La surface du néocortex des primates s'accroissant peu à peu, la largeur de la partie antérieure devint disproportionnée, notamment chez l'être humain. Comparé à celui d'autres primates et pré-hominidés, notre front est énorme, conçu pour héberger un très vaste cortex antérieur. Mais cette hypertrophie n'explique pas à elle seule les améliorations de nos capacités motrices, comparées à celles d'autres créatures. Nos dispositions à exécuter des mouvements exceptionnellement compliqués découlent du fait que notre cortex moteur entretient beaucoup plus de connexions avec nos muscles. Chez d'autres mammifères, le cortex frontal joue un rôle moins direct dans le comportement moteur. La plupart des autres animaux se fient largement aux parties archaïques du cerveau pour produire leurs comportements. En revanche, le cortex moteur obtient la plupart des contrôles moteurs du restant du cerveau. Si vous endommagez le cortex moteur d'un rat, l'animal ne présentera pas forcément un

déficit moteur sensible. Mais si vous endommagez le cortex moteur d'un être humain, il devient paralysé.

Les gens m'interrogent souvent sur les dauphins. Leur cerveau n'est-il pas énorme ? Certes, le néocortex d'un dauphin est vaste. La structure de leur cortex est plus rudimentaire que la nôtre (trois couches seulement au lieu des six que nous avons), mais à tout autre point de vue, leur cerveau est gros. Il est probable qu'un dauphin peut se souvenir et comprendre beaucoup de choses. Il reconnaît chacun des autres dauphins. Il possède sans doute une bonne mémoire de son existence, dans le sens autobiographique. Il connaît sûrement tous les coins et recoins des régions où il évolue. Mais bien qu'ils manifestent quelques comportements sophistiqués, les dauphins ne nous sont pas proches. Nous pouvons conjecturer que leur cortex a une influence peu déterminante sur leur comportement. Le fait est que le cortex a principalement évolué pour procurer une mémoire du monde. Un animal doté d'un vaste cortex pourrait percevoir le monde un peu comme vous et moi. Mais l'être humain est unique de par le rôle dominant, avancé, que le cortex joue dans son comportement. C'est pour cela que nous avons élaboré un langage complexe ainsi que des outils compliqués, ce qui n'est pas le cas des animaux. C'est pourquoi nous pouvons écrire des romans, surfer sur Internet, construire des avions de ligne et envoyer une sonde sur Titan.

Nous avons à présent une vue d'ensemble. La nature a d'abord créé des animaux comme les reptiles, dotés de sens élaborés et de comportements sophistiqués mais relativement rigides. Elle s'est ensuite rendu compte qu'en lui adjoignant un système mnémonique alimenté par des flux sensoriels, l'animal pouvait se souvenir d'expériences passées. Ainsi, lorsqu'il serait de nouveau confronté à une situation semblable ou similaire, la mémoire pourrait être rappelée, conduisant à une prédiction de ce qui pourrait arriver ensuite. L'intelligence et la compréhension ont commencé sous la forme de systèmes mnémoniques introduisant des prédictions

dans le flux sensoriel. Ces prédictions sont l'essence même de la compréhension : savoir quelque chose suppose que vous pouvez faire des prédictions à ce sujet.

Le cortex a évolué dans deux directions. Premièrement, sa surface s'est accrue et a gagné en sophistication au niveau des types de mémoires qu'il peut stocker. Il devint capable de mémoriser davantage d'informations et procéder à des prédictions fondées sur des relations plus complexes. Deuxièmement, il se mit à interagir avec le système moteur du cerveau archaïque. Pour prédire ce que vous verrez, entendrez ou ressentirez ensuite, il lui fallait savoir quelles actions étaient entreprises. Chez l'être humain, le cortex prend en charge la plus grande partie de notre comportement moteur. Au lieu de ne faire que des prédictions fondées sur le comportement du cerveau archaïque, le néocortex ordonne le comportement qui satisfera à ses prédictions.

Le cortex humain est particulièrement vaste et possède de ce fait une énorme capacité de mémorisation. Il prédit constamment ce que vous verrez, entendrez et ressentirez, généralement sans que vous en ayez conscience. Ces prédictions sont nos pensées, et quand elles s'associent à des inputs sensoriels, elles sont nos perceptions. J'appelle cette vision du cerveau le *cadre de mémoire-prédiction* de l'intelligence.

Si la Chambre chinoise de Searle avait été dotée d'un tel système mnémonique, capable de prédire le prochain idéogramme ainsi que ce qui se passerait ensuite dans l'histoire, nous pourrions avoir l'assurance que la chambre (NdT : qui serait ici une métaphore du cortex) comprend le chinois et comprend l'histoire. Nous pouvons à présent voir où Alan Turing s'est trompé. C'est la prédiction, et non le comportement, qui apporte la preuve de l'intelligence.

Nous sommes maintenant prêts à étudier en détail ce nouveau concept de cadre de mémoire-prédiction du cerveau. Pour faire des prédictions sur les événements futurs, notre néocortex doit stocker des séquences de patterns. Pour rappeler les mémoires

appropriées, il doit en extraire les patterns selon leur similarité avec les patterns passés (rappel auto-associatif). Et enfin, les différentes mémoires doivent être stockées sous une forme invariante afin que la connaissance des événements passés puisse être appliquée à de nouvelles situations similaires, mais pas identiques, à celles du passé. La manière dont le cortex physique accomplit cette tâche ainsi que l'étude approfondie de ses hiérarchies sont le sujet du prochain chapitre.

# 6

# LE FONCTIONNEMENT DU CORTEX

**Tenter** de comprendre le fonctionnement du cerveau équivaut à résoudre un puzzle géant. Deux approches sont envisageables. Si vous adoptez la démarche déductive, du général au particulier, vous commencez par l'image de ce que doit représenter le puzzle, et vous vous fondez dessus pour savoir quelles pièces doivent être ignorées et lesquelles il faut rechercher. Si vous adoptez l'autre approche, inductive, du particulier au général, vous vous concentrez sur chacune des pièces. Vous recherchez leurs caractéristiques propres et vérifiez leur concordance avec d'autres pièces du puzzle. Si vous ne possédez pas l'image du puzzle terminé, la méthode inductive est parfois la seule et unique façon de procéder.

Le puzzle qui se rapporte à la compréhension du mécanisme cérébral est particulièrement ardu. Ne disposant pas d'une bonne structure permettant de comprendre l'intelligence, les chercheurs ont été obligés de s'en tenir à l'approche inductive, du particulier au général. Mais pour un puzzle aussi complexe que le cerveau, la tâche est colossale, voire impossible. Pour

vous faire une idée de la difficulté, imaginez un puzzle de plusieurs milliers de pièces : beaucoup peuvent être interprétées de multiples manières, comme si elles comportaient une image au recto et au verso, et pas uniquement sur une seule face. La forme des pièces n'est pas nettement définie, de sorte que vous n'avez jamais la certitude que deux pièces s'ajustent ou non. Beaucoup ne sont pas utilisées dans la représentation finale, mais vous ne savez ni lesquelles ni combien. Chaque mois, de nouvelles pièces vous arrivent par la poste. Certaines remplacent des anciennes, accompagnées d'une lettre du fabricant du puzzle disant : « Je sais que vous travaillez depuis quelques années avec ces anciennes pièces, mais il s'est révélé qu'elles sont erronées. Désolé. Utilisez désormais celles-ci jusqu'à nouvel ordre. » Vous ne savez hélas pas à quoi ressemblera l'œuvre terminée. Pire, vous en avez peut-être une idée ou une autre, mais toutes sont fausses.

Cette analogie avec un puzzle évoque fort bien les difficultés auxquelles vous êtes confronté lorsque vous tentez d'élaborer une nouvelle théorie du cortex et de l'intelligence. Les pièces sont les données biologiques et comportementales que les scientifiques ont collectées depuis bien plus d'une centaine d'années. Chaque mois, de nouvelles publications apparaissent, qui ajoutent de nouvelles pièces au puzzle. Parfois, les données d'un chercheur contredisent celles d'un autre. Comme elles peuvent être interprétées de différentes manières, des désaccords surgissent sur pratiquement tout. Sans une structure déductive, aucun consensus n'est possible sur ce qu'il faut rechercher, sur ce qui est le plus important ou sur la façon d'interpréter la montagne d'informations qui s'est accumulée. Nos tentatives pour comprendre le cerveau se sont limitées à une approche inductive, du particulier au général. Or, nous avons besoin d'une structure déductive.

Le modèle de mémoire-prédiction peut jouer ce rôle. Il peut nous montrer comment commencer à réunir les pièces du puzzle. Pour procéder à des prédictions, le cortex doit pouvoir mémoriser et stocker ses connaissances au sujet des séquences d'événements.

Pour faire des prédictions d'événements nouveaux, le cortex doit former des représentations invariantes. Le cerveau a besoin de créer et stocker un modèle du monde tel qu'il est véritablement, indépendamment de la manière dont vous le percevez dans des conditions changeantes. Savoir ce que le cortex doit faire nous guide dans la compréhension de son architecture, notamment sa structure à la fois hiérarchique et à six couches.

Lors de l'étude de cette nouvelle structure, présentée ici pour la première fois, nous évoquerons des notions qui risquent de se révéler ardues pour le néophyte. Beaucoup des concepts que vous rencontrerez sont peu familiers, même à des experts en neurobiologie. Mais moyennant quelques efforts, je crois que n'importe qui sera en mesure d'assimiler les fondements de cette nouvelle structure. Les Chapitres 7 et 8, bien moins techniques, exposent les principales implications de cette théorie.

Notre comparaison avec un puzzle peut à présent être appliquée aux notions biologiques qui soutiennent notre hypothèse de la mémoire-prédiction. C'est un peu comme mettre de côté une grande partie du puzzle, sachant que le relativement peu de pièces restantes nous révélera la solution ultime. Lorsque nous savons ce qu'il faut chercher, la tâche est moins ardue.

Je tiens aussi à souligner que cette nouvelle structure est incomplète. Beaucoup d'éléments m'échappent encore. Mais j'en ai compris beaucoup d'autres grâce à un raisonnement déductif, à des expériences faites dans de nombreux laboratoires, et à ce que nous savons de l'anatomie du cerveau. Au cours des cinq à dix dernières années, les chercheurs travaillant dans les divers domaines de la neurobiologie ont exploré des idées semblables aux miennes, bien qu'ils aient utilisé une terminologie différente, et n'aient pas — pour autant que je sache — tenté d'organiser leurs découvertes au sein d'une structure globale. Ils parlent de processus déductifs et inductifs, expliquent comment les patterns se propagent à travers des régions sensorielles du cerveau et insistent sur l'importance des représentations invariantes. Par exemple, Gabriel

Kreiman et Christof Koch, neurobiologistes au Caltech, en collaboration avec le neurochirurgien Itzhak Fried, de l'UCLA, ont découvert des cellules qui s'activent chaque fois que le sujet voit une photo de Bill Clinton. L'un de mes objectifs est d'expliquer comment ces cellules « bill-clintoniennes » ont pu se créer. Bien sûr, toute théorie doit produire des prédictions vérifiables en laboratoire. Je suggère bon nombre de ces prédictions dans l'annexe. Maintenant que nous savons ce qu'il faut rechercher, ce système très complexe paraîtra plus simple.

Dans les sections qui suivent, nous nous plongerons de plus en plus profondément dans le fonctionnement du modèle de mémoire-prédiction du cortex. Nous commencerons par la structure et la fonction globales du néocortex, puis nous tenterons de comprendre comment les éléments plus petits s'inscrivent dans le tout.

## LES REPRÉSENTATIONS INVARIANTES

J'avais précédemment décrit le cortex comme une feuille contenant des cellules, de la taille d'une nappe de table, épaisse de deux millimètres, dans laquelle les connexions entre diverses régions organisent l'ensemble en structure hiérarchique. Voici à présent une autre description du cortex qui met en avant sa connectivité hiérarchique. Découpons la nappe de table en régions fonctionnelles distinctes — des parties du cortex spécialisées dans certaines tâches — et empilons-les comme des crêpes (la Figure 1 montre une coupe de ce tas). En vérité, le cortex ne se présente pas du tout ainsi, mais ce schéma vous aidera à visualiser la circulation de l'information. Quatre régions corticales sont représentées, dans lesquelles les inputs sensoriels entrent par en dessous, dans la région la plus basse, et se propagent vers le haut de région en région. Remarquez que l'information circule dans les deux sens.

La Figure 1 représente les quatre premières régions visuelles impliquées dans la représentation des objets : comment vous parvenez à voir et reconnaître un chat, une cathédrale, votre

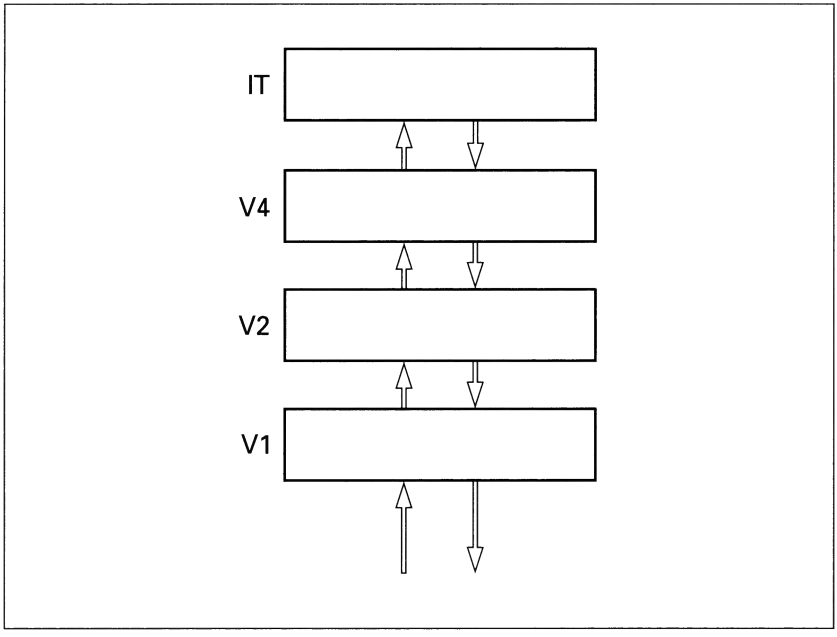



**Figure 1 : Les quatre premières régions visuelles de la reconnaissance des objets.**

mère, la Grande Muraille de Chine et les nommer. Les biologistes les ont appelées V1, V2, V4 et IT. Les inputs visuels, représentés par les flèches pointées vers le haut, à la Figure 1, proviennent des rétines et sont acheminés vers V1. Ces inputs peuvent être considérés comme des patterns constamment changeants, véhiculés par approximativement un million d'axones réunis en faisceaux pour former les nerfs optiques.

Nous avons déjà parlé des patterns temporels, mais il me semble utile de rappeler de quoi il s'agit car nous y ferons souvent allusion. Le cortex est une vaste feuille de tissus contenant des aires fonctionnelles spécialisées dans certaines tâches. Ces régions sont interconnectées par de larges faisceaux d'axones qui transfèrent l'information d'une région à une autre, en même temps. A tout moment, certains ensembles de fibres déclenchent des impulsions appelées « potentiels d'action » tandis que d'autres ne

le font pas. L'activité collective d'un ensemble de fibres est ce que nous entendons par le mot *pattern*. Les patterns arrivant en V1 peuvent être spatiaux, ce qui est le cas lorsque votre regard se pose un instant sur un objet, ou temporels, lorsque le regard parcourt cet objet.

Comme vous le savez, trois fois par seconde, les yeux se livrent à un mouvement rapide appelé « saccade » et à une immobilisation appelée « fixation ». Si l'on vous équipait d'un détecteur de mouvements oculaires, vous seriez surpris de découvrir combien vos saccades sont amples, bien que vous perceviez votre vision comme continue et stable. La Figure 2a montre le parcours du regard d'un observateur sur la photo d'un visage. Les fixations ne sont pas aléatoires. Imaginez que vous parvenez à visualiser le pattern d'activité qui arrive en V1, en provenance des yeux de l'observateur : il change complètement à chaque saccade. Plusieurs fois par seconde, le cortex visuel « voit » un pattern totalement renouvelé.

Peut-être pensez-vous qu'au cours des saccades c'est toujours le même sujet qui est vu, mais légèrement décalé. C'est un peu vrai, mais pas tant que ça. Les récepteurs photosensibles de la rétine sont irrégulièrement répartis. Ils sont surtout concentrés dans la fovéa, au milieu, et sont progressivement plus épars vers la périphérie. En revanche, dans le cortex, les cellules sont partout régulièrement réparties. Il en résulte que l'image rétinienne acheminée vers l'aire visuelle primaire V1 est extrêmement déformée. Quand le regard se fixe sur un nez au lieu d'un œil, l'input visuel est très différent, comme s'il provenait d'un objectif ultra grand-angulaire appelé *fish-eye* (œil de poisson) violemment secoué. Mais pour l'observateur, le visage ne paraît ni déformé ni secoué. Le plus souvent, vous ne vous rendez pas même compte que le pattern rétinien a complètement changé. Vous ne voyez qu'un visage. La Figure 2b montre cet effet sur la vision d'une plage. C'est une réaffirmation du mystère de la représentation invariante évoquée au Chapitre 4 à propos de la mémoire. Ce que vous « percevez » n'est pas ce que V1 voit.

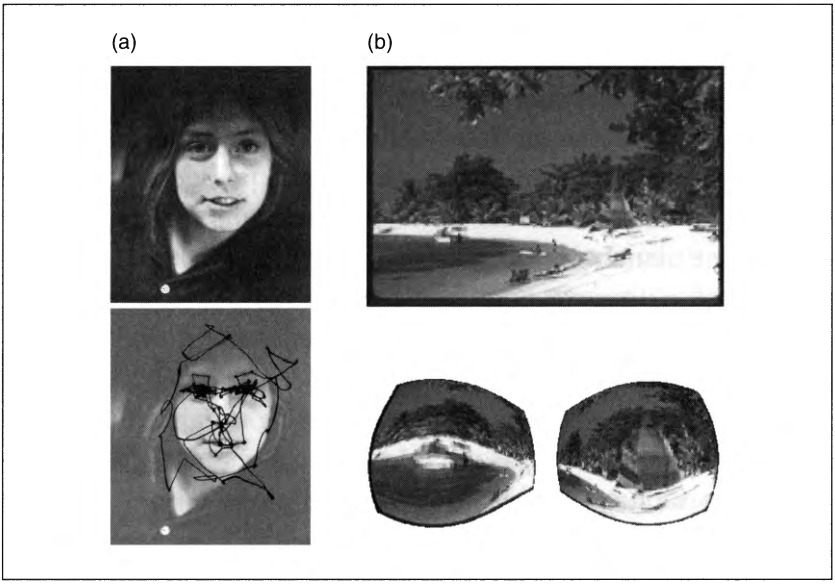



**Figure 2a : Saccades du regard sur la photo d'un visage.**
**Figure 2b : Distorsion causée par la répartition irrégulière des photorécepteurs sur la rétine.**

Comment votre cerveau peut-il même savoir qu'il « regarde » le même visage, et pourquoi ne savez-vous pas que les inputs sont changeants et déformés ?

Si nous placions une sonde en V1 et observions la réaction de chacune de ses cellules, nous découvririons que chacune d'elles en particulier ne réagit qu'en réponse à un input visuel provenant d'une minuscule partie de la rétine. Cette expérience, qui a été faite de nombreuses fois, est un des piliers de la recherche sur la vision. Chaque neurone V1 est doté de ce qu'il est convenu d'appeler « un champ réceptif » hautement spécifique à une partie minime du champ de vision total, c'est-à-dire du monde qui s'étend sous nos yeux. Les cellules V1 semblent n'avoir aucune connaissance de ce qu'est un visage, une voiture, un livre ou autre objet signifiant que nous voyons tout le temps. Tout ce qu'elles « connaissent » est une minuscule, microscopique partie du monde visuel.

Chaque cellule en V1 est réglée pour réagir à un type précis de patterns en entrée (inputs). Par exemple, une cellule en particulier s'activera vigoureusement si elle perçoit dans son champ réceptif une ligne ou une arête inclinée à trente degrés. Le fait que ce soit une ligne ou une arête n'a en soi que peu de signification. Il pourrait s'agir d'un élément appartenant à n'importe quel objet : une planche, le tronc d'un palmier au loin, un bord de la lettre *M* ou n'importe quoi d'autre. A chaque nouvelle fixation, le champ réceptif de la cellule en vient à s'arrêter sur une partie de l'espace visuel nouvelle et totalement différente. Lors de certaines fixations, la cellule s'active fortement, lors de certaines autres, elle s'active faiblement ou pas du tout. Par conséquent, chaque fois qu'une saccade se produit, de nombreuses cellules en V1 sont enclines à modifier leur activité.

Toutefois, quelque chose d'étonnant se produit quand vous plantez la sonde dans la région supérieure représentée à la Figure 1, la région IT (cortex inférotemporal). Elle contient des cellules qui s'activent et restent actives quand des objets entiers apparaissent dans le champ visuel. Nous trouverons par exemple une cellule nettement excitée chaque fois qu'un visage est visible. Elle reste active aussi longtemps que ce visage se trouve quelque part dans le champ de vision. Elle ne se désactive ni ne s'active à chaque saccade, comme les cellules en V1. Le champ réceptif de ces cellules IT recouvre la plus grande partie de l'espace visuel ; il est réglé pour ne s'activer que si un visage est visible.

Reformulons ce mystère. Au cours du franchissement des quatre niveaux corticaux de la rétine jusqu'en IT, les cellules sont passées d'un état de cellules de reconnaissance de minuscules caractéristiques rapidement changeantes, spatialement spécifiques, à des cellules de reconnaissance d'objet, spatialement non spécifiques, constamment excitées. La cellule IT nous informe que nous voyons un visage quelque part dans notre champ de vision. Elle s'active même si le visage est incliné, pivoté ou partiellement occulté. C'est une partie d'une représentation invariante de type « visage ».

Cette notion paraît bien simple : en quatre niveaux corticaux rapidement franchis, nous reconnaissons un visage. Aucun programme informatique ni aucune formule mathématique n'ont résolu ce problème avec la fiabilité et la généralité qu'autorise le cerveau humain. Maintenant que nous savons qu'il le résout en peu d'étapes, la réponse ne saurait être compliquée. L'un des principaux buts de ce chapitre est d'expliquer comment une « cellule de visage », celui de Bill Clinton ou quelqu'un d'autre, s'y prend. Nous y parviendrons, mais auparavant, nous aurons beaucoup de chemin à faire.

Examinons de nouveau la Figure 1. Vous constatez que l'information circule des régions les plus élevées vers les régions les plus basses par un réseau de connexions en feedback, ou rétropropagation. Ce sont des faisceaux d'axones qui s'étendent des régions les plus élevées comme IT vers des régions plus basses comme V4, V2 et V1. De plus, il y a autant sinon plus de connexions en feedback dans le cortex visuel qu'il y a de connexions en feedforward, ou antéropropagation.

Pendant de nombreuses années, la plupart des scientifiques ont ignoré ces connexions en feedback. Si votre compréhension du cerveau s'appuie seulement sur la manière dont le cortex reçoit les inputs, les traite et agit dessus, vous n'avez pas besoin de la notion de feedback. Tout ce dont vous avez besoin, ce sont des connexions en feedforward menant des parties sensorielles du cortex aux parties motrices. Mais quand vous commencez à réaliser que la fonction fondamentale du cortex est de faire des prédictions, vous devez introduire du feedback dans le modèle ; le cerveau doit en effet renvoyer le flux d'informations vers les zones qui ont d'abord reçu les inputs. La prédiction exige de comparer ce qui s'est passé à ce à quoi vous vous attendiez. Ce qui s'est véritablement passé monte, ce à quoi vous vous attendiez descend.

Le même processus d'antéropropagation et de rétropropagation se produit dans toutes les aires corticales impliquant la totalité des sens. La Figure 3 montre l'empilement des régions

visuelles placé à côté d'empilements semblables pour l'ouïe et le toucher. Elle montre aussi quelques régions corticales de niveau plus élevé, des aires associatives, qui reçoivent et intègrent les inputs provenant d'un ensemble de différents sens, comme l'ouïe, auxquels s'ajoutent le toucher et la vue. Alors que la Figure 1 est fondée sur une connectivité connue entre quatre régions connues du cortex, la Figure 3 est un diagramme purement conceptuel qui ne vise pas à représenter les régions corticales telles qu'elles sont véritablement. Dans le cerveau humain, des dizaines de régions corticales sont en effet interconnectées de toutes sortes de façons. A vrai dire, la plus grande partie du cortex humain est formée d'aires associatives. La représentation schématique montrée ici et dans les figures qui suivent sert à comprendre sans risquer d'être induit en erreur.

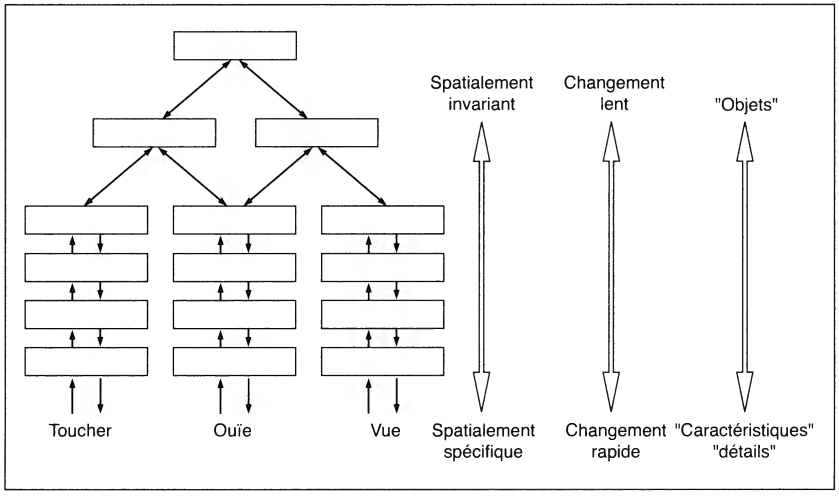



**Figure 3 : Formation de représentations invariantes pour l'ouïe, la vue et le toucher.**

---

La transformation — d'un changement rapide à un changement lent, et du spatialement spécifique au spatialement invariant — est bien décrite en ce qui concerne la vision. Bien que peu d'évidences contribuent à le prouver, nombreux sont les neurobiologistes qui

pensent que le même processus se produit dans toutes les aires sensorielles du cortex, et pas uniquement dans celle de la vision.

Prenons l'ouïe. Quand quelqu'un vous parle, les changements de pression d'air produits par le son se font très rapidement. Les patterns atteignant l'aire auditive primaire, A1, changent tout aussi vite. Mais si nous pouvions placer une sonde en amont dans le flux auditif, nous découvririons des cellules invariantes qui réagissent aux mots, voire dans certains cas à des phrases. Le cortex auditif serait doté d'un groupe de cellules qui réagissent lorsque vous entendez « merci », et d'un autre groupe de cellules qui seraient excitées par les mots « bonne journée ». De telles cellules restent actives pendant toute la durée de la formulation, à condition bien sûr que le sens des mots vous soit connu.

Les patterns reçus dans la première aire auditive peuvent varier considérablement. Un mot peut être dit avec différents accents, différentes hauteurs de voix ou à différentes vitesses. Mais, plus haut dans le cortex, ces fonctionnalités de bas niveau importent peu. Un mot est un mot, indépendamment des détails acoustiques. Il en va de même pour la musique. Vous pouvez écouter « Quand trois poules vont au champ » joué au piano, à la clarinette ou chanté par un enfant : dans chaque cas, la région A1 reçoit un pattern complètement différent. Mais une sonde placée dans une région plus élevée découvrirait des cellules qui réagissent régulièrement chaque fois que « Quand trois poules vont au champ » est joué, quels que soient l'instrument, le rythme ou tout autre détail. Cette expérience n'a évidemment pas été faite car elle est trop invasive pour être infligée à un être humain, mais si vous acceptez le postulat d'un algorithme cortical commun, il est certain que de telles cellules existent. Nous trouvons les mêmes sortes de biofeedback, de prédictions et de rappels invariants dans le cortex auditif que dans le système visuel.

Le toucher devrait se comporter de la même manière. Là encore, les expériences n'ont pas été faites, bien que des recherches soient menées sur des singes, en recourant à l'imagerie du

cerveau en haute résolution. Pendant que j'écris, je tiens un stylo. Je touche son capuchon et mon doigt effleure son agrafe métallique. Le pattern qui pénètre dans mon cortex somatosensoriel, envoyé par les capteurs tactiles situés dans la peau, varie rapidement lorsque mes doigts bougent. Mais c'est toujours un stylo que je perçois. A un moment, je pourrais plier l'agrafe métallique avec mes doigts, en faire autant plus tard avec d'autres doigts ou même le plier avec mes lèvres. Toutes ces actions produisent des inputs très différents qui parviennent à divers endroits de mon cortex somatosensoriel primaire. Toutefois, notre sonde découvrirait une fois de plus, dans des régions situées à plusieurs niveaux de distance de l'input primaire, des cellules qui réagiraient invariablement à « stylo ». Elles resteraient actives pendant que j'effleure le stylo, sans se préoccuper de savoir avec quels doigts ou quelles parties de mon corps je le touche.

L'ouïe et le toucher ne permettent pas de reconnaître un objet à partir d'un input momentané. Le pattern provenant des capteurs auditifs ou tactiles ne contient pas suffisamment d'informations, à un moment donné, pour vous apprendre ce que vous entendez ou touchez. Quand vous percevez une série de patterns auditifs comme une mélodie, un mot, une porte qui claque, ou quand vous percevez de façon tactile un objet comme un stylo ou son agrafe, le seul moyen d'identifier ces événements ou ces objets est le flux d'inputs qui s'étend dans la durée. Une mélodie ne peut pas être reconnue à partir d'une seule note, ni un stylo à partir d'un unique toucher. C'est pourquoi l'activité neuronale correspondant à la perception mentale d'objets ou de mots doit durer plus longtemps que celles de chaque pattern d'entrée. Ce n'est qu'un autre moyen de parvenir à la même conclusion que celle produite à un niveau plus élevé dans le cortex, les quelques changements dans la durée devenant perceptibles.

La vision, qui est aussi un flux d'inputs fondé sur la durée, se comporte de la même manière générale que l'ouïe et le toucher. Mais comme nous avons la capacité de reconnaître un objet du

premier coup d'œil, cela complique les choses. En fait, cette capacité à reconnaître instantanément des patterns spatiaux a induit en erreur, pendant plusieurs années, les chercheurs qui étudiaient la vision des animaux et des machines. Ils ont généralement ignoré l'importance du facteur « temps ». Les humains peuvent, en laboratoire, être entraînés à reconnaître des objets sans bouger les yeux, mais ce n'est pas la règle. La vision normale, comme lors de la lecture de ce livre, exige des mouvements oculaires constants.

## L'INTÉGRATION DES SENS

Qu'en est-il des aires associatives ? Jusqu'à présent, nous avons vu comment l'information monte vers une aire sensorielle particulière du cortex et en descend. Le flux descendant fournit l'input courant et fait des prédictions sur ce que sera la prochaine expérience sensorielle. Le même processus se produit entre les sens, c'est-à-dire entre la vue, l'ouïe et le toucher. Par exemple, ce que j'entends peut me permettre de prédire ce que je verrai ou ressentirai. En ce moment, j'écris dans ma chambre à coucher. Ma chatte Keo porte un collier qui tinte quand elle se promène. Je l'entends approcher dans le couloir. Grâce à cet input auditif, je reconnais ma chatte, je tourne ma tête vers le couloir, et voilà Keo qui entre. Je m'attendais à la voir à cause du son. Si elle n'était pas entrée, ou si un autre animal était apparu, j'aurais été surpris. Dans cet exemple, un input auditif a d'abord produit une reconnaissance auditive de Keo. L'information est remontée le long de la hiérarchie auditive jusqu'à une aire associative qui lie la vue à l'ouïe. La représentation est ensuite redescendue le long des hiérarchies auditives et visuelles, conduisant aux prédictions à la fois auditives et visuelles. La Figure 4 illustre ce cheminement.

Voici un autre exemple : plusieurs fois par semaine, je vais au travail à bicyclette. Ces matins-là, je vais au garage, je prends le vélo, je lui fais faire demi-tour puis je m'engage sur la chaussée. Pendant toutes ces manœuvres, mon cerveau reçoit de nombreux inputs visuels, tactiles et auditifs. La bicyclette heurte le montant

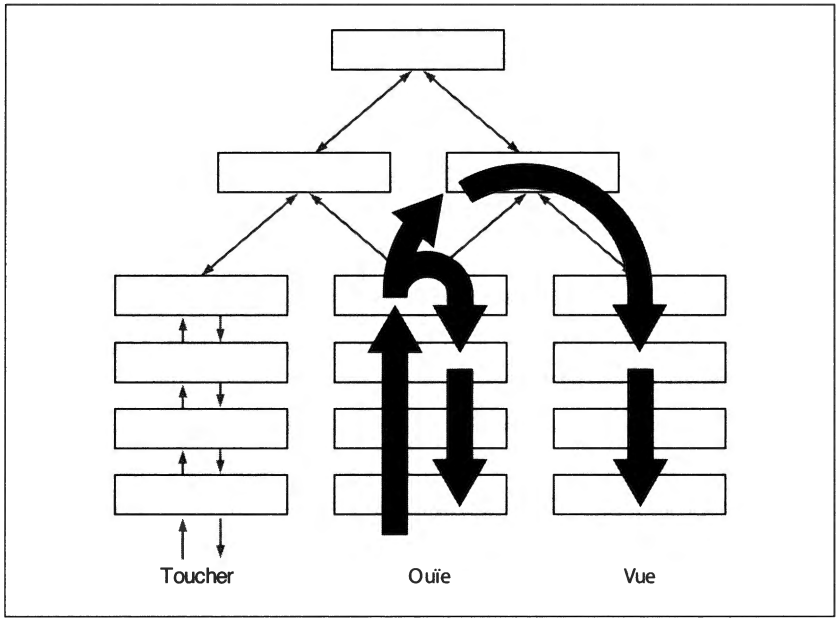



**Figure 4 : L'information monte puis descend le long des hiérarchies sensorielles afin d'établir des prédictions et produire une expérience sensorielle unifiée.**

---

de la porte, la chaîne émet un cliquetis, une pédale frôle ma jambe, la roue tourne en frottant sur le sol. Pendant que je sors la bicyclette du garage, mon cerveau est soumis à un déferlement de sensations provenant de la vue, de l'ouïe et du toucher. Chaque flux d'inputs sensoriels produit des prédictions pour les autres sens d'une manière incroyablement coordonnée. Tout ce que je vois suscite des prédictions précises sur ce que je vais ressentir et entendre, et sur d'autres sensations. Voir la bicyclette cogner contre le montant fait que je m'attends à entendre un bruit particulier et ressentir le rebond du vélo. Sentir le contact avec la pédale m'incite à regarder vers le bas et prédire que la pédale est tout près de l'endroit où je l'ai sentie. Ces prédictions sont si précises que si un seul de ces inputs était un tant soit peu mal coordonné ou inhabituel, je m'en rendrais compte immédiatement. L'information monte dans les hiérarchies sensorielles et en

descend pour produire une expérience sensorielle unifiée impliquant des prédictions pour tous les sens.

Essayez cette expérience : cessez de lire et faites autre chose, n'importe quelle activité vous obligeant à bouger le corps et manipuler un objet. Par exemple, allez à l'évier et ouvrez le robinet. Pendant que vous le faites, essayez de prendre conscience de chaque bruit, de chaque sensation tactile et des changements de l'input visuel. Vous devrez vous concentrer. Chaque action est intimement liée à la vue, à l'ouïe et au toucher. Actionnez le robinet, et votre cerveau s'attendra à ressentir une pression sur les doigts et une résistance à l'action musculaire. Vous vous attendez à voir et ressentir le mouvement du robinet, et vous vous attendez à voir et entendre l'eau couler. Dès que le jet frappe l'évier, vous vous attendez à entendre un bruit différent, et voir et ressentir l'éclaboussure.

Chaque pas que nous faisons fait un bruit que nous anticipons toujours, consciemment ou non. Rien que tenir ce livre tend à produire beaucoup de prédictions sensorielles. Imaginez que vous ressentez et entendez le livre se fermer, mais que visuellement il reste ouvert. Vous seriez surpris et troublé. Comme nous l'avons vu avec la porte faussée au Chapitre 5, vous faites des prédictions constantes concernant le monde qui vous entoure, qui sont coordonnées par tous vos sens. Quand je me concentre sur toutes les petites sensations, je suis stupéfait de constater à quel point nos prédictions perceptuelles sont pleinement intégrées. Bien que ces prédictions puissent paraître simples ou anodines, essayez de mesurer combien elles sont envahissantes, et comment elles ne peuvent survenir que grâce à une totale coordination des patterns qui montent et descendent dans la hiérarchie corticale.

Quand vous aurez compris à quel point les sens sont interconnectés, vous pourrez en conclure que le néocortex tout entier, toutes les aires sensorielles et associatives agissent comme une seule entité. Eh oui, nous avons un cortex visuel, mais ce n'est

qu'un élément d'un seul système sensoriel global : la vue, l'ouïe, le toucher et les autres sens, tous combinés, dont les flux montent et descendent le long d'une hiérarchie à branches multiples.

Un autre point : toutes les prédictions sont acquises par l'expérience. Nous nous attendons à ce que l'agrafe du stylo que l'on relâche claque, aujourd'hui et dans le futur, parce que c'est ce qui s'est produit dans le passé. Le bruit de la bicyclette que l'on sort du garage, ce que l'on voit et ressent, tout cela nous est prévisible. Nous ne sommes pas nés avec ce savoir ; nous l'avons acquis grâce aux énormes capacités de notre cortex à se souvenir des patterns. S'il existe des patterns homogènes parmi les inputs acheminés vers notre cerveau, notre cortex saura s'en servir pour prédire les événements futurs.

Bien que les Figures 3 et 4 ne représentent pas le cortex moteur, vous pouvez imaginer qu'il est lui aussi constitué d'un empilement hiérarchique semblable à l'empilement sensoriel, relié au système sensoriel par des aires associatives, mais avec cependant des connexions plus étroites avec le cortex somatosensoriel afin de percevoir les mouvements du corps. Ainsi, le cortex moteur se comporte en grande partie de la même manière qu'une région sensorielle. Un input dans n'importe quelle aire sensorielle peut monter jusqu'à une aire associative, qui peut inciter un pattern à descendre le long du cortex moteur, produisant un comportement. Tout comme il est capable d'inciter des patterns à descendre vers les parties du cortex de l'ouïe et du toucher, l'input visuel peut aussi les inciter à descendre le long des parties motrices du cortex. Dans le premier cas, nous interprétons ces patterns descendants comme des prédictions. Dans le cortex moteur, nous les interprétons comme des commandes motrices. Comme l'avait mentionné Mountcastle, le cortex moteur ressemble au cortex sensoriel. C'est pourquoi la manière dont le cortex traite les prédictions sensorielles descendantes est analogue à la manière dont il traite les commandes motrices descendantes.

Nous verrons d'ici peu qu'il n'existe pas de pures aires motrice ou sensorielle dans le cortex. Les patterns s'écoulent à la fois quelque part et n'importe où, et ils redescendent de n'importe quelle aire de la hiérarchie, suscitant des prédictions ou des comportements moteurs. Bien que le cortex moteur possède quelques caractéristiques spéciales, il est admis de penser qu'il n'est qu'une partie d'un vaste système de mémoire-prédiction hiérarchique. Il est quasiment comme n'importe quel autre sens. Voir, entendre, toucher et agir sont profondément entrelacés.

## UNE VISION NOUVELLE DE V1

La prochaine étape de l'élucidation de l'architecture du cortex exige un regard nouveau sur les régions corticales. Nous savons que les régions les plus élevées de la hiérarchie forment des représentations invariantes. Mais pourquoi cette fonction importante ne se produit-elle qu'au niveau le plus élevé ? La notion de symétrie de Mountcastle à l'esprit, j'ai commencé à explorer les différentes manières par lesquelles les régions corticales peuvent être connectées.

La Figure 1 montre les quatre régions classiques du cheminement visuel, V1, V2, V4 et IT. La région V1 se trouve tout en bas de l'empilement, suivie de V2, V4 et, tout en haut, IT. Conventionnellement, chacune est vue et représentée comme une seule région continue. C'est pourquoi toutes les cellules en V1 sont censées faire les mêmes choses, bien qu'en différentes parties du champ visuel. Toutes les cellules en V2 font les mêmes sortes de tâches, et toutes les cellules en V4 sont elles aussi spécialisées.

Dans cette représentation traditionnelle, quand l'image d'un visage entre dans la région V1, les cellules qui s'y trouvent en produisent un grossier dessin fait de droites (ou segments de lignes) et autres caractéristiques élémentaires. Ce dessin est transmis à V2, qui procède à une analyse plus détaillée des caractéristiques faciales. Puis il est transmis à V4, et ainsi de suite. L'invariance — et la reconnaissance de l'objet — n'est obtenue que quand l'input atteint le sommet, en IT.

Malheureusement, quelques problèmes affectent cette vision première des régions corticales comme V1, V2 et V4. Là encore, pourquoi les représentations invariantes ne devraient-elles se manifester qu'en IT ? Si toutes les zones corticales exécutent la même fonction, pourquoi IT devrait-elle faire bande à part ?

Deuxièmement, un visage peut apparaître du côté gauche de V1 ou du côté droit, et dans les deux cas vous le reconnaissez. Pourtant, des expériences ont clairement montré que des zones non adjacentes de V1 ne sont pas directement connectées : la partie gauche de V1 ne peut pas savoir directement ce que voit la partie droite. Revenons en arrière et réfléchissons-y. Les différentes parties de V1 font indubitablement des choses identiques, car toutes peuvent participer à la reconnaissance d'un visage, mais elles sont en même temps physiquement indépendantes. Des sous-régions, ou groupes, de V1 sont physiquement déconnectées mais font la même chose.

Enfin, des expériences ont montré que toutes les régions supérieures du cortex reçoivent des inputs convergents provenant de deux régions sensorielles ou plus situées en dessous (voir Figure 3). Dans la réalité, des dizaines de zones peuvent converger en une aire associative. Mais dans la représentation traditionnelle, les régions sensorielles inférieures comme V1, V2 et V4 semblent présenter une connectivité différente. Chacune paraît n'avoir qu'une seule source d'inputs — une flèche montante —, sans convergence évidente des inputs issus des diverses zones. V2 reçoit un input de V1 et c'était ainsi. Pourquoi certaines régions corticales recevraient des inputs convergents et d'autres pas ? Ceci aussi va à l'encontre de l'idée de Mountcastle d'un algorithme cortical commun.

Pour ces raisons et pour d'autres, j'en suis arrivé à penser que V1, V2 et V4 ne doivent pas être considérées comme des régions corticales seules. En réalité, chacune est un ensemble de plusieurs sous-régions plus petites. Revenons à l'analogie avec la nappe de table, censée représenter un cortex entier étalé. Supposons que

nous tracions dessus les limites de toutes les zones fonctionnaires. La plus vaste est incontestablement V1, la région principale de la vue, suivie de V2. Elles sont énormes comparées à la plupart des régions. Ce que je voudrais suggérer est que V1 devrait en réalité être considérée comme un ensemble de nombreuses très petites régions. Sur la nappe, elle ressemblerait à un patchwork de zones toutes contenues dans V1. Autrement dit, V1 est faite de nombreuses petites aires corticales séparées qui ne sont connectées qu'indirectement à leurs voisines, à travers des régions situées plus haut dans la hiérarchie. De toutes les aires de la vision, V1 posséderait le plus grand nombre de sous-régions. V2 serait composée de moins de sous-régions, légèrement plus grandes. Il serait de même pour V4. Mais tout en haut de la hiérarchie, IT serait d'un seul tenant, ce qui expliquerait pourquoi les cellules en IT ont une vue d'ensemble de la totalité du champ visuel.

Il y a là une intéressante symétrie. Jetez un coup d'œil à la Figure 5, qui montre la même hiérarchie qu'à la Figure 3, à la différence près qu'elle révèle les hiérarchies sensorielles que je viens de décrire. Remarquez la similitude du cortex partout. Choisissez n'importe quelle région et vous découvrirez que plusieurs autres, placées au-dessus, l'alimentent en inputs sensoriels convergents. Celle qui les reçoit renvoie des projections aux régions inférieures, qui leur indiquent quels patterns elles doivent s'attendre à voir. Les aires associatives situées plus haut unifient l'information provenant de multiples sens comme la vue et le toucher. Une région inférieure, comme une sous-région en V2, unifie l'information des sous-régions séparées présentes à l'intérieur de V1. Une région ne connaît pas — en fait, ne peut pas connaître — la signification du moindre de ces inputs. Une sous-région en V2 n'a pas à savoir qu'elle traite les inputs visuels de multiples parties de V1. Une aire associative n'a pas besoin de savoir qu'elle traite les inputs de la vue et de l'ouïe. La tâche de chaque région corticale consiste plutôt à découvrir les liens entre les inputs, mémoriser des successions de corrélations entre eux, et faire appel à la

mémoire pour prédire comment les inputs se comporteront dans le futur. Un cortex est un cortex. Le même processus se produit partout : c'est l'algorithme cortical commun.

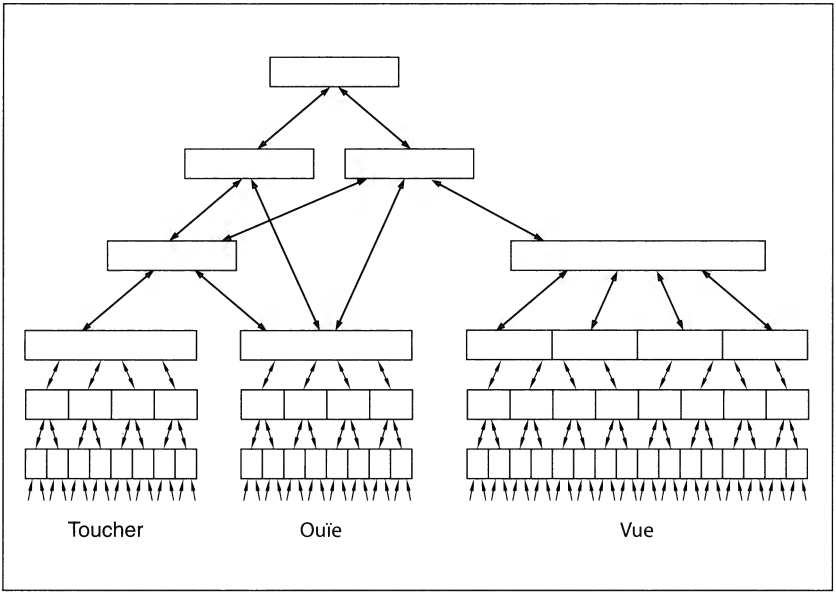



**Figure 5 : Une autre vision de la hiérarchie corticale.**

---

Ce nouveau schéma hiérarchique aide à comprendre le processus de création des représentations invariantes. Examinons de plus près ce qu'il en est de la vision. Au premier niveau du processus, l'espace visuel gauche est différent de l'espace visuel droit, de la même manière que l'ouïe est différente de la vue. La région V1 gauche et la région V1 droite ne forment un même type de représentation que parce qu'elles ont été exposées à des patterns identiques au cours de l'existence. A l'instar de l'ouïe et de la vue, elles peuvent être considérées comme des influx sensoriels séparés qui sont réunis au niveau plus élevé.

De même, les petites régions situées en V2 et V4 sont des aires associatives de la vue (ces sous-régions peuvent se chevaucher, mais cela ne change rien au principe fondamental). Interpréter le

cortex visuel de cette manière ne contredit ni ne modifie rien de ce que nous savons de son anatomie. L'information parcourt dans un sens et dans un autre toutes les branches de l'arborescence qu'est la mémoire hiérarchique. Un pattern dans le champ visuel gauche peut conduire à une prédiction dans le champ visuel droit exactement de la même manière que le grelot du collier de ma chatte entraîne la prédiction visuelle de son entrée imminente dans la chambre à coucher.

Le résultat le plus important de ce nouveau schéma de la hiérarchie corticale est que nous pouvons à présent dire que chaque et toutes les régions du cortex forment des représentations invariantes. Dans l'ancienne manière d'aborder cette question, nous n'avions pas de représentations invariantes complètes — comme des visages — tant que les inputs n'avaient pas atteint la couche supérieure IT, qui voit l'ensemble du champ visuel. Nous pouvons maintenant dire que les représentations invariantes sont omniprésentes. Elles sont formées dans chaque région corticale. L'invariance n'est pas un phénomène qui se manifeste comme par magie en allant dans les régions les plus hautes du cortex, comme IT. Chacune forme des représentations invariantes à partir des aires d'entrée situées hiérarchiquement dessous. De ce fait, les sous-régions en V4, V2 et V1 produisent des représentations invariantes fondées sur ce qu'elles reçoivent. Elles peuvent ne déceler qu'une minuscule partie du monde, et le vocabulaire des objets sensoriels auquel elles ont affaire est plus élémentaire, mais elles n'en exécutent pas moins la même tâche qu'en IT. De même, les aires associatives au-dessus de IT forment des représentations invariantes de patterns issus de multiples sens. Par conséquent, toutes les régions du cortex forment des représentations invariantes du monde qui se trouve hiérarchiquement dessous. Il y a là de la beauté.

Notre puzzle a changé. Nous n'avons plus à nous demander comment les représentations invariantes se forment au cours des quatre étapes, de bas en haut. Il nous faut plutôt nous demander

comment les représentations invariantes se forment dans chacune des régions corticales. C'est parfaitement sensé si nous admettons sérieusement l'existence d'un algorithme cortical commun. Si une région stocke des séquences de patterns, chaque région en stocke. Si une région crée des représentations invariantes, toutes les régions en créent. Redessiner la hiérarchie corticale comme à la Figure 5 rend cette interprétation possible.

## UN MODÈLE DU MONDE

Pourquoi le néocortex a-t-il une structure hiérarchique ?

Vous pouvez penser au monde, vous déplacer dans le monde et faire des prédictions parce que votre cortex a élaboré un modèle du monde. L'un des concepts les plus importants, dans ce livre, est que la structure hiérarchique du cortex stocke un modèle de la structure hiérarchique du monde réel. La structure imbriquée du monde réel est reproduite par la structure imbriquée de votre cortex.

Que faut-il entendre par structure hiérarchique ou imbriquée ? Pensez à la musique : les notes sont réunies pour former des intervalles. Les intervalles sont réunis pour former des phrases mélodiques. Les phrases sont réunies pour former des mélodies ou des chansons. Les chansons sont réunies dans des albums. Pensez à l'écriture : les lettres sont réunies en syllabes, les syllabes en mots. Les mots sont réunis pour former des propositions et des phrases. Regardez aussi autour de vous : vous verrez probablement des routes, des écoles, des immeubles… Les maisons ont des chambres, chaque chambre a des murs, un plafond, un plancher et une ou plusieurs fenêtres. Chaque fenêtre est composée d'éléments plus petits : du vitrage, des montants, des petits bois, des paumelles et une crémone. La crémone est faite de pièces comme la poignée, la tringle et les vis.

Observez autour de vous. Les patterns provenant de la rétine qui entrent dans le cortex visuel primaire sont combinés pour former des segments de ligne, qui s'assemblent pour obtenir des

formes plus complexes. Ces dernières se réunissent pour former des objets, un nez par exemple. Le nez est réuni avec les yeux et la bouche pour former le visage. Et le visage est réuni à d'autres parties du corps pour former la personne que vous voyez.

Tous les objets qui nous entourent sont ainsi composés d'assemblages de sous-objets ; ces derniers définissent l'objet. Quand nous attribuons un nom à un objet, nous le faisons parce qu'un ensemble de caractéristiques précises est réuni. Un visage est un visage justement parce qu'il a deux yeux, un nez et une bouche qui appartiennent au tout. Un œil est un œil justement parce qu'il a une pupille, un iris, une paupière, c'est-à-dire des éléments qui sont toujours ensemble. La même chose peut être dite des chaises, des voitures, des arbres, des parcs et des pays. Et enfin, une chanson est une chanson parce qu'une série d'intervalles se succèdent toujours en séquence.

Vu sous cette forme, le monde est comme une chanson : chaque objet qui s'y trouve est composé d'un ensemble d'objets plus petits, et la plupart des objets font partie d'objets plus grands. C'est ce que j'entends par « structure imbriquée ». Une fois que vous en avez pris conscience, vous découvrez des structures imbriquées partout. D'une manière analogue, votre mémoire des choses et la façon dont le cerveau se les représente sont toutes deux stockées dans la structure hiérarchique du cortex. La mémoire du lieu où vous habitez n'existe pas dans une région du cortex. Elle est répartie sur une hiérarchie de régions corticales qui reflètent la structure hiérarchique de l'habitation. Les relations à grande échelle sont stockées en haut de la hiérarchie, les relations à petite échelle en bas.

La conception du cortex et la méthode par laquelle il apprend révèlent naturellement les relations hiérarchiques du monde. Vous n'êtes pas né avec la connaissance du langage, des habitations ou de la musique. Le cortex est équipé d'un astucieux algorithme d'apprentissage qui découvre naturellement l'existence de toutes structures hiérarchiques et se les approprie. Lorsque

aucune structure n'est décelée, nous sommes plongés dans le trouble, voire dans le chaos.

Vous ne pouvez saisir qu'un sous-ensemble du monde à un moment donné. Vous ne pouvez être que dans une seule pièce de votre habitation, regardant dans une seule direction. A cause de la hiérarchie du cortex, vous pouvez savoir que vous êtes dans votre maison, dans le salon, regardant la fenêtre, même si à cet instant précis vos yeux sont fixés sur la poignée de la crémone. Les régions supérieures du cortex entretiennent une représentation de votre habitat tandis que les régions inférieures perçoivent une fenêtre. De même, la structure hiérarchique vous permet de savoir que vous écoutez à la fois une chanson et un album de musique, même si à chaque instant vous n'entendez qu'une seule note qui, à elle toute seule, n'exprime quasiment rien. La structure hiérarchique vous permet de savoir que vous êtes en compagnie de quelqu'un qui vous est cher même si vos yeux ne sont momentanément fixés que sur sa main. Les régions plus élevées du cortex conservent une trace de la perception globale tandis que les aires inférieures s'attachent activement aux petits détails rapidement changeants.

Comme nous ne pouvons toucher, entendre et voir qu'une très petite partie du monde à la fois, à un instant donné, l'information acheminée vers le cerveau lui parvient naturellement sous la forme d'une succession de patterns. Le cortex veut mémoriser ces séquences qui lui parviennent sans cesse. Dans certains cas, à l'instar d'une mélodie, la succession de patterns lui parvient dans un ordre immuable, celui des intervalles musicaux. La plupart d'entre nous connaissent bien ce genre de séquence. Mais j'utiliserai désormais le mot *séquence* d'une manière plus générale, plus proche dans sa signification du terme mathématique *ensemble*. Une séquence est un ensemble de patterns qui se succèdent généralement, mais pas toujours, dans un ordre défini. Ce qui est important est que les patterns d'une séquence en suivent ou en précèdent d'autres, même si ce n'est pas dans un ordre fixe.

Quelques exemples devraient clarifier ces propos. Quand je regarde votre visage, la séquence de patterns d'entrée que je vois n'est pas fixe mais définie par mes saccades oculaires. Je regarde une fois dans l'ordre « œil œil nez bouche », et un moment plus tard dans l'ordre « bouche œil nez œil ». Les composants du visage sont une séquence. Ils sont statistiquement liés et tendent à se produire ensemble, bien que l'ordre puisse varier. Si vous percevez « visage » en fixant « nez », il est fort probable que le prochain pattern soit « œil » ou « bouche », mais sans doute pas « stylo » ou « voiture ».

Chaque région du cortex reçoit un flux de ce genre de patterns. S'ils sont en relation les uns avec les autres de telle manière que la région corticale puisse apprendre à prédire le pattern qui suivra, la région forme une représentation persistante, ou mémoire, pour cette séquence. L'apprentissage des séquences est l'ingrédient le plus fondamental pour la formation des représentations invariantes des objets du monde réel.

Les objets du monde réel peuvent être matériels, comme un lézard, un visage ou une porte, ou abstraits comme un mot ou une théorie. Le cerveau traite le matériel et l'abstrait de la même manière. Tous deux ne sont que des séquences de patterns qui se déroulent d'une manière prédictible. La répétition à maintes reprises de certains patterns d'entrée permet à la région corticale de savoir que ces expériences sont causées par un objet réel du monde.

La prédictibilité est la définition même de la réalité. Si une région du cortex découvre qu'elle peut naviguer de façon fiable et prédictible parmi les patterns grâce à une série de mouvements physiques comme les saccades oculaires ou l'effleurement avec le doigt, ou peut les prédire avec exactitude tandis qu'ils se succèdent, comme c'est le cas pour la musique ou des phrases stéréotypées, le cerveau considère qu'il existe des relations causales entre les patterns. La probabilité de nombreux patterns se succédant avec une même relation, de façon répétée et par pure coïncidence, est des plus minimes. Une séquence de patterns prédictible

doit faire partie d'un objet plus grand qui existe réellement. Ainsi, une prédictibilité fiable est un moyen sûr et certain de savoir que différents événements sont physiquement liés. Chaque visage a des yeux, des oreilles, une bouche et un nez. Si le cerveau perçoit un œil puis, par des mouvements de saccade, un autre œil puis la bouche, il a la certitude que c'est un visage qui est vu.

Si les régions corticales pouvaient parler, elles diraient : « Je perçois beaucoup de patterns différents. Parfois, j'arrive à prédire ce que sera le prochain. Ces patterns sont assurément liés les uns aux autres. Ils se produisent toujours ensemble et je peux en toute quiétude passer de l'un à l'autre. Donc, chaque fois que je percevrai un de ces événements, j'y ferai référence par un nom qui leur sera commun. C'est ce nom de groupe, et non les patterns en particulier, que je transmettrai aux régions plus élevées du cortex. »

De ce fait, il est possible de dire du cerveau qu'il stocke des séquences de séquences. Chaque région du cortex apprend des séquences, développe ce que j'appelle des « noms » pour les séquences qu'il connaît, et transmet ces noms à la région immédiatement supérieure dans la hiérarchie corticale.

## LES SÉQUENCES DE SÉQUENCES

Lorsque l'information remonte des régions sensorielles primaires vers des niveaux plus élevés, un nombre toujours moindre de changements se produit. Dans les aires visuelles primaires comme V1, l'ensemble de cellules actives se modifie rapidement tandis que de nouveaux patterns se forment sur la rétine plusieurs fois par seconde. Dans l'aire visuelle IT, les cellules qui déclenchent des patterns sont plus stables. Que se passe-t-il à cet endroit ? Chaque région du cortex possède un répertoire de séquences qui lui sont connues, analogue au répertoire musical d'un artiste. Les régions stockent des séquences concernant tout et n'importe quoi : le bruit des vagues sur une plage, le visage maternel, le trajet de la maison jusqu'à la boutique du coin, comment s'épelle le mot « pop-corn », comment battre des cartes…

Nous nommons tout ce que nous connaissons, et de la même manière, chaque région corticale nomme chacune des séquences qu'elle connaît. Ce « nom » est un groupe de cellules dont les excitations collectives représentent l'ensemble des objets de la séquence (peu importe pour le moment comment le groupe de cellules représentant la séquence est sélectionné ; nous verrons cela plus tard). Ces cellules demeurent actives aussi longtemps que la séquence est jouée, et c'est ce « nom » qui est transmis à la prochaine région, dans la hiérarchie. Tant que les patterns d'entrée appartiennent à une séquence prédictible, la région présente un « nom » constant à la région située au-dessus.

C'est comme si la région disait : « Voici le nom de la séquence de ce que j'écoute, je vois ou je touche. Il est inutile de connaître chacune des notes, traits ou textures. Si quelque chose d'imprévu se produit, je vous le ferai savoir. » Plus spécifiquement, nous pouvons imaginer la région IT, en haut de la hiérarchie visuelle, relayant l'information vers une aire associative située au-dessus : « Je vois un visage. Oui, à chaque saccade les yeux en fixent une autre partie. Je vois ces différentes parties se succéder, mais c'est néanmoins toujours le même visage. Si je découvre quelque chose d'autre que ce qui est attendu, je vous le ferai savoir. » De cette manière, une séquence prédictible d'événements est identifiée par un « nom » : un pattern constant de cellules excitées. Ceci se produit sans cesse du bas en haut de la pyramide hiérarchique. Une région peut par exemple reconnaître la séquence de sons qui forme un phonème (les sons formant un mot) et transmettre un pattern représentant le phonème à la région au-dessus. La prochaine région supérieure reconnaît la séquence de patterns, ce qui lui permet de créer des mots. La prochaine région supérieure reconnaît des séquences de mots pour former des phrases, et ainsi de suite. Gardez à l'esprit qu'une « séquence » dans les régions les plus basses peut être toute simple, par exemple une ligne se déplaçant dans l'espace.

En réduisant les séquences prédictibles en « objets nommés » dans chaque région de la hiérarchie, nous obtenons une stabilité de plus en plus grande en remontant dans la hiérarchie. C'est ce qui produit des représentations invariantes.

L'effet opposé se produit lorsque le pattern descend dans la hiérarchie : les patterns stables se « décomposent » en séquences. Supposons que vous ayez appris le texte du discours de Gettysburg, qu'Abraham Lincoln prononça le 19 novembre 1863 sur le lieu de la plus meurtrière bataille de la guerre de Sécession, et que vous désiriez le réciter. Dans une région du langage située haut dans votre cortex est stocké un pattern qui représente le célèbre discours. Il est d'abord décomposé dans une mémoire contenant des séquences de phrases. Dans la région juste en dessous, chaque phrase est décomposée dans une mémoire contenant des séquences de mots. A ce point, le pattern décomposé se scinde et descend d'une part jusqu'à la partie auditive du cortex et d'autre part jusqu'à la partie motrice. Le long du cheminement moteur, chaque mot est décomposé en une séquence mémorisée de phonèmes. Enfin, dans la région tout en bas, chaque phonème est décomposé en une séquence de commandes musculaires qui produisent les sons. Plus vous descendez dans la hiérarchie, plus les patterns changent rapidement. Un unique pattern constant, tout en haut de la hiérarchie motrice, conduit en définitive à une séquence de sons vocaux plus complexe et plus longue.

L'invariance œuvre à notre avantage tandis que l'information redescend dans la hiérarchie. Si vous désiriez taper le discours de Gettysburg au lieu de le prononcer, vous commenceriez par le même pattern en haut de la hiérarchie. Il serait décomposé en phrases dans la région juste dessous, puis les phrases seraient décomposées en mots dans la région encore en dessous. Jusque-là, il n'y a aucune différence entre taper ou dire le discours. Mais, dès le niveau en dessous, le cortex moteur choisit une autre voie. Les mots sont décomposés en lettres, et les lettres en commandes musculaires qui actionnent les doigts chargés de taper le texte :

« Voici quatre-vingt-sept ans, nos pères ont apporté sur ce continent une nouvelle nation… » Les mémorisations des mots sont gérées en tant que représentation invariante ; peu importe que vous désiriez la réciter, la dactylographier ou l'écrire à la main. Remarquez qu'il n'est pas nécessaire de mémoriser le discours deux fois, une pour le prononcer et l'autre pour l'écrire. Une seule mémoire du discours peut adopter diverses formes comportementales. A chaque région, un pattern invariant peut bifurquer et descendre par un chemin différent.

Pour plus d'efficacité, les représentations des objets simples, en bas de la hiérarchie, peuvent être réutilisées à volonté pour des séquences situées à différents niveaux. Par exemple, nous n'avons pas à apprendre un ensemble de mots pour le discours de Gettysburg et un autre ensemble de mots complètement différents pour le discours de Martin Luther King, « J'ai fait un rêve », bien que dans les deux textes l'on trouve quelques mots communs. Une hiérarchie de séquences imbriquées autorise le partage et la réutilisation des objets de bas niveau, comme des mots, des phonèmes et des lettres, pour n'en citer que quelques-uns. C'est un moyen extrêmement efficace pour stocker l'information concernant le monde et ses structures, très différent du fonctionnement d'un ordinateur.

La même décomposition de séquences se produit dans les régions sensorielles et motrices. Le processus permet de percevoir et de comprendre les objets sous différents angles. Quand vous allez vers le réfrigérateur pour prendre une crème glacée, votre cortex visuel est actif à plusieurs niveaux. Au niveau le plus élevé, vous percevez la constante « réfrigérateur ». Dans les régions inférieures, cette attente visuelle est décomposée en une série d'inputs visuels plus localisés. La vue du réfrigérateur est composée de fixations oculaires sur la poignée de la porte, sur le bac à glace, sur les « magnets » qui ornent la porte, sur un dessin d'enfant collé dessus, et ainsi de suite. Lors des quelques millisecondes qui s'écoulent lorsque les saccades se fixent tour à tour sur

les divers éléments du réfrigérateur, des prédictions concernant le résultat de chaque fixation cascadent le long de la hiérarchie visuelle. Tant que ces prédictions sont confirmées d'une saccade à une autre, les régions visuelles supérieures sont satisfaites du regard que vous portez sur le réfrigérateur. Remarquez que dans ce cas, contrairement à l'ordre immuable des mots du discours de Gettysburg, la séquence que vous voyez en regardant le réfrigérateur n'est pas fixe ; le flux des inputs et les patterns de mémoire extraits dépendent de vos propres actions. Ainsi donc, dans un tel cas, le pattern qui a été décomposé n'est pas une séquence rigide, mais le résultat final est le même : des patterns de haut niveau variant lentement, qui se décomposent en patterns de bas niveau rapidement changeants.

La manière par laquelle vous mémorisez des séquences et les représentez par des noms, tandis que l'information monte et descend dans votre hiérarchie corticale, peut être comparée à des ordres militaires transmis hiérarchiquement. Le général dit : « Déplacez les troupes vers la Floride pour les quartiers d'hiver. » Ce simple ordre de haut niveau est décomposé en une séquence d'ordres plus détaillés en descendant dans la hiérarchie. Les subordonnés du général reconnaissent que l'ordre exige une séquence de phases comme la préparation à lever le camp, le transport vers la Floride et les préparatifs à l'arrivée. Chacune de ces phases se décompose en étapes plus spécifiques confiées à des subordonnés. Tout en bas de la hiérarchie, des milliers de soldats entreprennent des dizaines de milliers d'actions dont résulte le déplacement des troupes. Des rapports sont établis à chaque niveau. En remontant dans la hiérarchie, ils font l'objet d'une synthèse, puis de divers résumés, jusqu'à ce que tout en haut de la hiérarchie, le général reçoive une brève note disant : « Le déplacement vers la Floride se déroule comme prévu. » Le général n'a pas besoin des détails.

Il existe une exception à cette règle. Si un incident se produit qui ne peut pas être pris en charge par les subordonnés, le

problème monte dans la hiérarchie jusqu'à ce que quelqu'un sache ce qu'il faut faire. L'officier capable de résoudre le problème ne considère pas l'incident comme une exception. Ce qui était un problème qui, aux yeux des subordonnés, n'a pas été anticipé n'est pour lui que la prochaine tâche attendue. L'officier donne alors de nouveaux ordres aux subordonnés. Le néocortex se comporte de la même façon. Comme nous le verrons d'ici peu, lorsque des événements — autrement dit, des patterns — se produisent, qui n'ont pas été anticipés, l'information à leur sujet remonte dans la hiérarchie corticale jusqu'à ce qu'une région parvienne à les résoudre. Si les régions basses du cortex ne parviennent pas à prédire quels patterns elles voient, elles les considèrent comme des erreurs et les font monter dans la hiérarchie. Ceci se répète jusqu'à ce qu'une région parvienne à anticiper le pattern.

---

De par leur conception, chaque région corticale s'efforce de stocker des séquences et de s'en rappeler. Mais cette description du cerveau est encore trop simpliste. Nous devons ajouter un brin de complexité au modèle.

Les inputs ascendants vers une région du cortex sont des patterns d'entrée acheminés par des milliers ou des millions d'axones. Ces derniers proviennent de différentes régions et contiennent toutes sortes de patterns. Le nombre de patterns susceptibles de s'inscrire sur un millier d'axones est plus élevé que le nombre de molécules dans l'univers. Au cours d'une vie, une région ne verra qu'une quantité infime des patterns possibles.

Voici la question : quand une seule région stocke des séquences, de quoi sont-elles faites ? La réponse est que la région classe d'abord les inputs selon un nombre limité de possibilités, puis elle y cherche des séquences. Mettez-vous dans la peau d'une seule région corticale. Votre tâche consiste à trier des bouts de papier coloré. Devant vous se trouvent dix godets avec, collé sur

chacun d'eux, un échantillon de couleur. Il y a un godet pour le vert, un pour le jaune, un autre pour le rouge, et ainsi de suite. On vous remet les bouts de papier coloré, un par un, que vous devez trier selon leur couleur. Chaque papier est légèrement différent. Comme il existe un nombre infini de couleurs, deux bouts de papier n'ont jamais exactement la même teinte. Parfois, il est facile de savoir dans quel godet il faut le placer, mais parfois c'est moins évident. Un papier entre le rouge et l'orange peut être mis dans l'un ou l'autre des deux godets, mais vous devez faire un choix, même s'il vous paraît quelque peu hasardeux (l'exercice consiste à montrer que le cerveau doit obligatoirement classer des patterns. Les régions du cortex le font, sauf qu'il n'y a rien d'équivalent à des godets pour y loger les patterns).

Maintenant, on vous demande en plus de rechercher des séquences. Vous remarquez que la séquence rouge-rouge-vert-pourpre-orange-vert apparaît fréquemment. Vous l'appelez « séquence rrvpov ». Remarquez que la reconnaissance des séquences serait impossible si vous n'aviez pas préalablement classé chaque bout de papier. Sans les avoir répartis dans l'une des dix catégories, vous ne pourriez affirmer que deux séquences sont identiques.

Vous voilà à pied d'œuvre, prêt pour la tâche. Vous examinerez tous les patterns d'entrée — les morceaux de papier coloré provenant des régions corticales inférieures —, vous les classerez puis vous rechercherez les séquences. Les deux étapes de classification et de formation de séquences sont indispensables pour créer des représentations invariantes ; chaque région du cortex le fait.

Le processus de formation de séquences est avantageux lorsqu'un input est ambigu, comme c'est le cas du bout de papier entre le rouge et l'orange. Vous devez choisir un godet même si vous n'êtes pas sûr que ce soit le rouge ou l'orange qui l'emporte. Si vous connaissez la séquence la plus probable pour cette série d'inputs, vous pourrez utiliser ce savoir pour décider comment classer les inputs ambigus. Si vous pensez qu'une séquence

« rrvpov » est en cours parce que vous avez reçu deux rouges, un vert et un pourpre, vous vous attendez à ce que le prochain papier soit orange. Or, ce papier qui arrive n'est pas orange. Sa couleur se situe plutôt entre le rouge et l'orange. Il serait même plutôt rouge. Mais il vous est familier, et comme vous attendez une séquence « rrvpov », vous placez le papier dans le godet orange. Vous vous servez du contexte des séquences connues pour résoudre une ambiguïté.

Ce phénomène se produit tout le temps dans notre vie quotidienne. Quand des gens parlent, les mots qu'ils disent ne peuvent souvent pas être compris s'ils sont extraits du contexte. Pourtant, quand vous entendez un mot ambigu dans une phrase, vous ne butez pas sur son ambiguïté, vous le comprenez. De même, des mots manuscrits sont souvent incompréhensibles lorsqu'ils sont hors de leur contexte, mais parfaitement clairs dans la phrase complète. La plupart du temps, vous ne vous rendez pas compte que vous complétez une information ambiguë ou partielle à partir d'une mémoire contenant des séquences mémorisées. Vous entendez ce que vous vous attendez à entendre et vous voyez ce que vous vous attendez à voir, du moins lorsque ce que vous écoutez ou ce que vous regardez correspond à une expérience passée.

Remarquez aussi que les séquences mémorisées permettent non seulement de résoudre une ambiguïté de l'input courant, mais aussi de prédire l'input qui doit se produire ensuite. Pendant que la région corticale dont vous jouez le rôle trie des papiers colorés, vous pouvez dire au personnage Input qui vous donne chaque papier : « Dis donc, si jamais tu ne sais pas ce qu'il faut me donner ensuite, d'après ma mémoire, ça devrait être un papier orange. » En reconnaissant une séquence de patterns, la région corticale prédit le prochain pattern d'entrée et indique à la région en dessous à quoi elle doit s'attendre.

Une région du cortex apprend non seulement des séquences familières, mais aussi à modifier ses classifications. Supposons

que vous commenciez avec un ensemble de godets étiquetés « vert », « jaune », « rouge », « pourpre » et « orange ». Vous êtes prêt à reconnaître la séquence « rrvpov » ainsi que d'autres combinaisons de ces couleurs. Comment ferez-vous si une teinte se révèle trop différente ? Que ferez-vous si, chaque fois que vous rencontrez la séquence « rrvpov », le pourpre tire trop sur le violet ? Cette couleur étant plus proche de l'indigo, vous réétiquetterez le godet du pourpre en « indigo ». A présent, les godets correspondent mieux à ce que vous voyez. Vous avez réduit l'ambiguïté. Le cortex est flexible.

Dans les régions corticales, les classifications bas-haut et les séquences haut-bas interagissent constamment et changent tout au long de votre vie. C'est l'essence de l'apprentissage. A vrai dire, toutes les régions du cortex sont façonnables ; c'est pourquoi elles peuvent être modifiées par l'expérience. C'est par la formation de nouvelles classifications et de nouvelles séquences que vous vous souvenez.

Pour finir, voyons comment ces classifications et ces prédictions interagissent avec la région juste au-dessus. Une autre partie de notre tâche corticale étant de transmettre au prochain niveau supérieur le nom de la séquence que vous voyez, vous lui passez un bout de papier avec les lettres « rrvpov » écrites dessus. Par elles-mêmes, ces lettres ne signifient pas grand-chose pour la prochaine région. Le nom n'est qu'un pattern qui doit être combiné avec d'autres inputs, classé puis placé dans une séquence d'ordre plus élevé. Comme vous, la région supérieure conserve une trace des séquences qu'elle voit. A un certain point, elle vous dira peut-être : « Dis donc, si jamais tu ne sais pas ce qu'il faut me donner ensuite, d'après ma mémoire, je prédis que ça devrait être une séquence rrvpov. » C'est une injonction manifeste de ce que vous devez rechercher dans votre propre flux d'inputs. Vous ferez de votre mieux pour interpréter ce que vous voyez selon cette séquence.

L'expression *classification de patterns* est apparentée à l'expression « classification de forme » utilisée dans les domaines

de l'intelligence artificielle et de la vision des machines. Pour que des machines puissent reconnaître des objets, les roboticiens créent un gabarit — disons l'image d'une tasse, ou la forme prototypique d'une tasse —, puis ils apprennent à la machine à faire correspondre ses données en entrée avec la tasse. Si une concordance proche est découverte, l'ordinateur considère qu'il a trouvé la tasse. Mais notre cerveau ne possède pas de gabarit comme celui-ci, et les patterns que chaque région corticale reçoit en entrée ne sont pas du tout des images. Vous ne mémorisez pas des instantanés photographiques de ce qui s'est inscrit sur la rétine, ou des instantanés des patterns produits par la cochlée ou la peau. La hiérarchie du cortex s'assure que les mémorisations des objets sont réparties sur toute la hiérarchie ; elles ne sont pas stockées à un unique endroit. Donc, comme chaque région de la hiérarchie forme des mémoires invariantes, ce qu'une région donnée du cortex apprend, ce sont des séquences de représentations invariantes. Vous ne trouverez jamais l'image d'une tasse ou de tout autre objet stockée telle quelle dans le cerveau.

Contrairement à la mémoire d'un appareil photo numérique, votre cerveau se souvient du monde tel qu'il est et non tel qu'il apparaît. Quand vous pensez au monde, vous vous souvenez de séquences de patterns correspondant à ce que les objets sont dans le monde et comment ils se comportent, et non comment ils apparaissent par le truchement d'un sens particulier à un moment donné. Les séquences au travers desquelles vous appréhendez les objets du monde reflètent la structure invariante du monde lui-même. L'ordre dans lequel vous appréhendez des éléments du monde est déterminé par la structure du monde. Par exemple, vous pouvez entrer directement dans un avion de ligne en empruntant la passerelle, mais jamais à partir du comptoir d'enregistrement. La séquence par laquelle le monde se présente à vous *est* la structure réelle du monde, et c'est de cela que le cortex tient à se souvenir.

N'oubliez pas, toutefois, qu'en propageant le pattern jusqu'en bas de la hiérarchie, une représentation invariante dans n'importe

quelle région du cortex peut être transformée en une prédiction détaillée de la manière dont elle apparaîtra à nos sens. De même, une représentation invariante du cortex moteur peut être transformée en commandes motrices détaillées, spécifiques d'une situation, en propageant le pattern jusqu'en bas de la hiérarchie.

## À QUOI RESSEMBLE UNE RÉGION DU CORTEX ?

Nous allons à présent porter notre attention vers une région du cortex, c'est-à-dire l'un des rectangles visibles à la Figure 5. Plus loin, la Figure 6 montre l'une d'elles plus en détail. Mon intention est de vous montrer comment les cellules d'une région du cortex peuvent apprendre et se souvenir de séquences de patterns, ce qui est l'élément le plus essentiel pour former des représentations invariantes et faire des prédictions. Nous commencerons par décrire à quoi ressemble une région corticale, et comment elles sont toutes réunies. Leur taille varie considérablement ; les plus grandes sont les aires sensorielles primaires. Par exemple, celle de V1, située dans la partie occipitale du cerveau, est d'environ $9 \times 12$ cm. Mais comme je l'ai exposé précédemment, cette aire est en réalité composée de nombreuses régions plus petites dont la taille n'excède pas quelques millimètres carrés. Pour le moment, nous présumerons que la taille d'une aire corticale typique est celle d'une petite pièce de monnaie.

Souvenez-vous des six cartes à jouer évoquées au Chapitre 3, dont chacune représente une couche différente du tissu cortical. Pourquoi ce terme de « couche » ? Si vous placez un fragment de cortex sous un microscope, vous découvrirez que la densité et la forme des cellules varient de haut en bas. Ce sont ces différences qui définissent les couches. Celle d'au-dessus, appelée couche 1, est la plus reconnaissable des six. Comprenant très peu de cellules, elle est principalement constituée d'un tapis d'axones s'étendant parallèlement à la surface corticale. Les couches 2 et 3 lui ressemblent beaucoup. Elles contiennent une forte densité de cellules pyramidales. La couche 4 contient des cellules en forme d'étoiles, et la

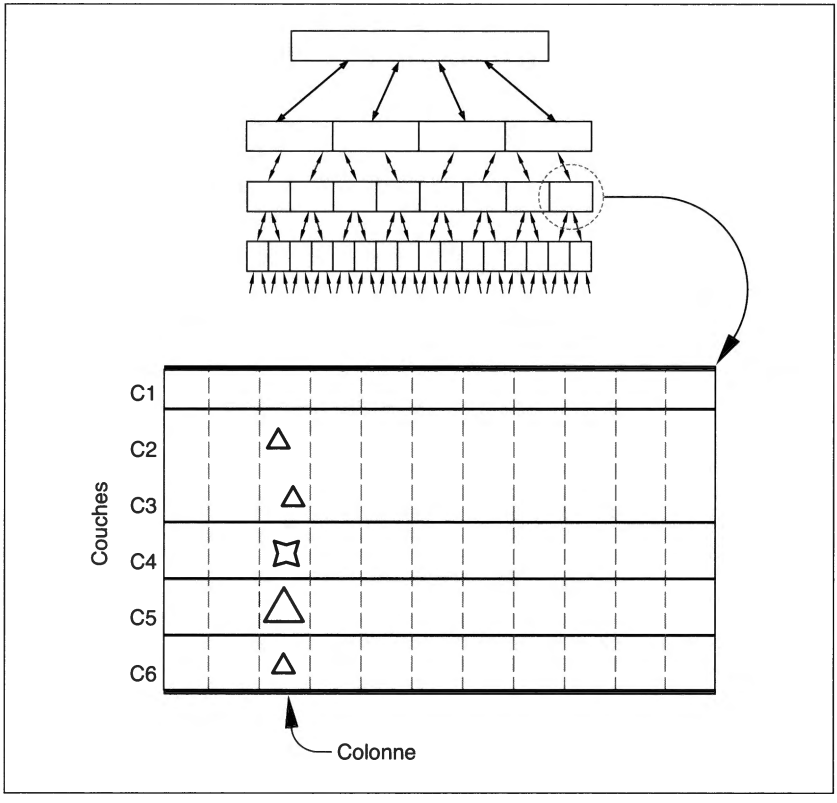



**Figure 6 : Couches et colonnes d'une région du cortex.**

couche 5 des cellules pyramidales normales ainsi qu'une variété de cellules géantes elles aussi pyramidales. Tout en bas, la couche 6 se caractérise par plusieurs types de neurones uniques.

Visuellement, nous plaçons les couches à l'horizontale, mais le plus souvent, les scientifiques s'intéressent à des colonnes de cellules perpendiculaires aux couches. Considérez ces colonnes comme des « unités » verticales de cellules agissant ensemble. Sachez aussi que la notion de *colonne* est très discutée parmi les neurobiologistes. Leur taille, leur fonction et leur importance prêtent à controverse. Mais dans le cadre de cet ouvrage, rien ne vous empêche de penser en termes de colonne architecturale.

A l'intérieur de chaque colonne, les couches sont reliées par des axones qui montent et qui descendent, formant des synapses sur toute leur longueur. Les colonnes ne sont pas parfaitement délimitées — rien, dans le cortex, n'est simple —, mais leur existence peut être déduite à partir de plusieurs évidences.

Une bonne raison de parler de colonnes est que, dans chacune d'elles, les cellules verticalement alignées sont enclines à être activées par un même stimulus. Si nous observons de près les colonnes en V1, nous découvrons que certaines réagissent à des droites, ou segments de ligne, penchées dans une direction (/) tandis que d'autres réagissent à des droites penchées dans une autre direction (\). Les cellules à l'intérieur de chaque colonne sont fortement connectées ; c'est pourquoi la colonne entière réagit à un même stimulus.

Spécifiquement, une cellule active dans la couche 4 fait que les cellules situées au-dessus d'elle, dans les couches 3 et 2, s'activent également, ce qui entraîne aussi l'activation des cellules en dessous, dans les couches 5 et 6. A l'intérieur d'une colonne de cellules, l'activité se propage à la fois vers le haut et vers le bas.

Une autre bonne raison de parler de colonnes découle de la manière dont le cortex se forme. Dans un embryon, des cellules précurseurs isolées migrent à partir d'une cavité interne du cerveau jusqu'à l'endroit où le cortex se formera. Chacune de ces cellules se divise pour créer des éléments d'une centaine de neurones environ, c'est-à-dire des microcolonnes interconnectées verticalement de la manière que je viens d'expliquer. Le terme *colonne* est souvent utilisé assez librement pour décrire différentes manifestations : la connectivité verticale générale ou des groupes de cellules spécifiques issues du même géniteur. En ce qui concerne cette dernière définition, nous pouvons dire que le cortex humain comprend au bas mot plusieurs centaines de millions de microcolonnes.

Pour mieux visualiser cette structure colonnaire, imaginez une seule microcolonne du diamètre d'un cheveu. Prenez un

millier de cheveux et coupez-les en tout petits morceaux de deux millimètres de long. Tassez bien tous ces cheveux ou colonnes, serrez-les bien fort et collez-les ensemble pour en faire une sorte de brosse extrêmement dense. Confectionnez ensuite une feuille faite de longs cheveux très fins — qui représentent les axones de la couche 1 — et collez-la par-dessus le tapis de cheveux ras. Ce bricolage capillaire est un modèle très simpliste d'une région corticale de la taille d'une pièce de monnaie. L'information circule principalement dans la direction des cheveux : à l'horizontale dans la couche 1, à la verticale dans les couches 2 à 6.

Après vous avoir fait part d'un autre détail concernant les colonnes que vous devez absolument connaître, nous découvrirons à quoi elles servent. Un examen de près révèle qu'au moins 90 % des synapses des cellules, dans chaque colonne, viennent d'en dehors de la colonne elle-même. Certaines connexions sont issues des colonnes voisines, d'autres de lieux situés à mi-parcours sur le cerveau. Dans ce cas, comment pouvons-nous parler de l'importance des colonnes si une telle part du câblage cortical s'étend littéralement sur de si vastes zones ?

La réponse réside dans le modèle de mémoire-prédiction. En 1979, lorsque Vernon Mountcastle avait argué qu'il existe un algorithme cortical unique, il avait aussi émis l'hypothèse que la colonne corticale est l'unité de base des computations — ou calculs — intracorticales. Il ne savait cependant pas quelle fonction exécute une colonne. Il pensait que cette dernière est l'unité de base de la prédiction. Pour qu'une colonne puisse prédire quand elle doit être active, elle doit savoir ce qui se passe ailleurs, d'où les connexions synaptiques ici et là-bas.

Nous entrerons d'ici peu dans les détails. Mais auparavant, voici une brève explication de la nécessité de cette sorte de câblage dans le cerveau. Pour prédire la prochaine note d'une chanson, vous devez connaître le nom de cette chanson, où vous en êtes dans son écoute, la durée écoulée depuis la dernière note, et ce qu'était cette dernière note. Le grand nombre des synapses

connectant les cellules d'une colonne à d'autres parties du cerveau fournit à chaque colonne le contexte dont elle a besoin pour prédire son activité dans différentes situations.

---

Le prochain point auquel nous devrons réfléchir est comment ces régions corticales de la taille d'une petite pièce de monnaie — et leurs colonnes — envoient et reçoivent l'information montant et descendant le long de la hiérarchie corticale. Nous examinerons d'abord le flux montant qui emprunte la route relativement directe représentée à la Figure 7. Imaginez que nous examinons une région corticale et ses dizaines de colonnes. Nous zoomons sur l'une de ces dernières. Les inputs convergents parvenant des régions plus basses arrivent toujours à la couche 4, la couche d'entrée principale. En passant, ils forment aussi une connexion dans la couche 6 (nous verrons plus tard en quoi c'est important). Les cellules de la couche 4 envoient ensuite des projections vers le haut, vers les couches 2 et 3 des mêmes colonnes. Quand une colonne projette des informations vers le haut, beaucoup de cellules des couches 2 et 3 étendent des axones vers la couche d'entrée de cette région située juste au-dessus. De ce fait, l'information circule de région en région en remontant dans la hiérarchie.

L'information qui descend le long de la hiérarchie corticale emprunte un chemin moins direct, comme le révèle la Figure 8. Les cellules de la couche 6 sont des cellules de sortie qui envoient des projections vers le bas à partir d'une colonne corticale, et des projections vers la couche 1 dans des régions hiérarchiquement inférieures. Ici, dans la couche 1, l'axone s'étend sur de grandes distances dans la région corticale inférieure. De ce fait, l'information qui descend le long de la hiérarchie à partir d'une colonne est potentiellement capable d'activer beaucoup de colonnes dans les régions qui se trouvent en dessous. Il y a très peu de cellules dans la couche 1, mais celles des couches 2, 3 et 5 ont des dendrites dans cette couche 1, ce qui permet à ces cellules d'être excitées par le

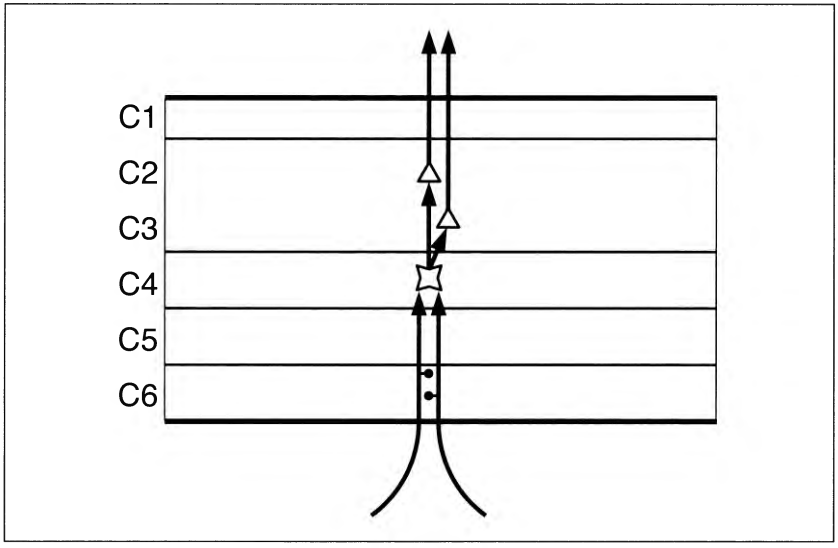



**Figure 7 : Flux d'informations montant à travers une région du cortex.**

feedback se propageant à travers toute la couche 1. Les axones en provenance des cellules des couches 2 et 3 forment des synapses dans la couche 5 au moment où elles quittent le cortex ; on pense qu'elles excitent les cellules des couches 5 et 6. Nous pouvons donc dire qu'une information qui descend le long de la hiérarchie suit un chemin moins direct. Elle peut se ramifier dans beaucoup de directions dès sa dispersion à la couche 1. L'information en feedback commence dans une cellule de la couche 6, dans la région supérieure. Elle se propage à travers la couche 1, dans la région inférieure. Certaines cellules des couches 2, 3 et 5, dans la région inférieure, sont excitées, et certaines d'entre elles excitent les cellules de la couche 6, qui envoient des projections vers la couche 1 dans des régions hiérarchiquement inférieures, et ainsi de suite (ce processus est plus limpide en le suivant à la Figure 8).

Convertir une représentation invariante en une prédiction spécifique exige de pouvoir décider, moment après moment, par quel moyen envoyer le signal tandis qu'il se propage vers le bas de la hiérarchie. La couche 1 offre un moyen de convertir une

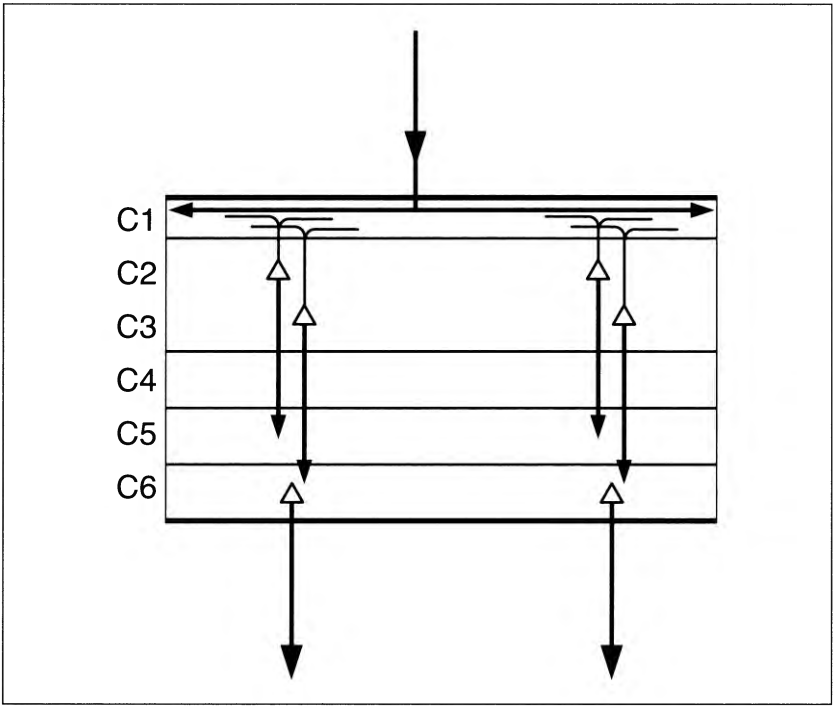



**Figure 8 : Flux d'informations descendant à travers une région du cortex.**

---

représentation invariante en une représentation plus détaillée et plus spécifique. Rappelez-vous qu'il est possible de se souvenir du discours de Gettysburg soit oralement, soit en l'écrivant. Une représentation commune se déplace le long de l'un des deux chemins, l'un pour la voix, l'autre pour l'écriture. De même, quand j'entends la prochaine note d'une mélodie, mon cerveau doit prendre un intervalle générique, comme une quinte, et le convertir dans la note spécifique correcte, comme un do ou un sol. Le flux horizontal d'activité à travers la couche 1 produit le mécanisme nécessaire à cette fin. Pour que des prédictions invariantes de haut niveau puissent se propager jusqu'en bas du cortex et devenir des prédictions spécifiques, nous devons disposer d'un mécanisme autorisant le flux de patterns à se ramifier à chaque niveau. La couche 1 se prête bien à cela. Nous pourrions

prédire que nous en avons besoin même si nous ne savions pas qu'elle existe.

Un dernier détail anatomique : quand les axones quittent la couche 6 pour s'étendre vers d'autres destinations, ils sont gainés d'une substance grasse, la myéline. Ce que l'on appelle la « matière grise » s'apparente aux gaines isolantes qui protègent l'installation électrique d'une habitation. Elle empêche les signaux de se mêler et les fait voyager plus vite, à une vitesse de plus de 300 km/h. Quand les axones quittent la matière grise, ils entrent dans une nouvelle colonne corticale, à la couche 6.

---

Enfin, les régions corticales communiquent entre elles autrement, de façon indirecte.

Avant d'entrer dans les détails, je tiens à rappeler la notion de mémoire auto-associative exposée au Chapitre 2. Comme vous vous en souvenez sans doute, des mémoires auto-associatives peuvent servir à stocker des séquences de patterns. Quand l'output d'un groupe de neurones artificiels est rétropropagé (feedback) pour former l'input de tous les neurones, et que ce feedback est retardé, les patterns apprennent à se succéder en séquence. Je crois que le cortex utilise ce même mécanisme élémentaire pour stocker des séquences, bien que d'une manière un peu plus alambiquée. Au lieu de former une mémoire auto-associative à partir de quelques neurones, il la forme à partir de colonnes corticales. L'output de toutes les colonnes est rétropropagé vers la couche 1. De cette manière, cette dernière contient des informations indiquant quelles colonnes étaient juste actives dans la région du cortex.

Parcourons ces éléments que montre la Figure 9. On sait depuis des années que les cellules particulièrement grandes de la couche 5, dans le cortex moteur (M1), établissent un contact direct avec les muscles et les régions motrices de la moelle épinière. Ces cellules commandent littéralement les muscles,

produisant les mouvements. Chaque fois que vous parlez, tapez du texte ou exécutez quelque comportement sophistiqué, ces cellules sont activées et désactivées d'une façon très coordonnée, entraînant la contraction des muscles.

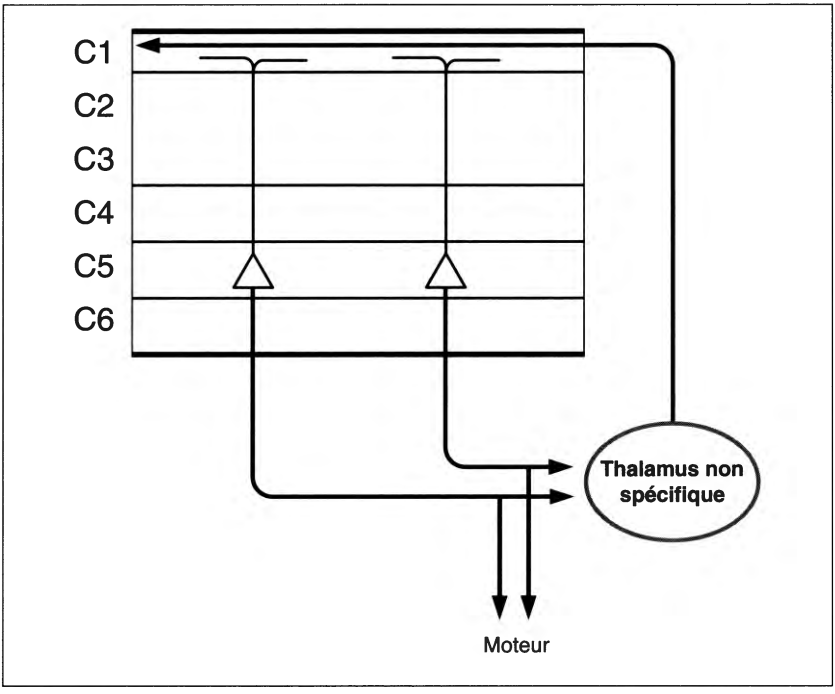



**Figure 9 : Comment l'état courant et le comportement moteur courant communiquent largement entre eux *via* le thalamus.**

Des chercheurs ont récemment découvert que les cellules géantes de la couche 5 sont susceptibles de jouer un rôle primordial dans le comportement d'autres parties du cortex, et pas uniquement dans les régions motrices. Par exemple, ces cellules géantes de la couche 5 situées dans le cortex visuel envoient des projections dans la partie du cerveau qui met les yeux en mouvement. Ainsi, les aires sensorielles visuelles du cortex, comme V2 et V4, traitent non seulement les inputs visuels, mais contribuent à déterminer le mouvement des yeux lui-même, et par conséquent ce que vous

voyez. Les grandes cellules de la couche 5 se trouvent dans tout le néocortex, dans chaque région, laissant à penser qu'elles jouent un rôle prédominant dans toutes sortes de mouvements.

Outre leur rôle comportemental, les axones des grandes cellules de la couche 5 sont scindés en deux. Une branche s'étend vers une partie du cerveau appelée thalamus, représenté à la Figure 9. Le thalamus humain a la forme et la taille de deux œufs de moineau. Il se trouve au centre du cerveau, au-dessus du cerveau archaïque, enveloppé par la matière grise et le cortex.

Le thalamus reçoit de nombreux axones provenant de diverses parties du cerveau et en renvoie vers ces mêmes aires. Ces connexions sont bien connues, mais le thalamus lui-même est une structure compliquée dont le rôle n'est pas très clair. Il est cependant essentiel, car s'il est endommagé le sujet est réduit à une existence végétative.

Deux fibres relient le thalamus au cortex, mais une seule nous intéresse. Elle part des grandes cellules de la couche 5 qui envoient des projections vers une classe de cellules thalamiques dites « non spécifiques ». Les cellules non spécifiques renvoient des axones vers la couche 1 à travers un grand nombre de régions différentes du cortex. Par exemple, les cellules de la couche 5 réparties dans les régions V2 et V4 étendent des axones vers le thalamus, et le thalamus renvoie des informations à la couche 1 à travers V2 et V4. D'autres parties du cortex en font autant. Les cellules de la couche 5 que l'on trouve dans de nombreuses régions corticales envoient des projections vers le thalamus, qui renvoie des informations vers la couche 1 à travers ces mêmes régions associées. Je pense que ce circuit est exactement comme les feedbacks retardés qui permettent aux modèles de mémoire auto-associative d'apprendre les séquences.

J'ai mentionné deux inputs vers la couche 1. Les régions plus hautes du cortex diffusent l'activité dans toute la couche 1, dans les régions plus basses. Les colonnes actives à l'intérieur d'une région diffusent elles aussi une activité à travers la couche 1, dans

la même région, *via* le thalamus. Nous pouvons imaginer que ces inputs vers la couche 1 peuvent être le titre d'une chanson (input par le dessus) et où nous en sommes dans l'écoute de la chanson (activité retardée des colonnes actives dans la même région). De ce fait, la couche 1 véhicule la majorité de l'information dont nous avons besoin pour prédire quand une colonne devrait être active, c'est-à-dire le nom de la séquence et où nous en sommes dans la séquence. En recourant à ces deux signaux dans la couche 1, une région du cortex peut apprendre et se souvenir de multiples séquences de patterns.

## LE FONCTIONNEMENT D'UNE RÉGION CORTICALE

Ces trois circuits en tête — des patterns convergents remontant la hiérarchie corticale, des patterns divergents descendant la hiérarchie corticale et un feedback retardé à travers le thalamus —, nous commençons à voir comment une région du cortex exécute les fonctions dont elle a besoin. Ce que nous savons se réduit à ces questions :

1. Comment une région du cortex classe-t-elle ses inputs (souvenez-vous des godets) ?
2. Comment apprend-t-elle les séquences de patterns (comme les intervalles d'une mélodie ou la structure « yeux nez yeux » d'un visage) ?
3. Comment forme-t-elle un pattern constant, c'est-à-dire le « nom » d'une séquence ?
4. Comment effectue-t-elle des prédictions spécifiques (être à l'heure à l'arrivée du train, prédire une note spécifique d'une mélodie) ?

Commençons par présumer que les colonnes d'une région du cortex sont comme les godets que nous avions utilisés pour classer des inputs symbolisés par des papiers colorés. Chaque colonne représente l'étiquette d'un godet. Dans chaque colonne, les cellules de la couche 4 reçoivent des fibres d'entrée provenant

de diverses régions situées dessous, et s'excitent dès qu'elles ont reçu la bonne combinaison d'inputs. Quand les cellules de la couche 4 s'excitent, c'est comme si elles « votaient » pour dire que l'input correspond à l'étiquette. Comme dans l'analogie du tri des bouts de papier, les inputs peuvent être ambigus, de sorte que plusieurs colonnes sont susceptibles de correspondre à l'input. Or, la région du cortex doit pouvoir choisir une seule et unique interprétation : le papier est soit rouge, soit orange, mais pas les deux à la fois. Une colonne recevant un input fort doit absolument empêcher l'excitation des autres cellules.

Le cerveau est doté de cellules inhibitrices chargées de cette tâche. Elles inhibent les autres neurones du voisinage, ne permettant qu'à un seul d'entre eux d'être l'élu. Ces cellules spéciales affectent la zone entourant une colonne. Ainsi, même si les inhibitions sont nombreuses, beaucoup de colonnes d'une région peuvent néanmoins être actives simultanément (en fait, dans le cerveau, rien n'est jamais représenté par un seul neurone ou par une seule colonne). Pour que cela soit plus clair, considérez que, dans une région, une seule et unique colonne seulement peut être activée. Mais en votre for intérieur, sachez que beaucoup de colonnes peuvent être actives en même temps. Le véritable processus auquel recourt une région du cortex pour classer des inputs, et comment elle apprend à le faire, est compliqué et n'est pas encore complètement élucidé. Je ne vous infligerai pas les détails. Je préfère que vous partiez du principe que les régions du cortex ont classé leurs inputs sous forme d'activité dans un ensemble de colonnes. Nous pourrons alors nous concentrer sur la formation des séquences et comment elles sont nommées.

Comment notre région corticale stocke-t-elle la séquence de ces patterns classifiés ? J'ai déjà la réponse à cette question, mais nous allons à présent entrer dans les détails. Imaginons que vous êtes une colonne de cellules, et qu'un input provenant d'une région inférieure excite l'une des cellules de la couche 4. Vous êtes

content, et la cellule de la couche 4 entraîne à son tour l'excitation des cellules dans les couches 2 et 3, puis 5 et ensuite 6. Commandée par en dessous, la colonne tout entière devient active. Les cellules des couches 2, 3 et 5 possèdent des milliers de synapses dans la couche 1. Si l'une d'elles est active lorsque les cellules des couches 2, 3 et 5 sont excitées, les synapses sont renforcées. Si cela se produit fréquemment, ces synapses de la couche 1 deviennent suffisamment fortes pour entraîner l'excitation des cellules des couches 2, 3 et 5, et cela même si une cellule de la couche 4 ne s'est pas excitée. Cela signifie qu'une partie de la colonne peut devenir active sans recevoir un input d'une région plus basse du cortex. De cette manière, les cellules des couches 2, 3 et 5 apprennent à « anticiper » le moment où elles doivent s'exciter en se fondant sur le pattern présent dans la couche 1. Avant d'apprendre, la colonne ne peut devenir active que si elle est commandée par une cellule de la couche 4. Après l'apprentissage, la colonne peut devenir partiellement active *via* la mémoire. Quand une colonne devient active *via* les synapses de la couche 1, elle est par anticipation commandée d'en dessous. Ceci est une prédiction. Si la colonne pouvait parler, elle dirait : « Quand j'ai été active dans le passé, cet ensemble particulier de synapses de ma couche 1 était actif. Par conséquent, si je vois de nouveau cet ensemble particulier, je m'excite à l'avance. »

Rappelez-vous que la moitié de l'input vers la couche 1 provient des cellules de la couche 5, situées dans les colonnes et régions corticales voisines. Cette information représente ce qui s'était passé dans les moments précédents. Elle représente les colonnes qui étaient actives avant que votre colonne le devienne. Elle représente l'intervalle précédent d'une mélodie, ou la dernière chose vue, ou la dernière sensation éprouvée, ou le phonème précédent du discours que j'entends. Si l'ordre dans lequel ces patterns se produisent est régulier, les colonnes apprennent cet ordre. Elles s'exciteront les unes après les autres en une séquence correcte.

L'autre moitié de l'input vers la couche 1 provient des cellules de la couche 6, dans des régions hiérarchiquement plus élevées. Cette information est plus stable. Elle représente le nom de la séquence actuellement en cours. Si vos colonnes sont des intervalles musicaux, c'est le nom de la mélodie. Si les colonnes sont des phonèmes, c'est le mot que vous entendez. Si les colonnes sont un mot prononcé, le signal venu d'en haut est le texte que vous récitez. De ce fait, l'information dans la couche 1 représente à la fois le nom d'une séquence et le dernier élément de la séquence. De cette manière, une colonne particulière peut sans confusion être partagée entre beaucoup de séquences différentes. Les colonnes apprennent à s'exciter selon le contexte approprié et dans l'ordre correct.

Avant de poursuivre, je dois préciser que les synapses présentes dans la couche 1 ne sont pas les seules à participer à l'apprentissage, lorsqu'une colonne doit devenir active. Comme je l'ai mentionné précédemment, les cellules reçoivent des inputs de beaucoup de colonnes environnantes, et en envoient vers beaucoup de colonnes environnantes. Rappelez-vous que plus de 90 % de toutes les synapses proviennent de cellules situées hors de la colonne, et la plupart de ces synapses ne sont pas dans la couche 1. Par exemple, les cellules des couches 2, 3 et 5 ont certes des milliers de synapses dans la couche 1, mais aussi des milliers de synapses dans leurs propres couches. L'idée générale est que les cellules recherchent toute information qui les aiderait à prédire à partir de quel moment elles seront commandées d'en dessous. Usuellement, l'activité dans les colonnes proches est fortement corrélée, d'où les nombreuses connexions directes vers les colonnes voisines. Par exemple, une ligne qui se déplace dans le champ visuel active des colonnes successives. Mais bien souvent, l'information nécessaire pour prédire l'activité d'une colonne est plus globale, et c'est là que les synapses de la couche 1 jouent un rôle. Si vous étiez une cellule ou une colonne, vous ne connaîtriez pas la signification de chacune des nombreuses synapses ; tout ce que

vous sauriez c'est qu'elles peuvent vous aider à prédire le moment où vous devriez être activée.

---

Voyons maintenant comment une région du cortex forme le nom d'une séquence apprise. Imaginez une fois de plus que vous êtes une région du cortex. Vos colonnes actives changent à chaque nouvel input. Vous avez réussi à apprendre l'ordre dans lequel vos colonnes deviennent actives, sachant que certaines cellules des colonnes s'activent avant l'arrivée d'inputs en provenance de régions inférieures. Quelle information enverrez-vous vers les régions du cortex hiérarchiquement supérieures ? Nous avons vu précédemment que les cellules de vos couches 2 et 3 envoient leurs axones vers la prochaine région la plus élevée. L'activité de ces cellules est l'input vers les régions plus élevées. Mais un problème se pose. Pour que cette hiérarchie puisse fonctionner, vous devez relayer un pattern constant durant les séquences apprises ; vous devez transmettre le nom d'une séquence, pas les détails. Avant d'apprendre une séquence, vous pouvez transmettre les détails, mais après avoir appris la séquence et être ainsi capable de prédire exactement les colonnes qui seront actives, vous devez relayer un pattern constant. Or, jusqu'à présent, je ne vous ai pas expliqué comment le faire. Tel que cela se présente en ce moment, vous transmettrez chaque pattern changeant, que vous puissiez le prédire ou non. Au fur et à mesure que les colonnes deviennent actives, leurs cellules de la couche 2 et de la couche 3 envoient un nouveau signal vers le haut de la hiérarchie. Le cortex doit disposer d'un moyen de préserver la constance de l'input envoyé vers la prochaine région au cours des séquences apprises. Il nous faut un moyen de neutraliser l'output des cellules de la couche 2 et de la couche 3 lorsqu'une colonne prédit son activité, ou au contraire rendre ces cellules actives lorsque la colonne ne parvient pas à prédire son activité. C'est le seul moyen pour confectionner un pattern de nom constant.

Nous n'en savons pas suffisamment sur le cortex pour expliquer exactement comment il s'y prend. J'ai imaginé plusieurs techniques. Je décrirai celle qui me convient actuellement le mieux, mais gardez à l'esprit que le concept est plus important que telle ou telle technique. La création d'un pattern de « nom » constant est obligatoire pour cette théorie. Tout ce que je peux dire en ce moment est qu'il existe un mécanisme plausible pour le nommage.

Une fois de plus, mettez-vous dans la peau — si l'on peut dire — d'une colonne, comme l'illustre la Figure 10. Nous voulons comprendre comment vous apprenez à présenter un pattern constant à la région immédiatement au-dessus lorsque vous parvenez à prédire votre activité, et un pattern changeant lorsque vous ne pouvez pas la prédire. Commençons par supposer que dans les couches 2 et 3 se trouvent plusieurs classes de cellules (outre plusieurs types de cellules inhibitrices, beaucoup d'anatomistes font la différence entre les types de cellules présentes dans ce qu'ils appellent des couches 3a et 3b ; c'est pourquoi cette hypothèse n'est pas déraisonnable).

Supposons aussi qu'une classe de cellules, appelées « cellules de la couche 2 », apprenne à rester en place au cours des séquences d'apprentissage. Ces cellules, en tant que groupe, représentent le nom de la séquence. Elles présenteront un pattern constant aux régions corticales plus élevées aussi longtemps que votre région sera capable de prédire quelles seront les prochaines colonnes actives. Si la région du cortex possède une séquence apprise composée de trois patterns différents, les cellules de la couche 2 de toutes les colonnes représentant ces trois patterns resteront actives aussi longtemps que nous serons dans cette séquence. Elles sont le nom de la séquence.

Supposons ensuite qu'il existe une autre classe de cellules, les cellules de la couche 3b qui ne s'excitent pas lorsque notre colonne prédit correctement son input, mais qui s'excitent en revanche quand elle ne prédit pas son activité. Une cellule de la couche 3b représente un pattern inattendu. Elle s'excite lorsqu'une colonne

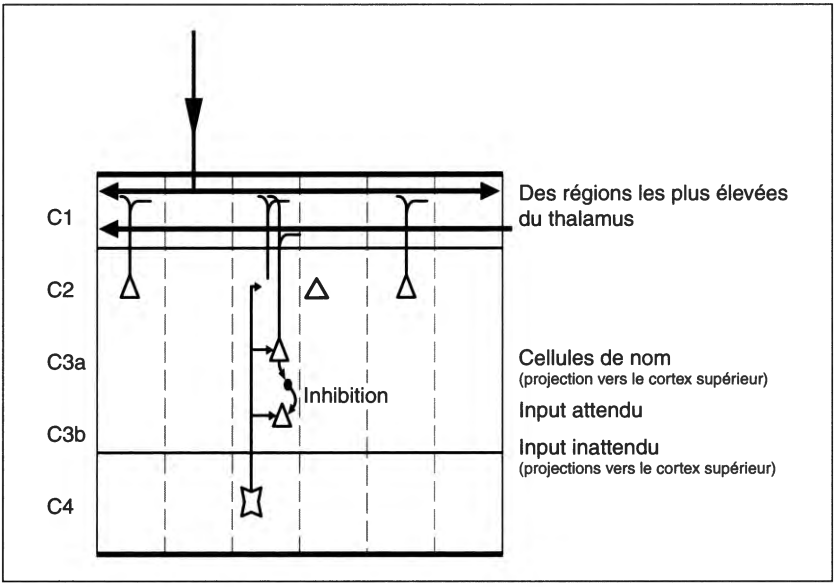



**Figure 10 : Formation du nom constant d'une séquence apprise.**

devient active de manière imprévue. Elle s'excitera chaque fois qu'une colonne devient active avant tout apprentissage préalable. Mais lorsqu'une colonne apprend à prédire son activité, la cellule de la couche 3b se tient tranquille. Les cellules de la couche 2 et de la couche 3b répondent toutes deux à nos exigences. Avant l'apprentissage, les deux cellules s'excitent ou cessent de l'être avec la colonne, mais après l'apprentissage, la cellule de la couche 2 est constamment active et celle de la cellule 3b reste tranquille.

Comment ces cellules apprennent-elles à faire cela ? Commençons d'abord par réfléchir à la manière de désactiver la cellule de la couche 3b lorsque sa colonne prédit correctement son activité. Supposons qu'une autre cellule se trouve juste au-dessus de celle de la couche 3b, c'est-à-dire dans la couche 3a. Cette cellule a aussi des dendrites dans la couche 1. Sa seule tâche est d'empêcher la cellule de la couche 3b de s'exciter lorsqu'elle reconnaît le pattern approprié dans la couche 1. Quand la cellule de la couche 3a voit le pattern appris dans la couche 1, elle active

rapidement une cellule inhibitrice qui empêche l'excitation de la cellule de la couche 3b. Il n'en faut pas davantage pour stopper l'excitation de la cellule de la couche 3b lorsque la colonne prédit correctement son activité.

Considérez à présent la tâche plus difficile consistant à faire en sorte que la cellule de la couche 2 reste active tout au long d'une séquence connue de patterns. C'est plus ardu car un ensemble diversifié de cellules de couche 2 dans différentes colonnes doit rester actif conjointement, même si chacune des colonnes individuelles n'est pas active. Voici comment, à mon avis, cela se produit. Les cellules de la couche 2 peuvent apprendre à être commandées uniquement à partir des régions hiérarchiquement supérieures du cortex. Elles peuvent former des synapses préférentiellement avec les axones des cellules de la couche 6, dans les régions au-dessus. Les cellules de la couche 2 représenteront par conséquent le pattern de nom constant de la région supérieure. Quand une région supérieure du cortex envoie un pattern à la couche 1 située plus bas, un ensemble de cellules de la couche 2, dans la région inférieure, devient actif, représentant toutes les colonnes qui font partie de la séquence. Comme les cellules de la couche 2 rétroprojettent aussi vers les régions plus élevées, elles forment un groupe de cellules semi-stables (il est improbable que ces cellules restent constamment actives. Elles s'excitent probablement de manière synchrone, quelque peu rythmique). C'est comme si les régions plus élevées envoyaient le nom d'une mélodie, par exemple, vers la couche 1, en dessous. Cet événement entraîne l'excitation d'un ensemble de cellules de la couche 2, une pour chacune des colonnes qui s'activent tour à tour au cours de l'audition de la mélodie.

La somme de tous ces mécanismes permet au cortex d'apprendre des séquences, de procéder à des prédictions et de former des représentations constantes, ou « noms », pour les séquences. Ce sont les opérations de base pour la formation des représentations invariantes.

Comment faisons-nous des prédictions au sujet d'événements que nous n'avons jamais connus auparavant ? Comment décidons-nous parmi de multiples interprétations d'un input ? Comment une région du cortex fait-elle des prédictions spécifiques à partir de mémoires invariantes ? J'ai donné des exemples de tout cela précédemment, comme la prédiction de l'exacte prochaine note d'une mélodie lorsque votre mémoire ne se souvient que des intervalles qui les séparent, l'exemple de l'heure d'arrivée du train ou la récitation du discours de Gettysburg. Dans tous ces cas, le seul moyen de résoudre le problème est d'utiliser la dernière information spécifique pour convertir une prédiction invariante en prédiction spécifique. Une autre manière d'exprimer ceci, en termes de cortex, serait de dire qu'il faut combiner des informations d'antéropropagation (l'input réel) en information de rétropropagation (une prédiction dans une forme invariante).

Voici un exemple simple de ce qui, à mon avis, se passe. Supposons qu'une région du cortex ait été informée qu'elle doit s'attendre à un intervalle musical d'une quinte. Les colonnes de cette région représentent tous les intervalles possibles comme do-mi, do-sol, ré-la, etc. Vous devez décider quelles colonnes doivent être actives. Quand la région située au-dessus signale qu'il faut s'attendre à une quinte, il en découle que les cellules de la couche 2 s'excitent dans toutes les colonnes qui sont des quintes, comme sol-do, ré-la et mi-si. Les cellules des colonnes de la couche 2 représentant d'autres intervalles ne sont pas actives. Vous devez à présent sélectionner une colonne parmi toutes les quintes possibles. Les inputs vers votre région sont des notes spécifiques. Si la dernière que vous avez entendue était un ré, toutes les colonnes représentant des intervalles impliquant un ré, comme ré-mi et ré-si, ont un input partiel. Nous avons donc, dans la couche 2, une activité dans toutes les colonnes qui sont des quintes, et dans la couche 4 un input partiel vers toutes les colonnes représentant un intervalle impliquant un ré. L'intersection de ces deux ensem-

bles représente notre réponse, c'est-à-dire la colonne représentant l'intervalle ré-la (voir Figure 11).

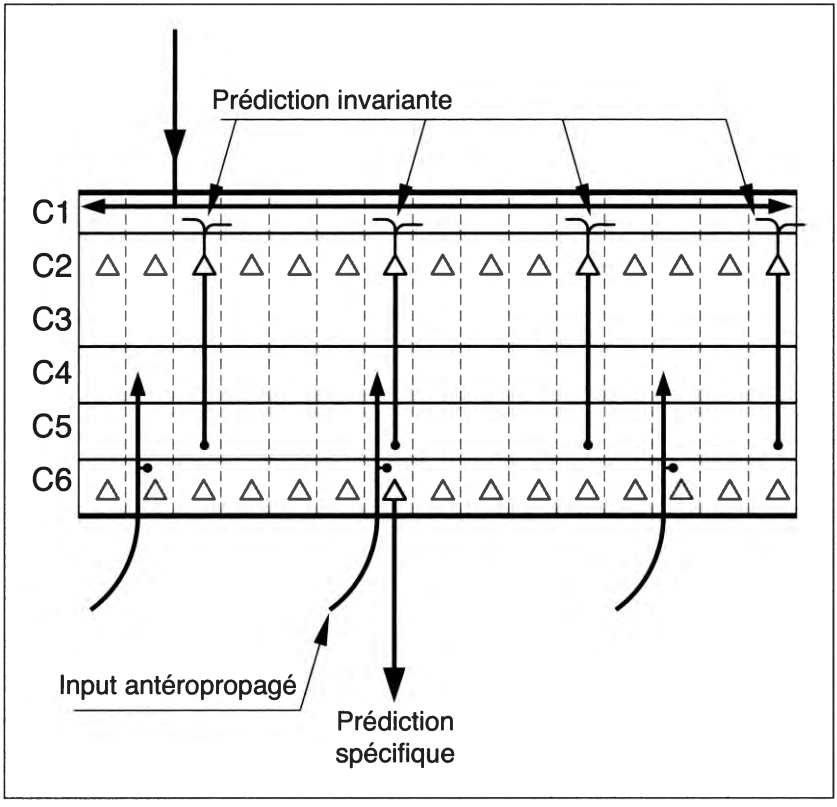



**Figure 11 : Comment une région du cortex procède à des prédictions spécifiques à partir de mémoires invariantes.**

Comment le cortex découvre-t-il cette intersection ? Rappelez-vous que j'ai mentionné le fait que les axones des cellules de la couche 2 et de la couche 3 forment généralement des synapses dans la couche 5 en quittant le cortex, et que des axones semblables s'approchant de la couche 4 à partir de régions inférieures du cortex produisent une synapse dans la couche 6. L'intersection de ces deux synapses (descendante et montante) fournit ce dont nous avons besoin. La cellule de la couche 6 qui reçoit ces deux inputs actifs

sera excitée. Une cellule de la couche 6 représente ce qu'une région du cortex croit qu'il se passe, c'est-à-dire une prédiction spécifique. Si cette cellule de la couche 6 pouvait parler, elle dirait : « Je fais partie d'une colonne qui représente quelque chose. Dans mon cas, c'est l'intervalle musical ré-la. D'autres colonnes peuvent avoir d'autres significations. Je parle de ma région du cortex. Quand je deviens active, cela veut dire que nous croyons que l'intervalle musical ré-la est soit en cours, soit sur le point de se produire. Je pourrais devenir active parce que l'input ascendant provenant des oreilles a incité la cellule de la couche 4 de ma colonne à exciter la colonne tout entière. Ou alors, mon activité pourrait signifier que nous reconnaissons une mélodie et que nous allons prédire son prochain intervalle spécifique. Dans les deux cas, ma tâche consiste à informer les régions inférieures du cortex de ce que nous pensons être en cours. Je représente notre interprétation du monde, qu'elle soit vraie ou seulement imaginée. »

Permettez-moi de décrire cela en recourant à une autre image mentale. Imaginez deux feuilles de papier pleines de petits trous. Ceux de l'une des feuilles représentent les colonnes dont les cellules de la couche 2 ou de la couche 3 sont actives, c'est-à-dire notre prédiction invariante. Les trous de l'autre feuille représentent les colonnes ayant reçu un input partiel de la région d'en dessous. Placez les deux feuilles de papier l'une sur l'autre : certains trous seront parfaitement superposés, d'autres non. Les trous superposés représentent les colonnes que nous pensons être actives.

Ce mécanisme produit non seulement des prédictions spécifiques, mais il résout aussi les ambiguïtés des inputs sensoriels. Très souvent, l'input vers une région sensorielle est ambigu, comme nous l'avons constaté avec les bouts de papier coloré, ou quand nous entendons un mot un peu confus. Ce mécanisme de concordance ascendante/descendante permet de décider entre plusieurs interprétations. Le choix effectué, l'interprétation est relayée vers la région du dessous.

A chaque moment de notre existence quotidienne, chaque région du cortex compare un ensemble de colonnes qui s'attendent à quelque chose, commandées par au-dessus, à un ensemble de colonnes observées commandées par en dessous. L'endroit où les deux ensembles se coupent est ce que nous percevons. Si l'input provenant d'en haut et les prédictions étaient parfaits, l'ensemble des colonnes perçues serait toujours contenu dans l'ensemble des colonnes prédites. Cet accord ne se produit pas souvent. La technique consistant à combiner une prédiction partielle avec un input partiel résout un input ambigu, elle comble un élément d'information manquant et décide entre diverses options. C'est ainsi que nous combinons un intervalle de hauteur de note invariant avec la dernière note entendue afin de prédire la prochaine note spécifique d'une mélodie. C'est ainsi que nous décidons qu'une image montre un vase plutôt que deux visages de profil se faisant face. C'est ainsi que nous divisons le flux moteur afin d'écrire ou de réciter le discours de Gettysburg.

Enfin, outre la projection vers les régions corticales inférieures, les cellules de la couche 6 peuvent renvoyer leur output vers les cellules de la couche 4 de leur propre colonne. Ce faisant, nos prédictions deviennent l'input. C'est ce qui se passe lorsque nous rêvassons ou pensons. Ceci nous permet de connaître les conséquences de nos propres prédictions. Nous procédons ainsi plusieurs fois par jour lorsque nous envisageons l'avenir, répétons le texte d'un discours ou quand nous nous inquiétons d'événements à venir. Stephen Grossberg, spécialiste de longue date de la modélisation du cortex, appelle cela le « feedback replié ». Je préfère le mot « imagination ».

---

Nous en arrivons au dernier sujet avant de clore cette section. J'ai signalé plusieurs fois que bien souvent, ce que nous voyons, entendons ou éprouvons dépend grandement de nos propres actions. Ce

que nous voyons dépend des saccades oculaires et de l'orientation de la tête. Ce que nous éprouvons dépend des mouvements de nos membres, y compris les doigts. Ce que nous entendons est parfois dépendant de ce que nous disons ou faisons.

C'est pourquoi, pour prédire ce que nous ressentirons ensuite, nous devons connaître les actions que nous entreprenons. Le comportement moteur et la perception sensorielle sont hautement interdépendants. Comment pouvons-nous faire des prédictions si ce que nous ressentirons ensuite est largement le résultat de nos propres actions ? Il existe fort heureusement une solution aussi surprenante qu'élégante à ce problème, bien que beaucoup de détails nous échappent encore.

La première découverte étonnante est que la perception et le comportement, c'est presque du pareil au même. Comme je l'ai déjà mentionné, la plupart sinon toutes les régions du cortex, y compris les aires visuelles, participent à la création du mouvement. Les cellules de la couche 5 qui effectuent des projections vers le thalamus puis vers la couche 1 semblent avoir une fonction motrice car elles effectuent des projections simultanément vers les aires motrices du cerveau archaïque. De ce fait, la connaissance de « ce qui vient de se produire » — tant sur le plan moteur que sensoriel — est disponible dans la couche 1.

La seconde surprise, qui est une conséquence de la première, est que le comportement moteur doit aussi être représenté par une hiérarchie de représentations invariantes. Vous produisez les mouvements nécessaires pour entreprendre une action particulière en pensant à l'exécuter sous une forme à détails invariants. Au cours de sa descente le long de la hiérarchie, l'ordre moteur est traduit en séquences complexes et détaillées requises pour effectuer l'activité prévue. Ceci se produit à la fois dans le cortex « moteur » et dans le cortex « sensoriel », ce qui estompe la distinction entre les deux. Si une région IT du cortex visuel perçoit « nez », la simple action de se représenter « œil » génère la saccade nécessaire pour que cette prédiction devienne réalité. La saccade

particulière nécessaire pour passer de la vision du nez à celle de l'œil varie selon le positionnement du visage. S'il est proche, la saccade oculaire est ample ; s'il est distant, la saccade est faible. Un visage penché impose une saccade sous un autre angle que si le visage est d'aplomb. Les détails de la saccade nécessaire sont déterminés pendant que la prédiction de la vision de « œil » se déplace en V1. La saccade devient de plus en plus spécifique en descendant, d'où un positionnement de la fovéa sur ou très près du point observé.

Examinons un autre exemple. Lorsque je me déplace physiquement du salon à la cuisine, tout ce que mon cerveau a à faire est de passer mentalement d'une représentation invariante du salon à la représentation invariante de la cuisine. Cette transition provoque un complexe déploiement de séquences. Le processus de génération de la séquence de prédictions de ce que je vais voir, ressentir et entendre en me rendant du salon à la cuisine produit la séquence de commandes motrices qui me fait aller d'une pièce à une autre et fait bouger mes yeux au cours de ce déplacement. Prédiction et comportement moteur fonctionnent de pair en tant que flux de patterns montant et descendant le long de la hiérarchie corticale. Aussi étrange que cela puisse paraître, lorsque votre propre comportement est impliqué, vos prédictions non seulement précèdent la sensation, mais la déterminent. Penser à aller au prochain pattern de la séquence provoque une prédiction en cascade de ce que vous devriez éprouver ensuite. Pendant qu'elle se déploie, la prédiction en cascade génère les commandes motrices nécessaires pour réaliser la prédiction. Penser, prédire et faire sont autant d'éléments d'un même déploiement de séquences descendant le long de la hiérarchie corticale.

« Faire » en pensant, ce qui est le déploiement parallèle de la perception et du comportement moteur, est l'essence même de ce qui s'appelle un comportement orienté vers le but. C'est le Graal de la robotique. Il est inscrit dans les tissus du cortex.

Il est bien sûr possible de désactiver notre comportement moteur. Je peux penser à la vision d'une chose sans véritablement la voir, et je peux penser à aller dans la cuisine sans véritablement m'y rendre. Mais, penser à faire quelque chose est littéralement le point de départ de comment nous agirons.

## FLUX ASCENDANT ET FLUX DESCENDANT

Revenons un peu en arrière et réfléchissons davantage à la manière dont l'information monte et descend le long de la hiérarchie corticale. Tandis que vous allez et venez, un flux d'inputs changeants parvient aux régions inférieures du cortex. Chaque région s'efforce de savoir si son flux d'inputs appartient à une séquence de patterns connue. Les colonnes essayent d'anticiper leur activité. Si elles y réussissent, elles transmettent un pattern stable, le nom de la séquence, à la région au-dessus. Là encore, c'est comme si la région disait : « J'écoute une chanson. Voici son nom. Je peux m'occuper des détails. »

Mais que se passerait-il si un pattern inattendu arrivait, une note imprévue ? Ou que se passerait-il si nous apercevions tout à coup quelque chose qui n'appartiendrait pas à un visage ? Le pattern inattendu est automatiquement transmis à la prochaine région supérieure du cortex. Ceci se produit naturellement car les cellules de la couche 3b qui ne font pas partie de la séquence attendue s'excitent. La région supérieure peut être capable de comprendre que ce nouveau pattern est la prochaine partie de sa propre séquence. Elle pourrait dire : « Je vois arriver une nouvelle note. C'est peut-être la première de la prochaine chanson de l'album. C'est bien possible ; je prédis donc que nous sommes passé à la prochaine chanson. Avis à la région d'en dessous : voici le nom de la prochaine chanson qui, à mon avis, sera entendue. » Mais si cette reconnaissance ne se produit pas, un pattern inattendu continuera à se propager vers le haut de la hiérarchie corticale jusqu'à ce que quelque région supérieure parvienne à l'interpréter comme faisant partie de la séquence

normale des événements. Plus le pattern inattendu doit monter haut, plus le nombre de régions du cortex impliquées dans la résolution de l'input inattendu s'accroît. Enfin, quand une région quelque part en haut de la hiérarchie estime qu'elle peut comprendre l'événement inattendu, elle génère une nouvelle prédiction. Cette dernière se propage vers le bas de la hiérarchie aussi loin qu'elle le peut. Si la nouvelle prédiction est erronée, une erreur est détectée, et de nouveau, la prédiction grimpe dans la hiérarchie jusqu'à ce qu'une région parvienne à l'interpréter comme faisant partie de son actuelle séquence active. Ce qui nous conduit à constater que dans la hiérarchie le flux des patterns observés est ascendant et celui des prédictions descendant. Idéalement, dans un monde qui serait connu et prévisible, la plupart des flux de patterns ascendants et descendants seraient rapides et se produiraient dans les régions inférieures du cortex. Le cerveau essaye de découvrir rapidement la partie de son modèle du monde qui serait en accord avec tout input inattendu. Ce n'est qu'alors qu'il comprendrait l'input et saurait à quoi s'attendre ensuite.

Quand je me promène dans une chambre familière de ma maison, peu d'erreurs se propagent vers la partie supérieure de mon cortex. Les séquences bien apprises concernant mon habitation peuvent être gérées dans les parties inférieures de la hiérarchie visuelle, somatosensorielle et motrice. Je connais si bien ma chambre que je pourrais y circuler les yeux fermés. Cette familiarité libère la majorité de mon cortex pour d'autres tâches comme mes réflexions sur le cerveau et l'écriture de livres. Mais si j'étais dans une chambre inconnue, surtout si elle est très différente de tout ce que j'aurais pu voir auparavant, il faudrait non seulement que je puisse voir où je vais, mais des patterns inattendus seraient sans cesse acheminés très haut dans la hiérarchie corticale. Moins l'expérience sensorielle en cours s'accorde avec les séquences apprises, et plus les erreurs qui surgissent sont nombreuses. Dans cette situation nouvelle, je ne pourrais plus penser au cerveau en

tant que sujet d'étude car mon cortex serait confronté aux problèmes d'orientation dans la chambre. C'est une expérience que connaissent bien les voyageurs qui débarquent dans un pays inconnu. Bien que les routes puissent ressembler à celles qui leur sont familières, les voitures roulent peut-être de l'autre côté de la chaussée, la monnaie est différente et la langue aussi, et rien que trouver une salle de bains peut accaparer toute la puissance corticale. Inutile d'essayer de répéter un discours en marchant dans un pays qui vous est complètement étranger.

La sensation de compréhension subite, le fameux « bon sang, mais c'est bien sûr ! », peut être comprise dans ce modèle. Supposons que vous regardiez une image ambiguë : faite de taches d'encre et de traits épars, elle n'évoque rien. La confusion naît de l'impossibilité, pour le cortex, de trouver une mémorisation concordant avec l'input. Nos yeux parcourent l'image. De nouveaux inputs remontent toute la hiérarchie corticale. Le cortex supérieur émet un grand nombre d'hypothèses et les teste, mais lorsqu'elles descendent dans la hiérarchie, toutes et chacune entrent en conflit avec l'input, et le cortex est obligé d'essayer de nouveau. Au cours de cet épisode marqué par la confusion, votre cerveau est complètement accaparé par la recherche de la compréhension de l'image. Vous procédez enfin à une prédiction de haut niveau qui se révèle exacte. A ce moment, cette prédiction faite tout en haut de la hiérarchie corticale réussit à se propager jusqu'au pied de la hiérarchie. En moins d'une seconde, chaque région a fourni une séquence qui correspond à la donnée. Plus aucune erreur ne remonte. Vous comprenez ce que représente l'image : une partie des taches forme un chien dalmatien, comme à la Figure 12.

## LE BIOFEEDBACK PEUT-IL VRAIMENT LE FAIRE ?

Nous savons depuis des décennies que dans la hiérarchie corticale les connexions sont réciproques. Si une région A envoie des projections vers une région B, dans ce cas, B envoie des projections vers A. Les fibres axoniques ramenant en arrière sont souvent



**Figure 12 : Vous voyez le dalmatien ?**

plus nombreuses que celles allant vers l'avant. Mais même si cette description est largement acceptée, le paradigme qui prévaut est que le feedback, ou plus exactement le biofeedback, joue un rôle mineur, ou « modulateur », dans le cerveau. L'idée selon laquelle un signal de biofeedback peut provoquer l'excitation instantanée et précise de divers ensembles de cellules de la couche 2 ne prévaut pas parmi les neurobiologistes.

Pourquoi en serait-il ainsi ? En partie, comme je l'ai mentionné, parce qu'il n'y a pas de véritable nécessité d'être concerné

par le feedback si vous n'acceptez pas le rôle central de la prédiction. Si vous pensez que l'information s'écoule directement à travers le système moteur, à quoi bon un biofeedback ? Une autre bonne raison d'ignorer le biofeedback est que le signal de feedback est diffusé sur de larges zones de la couche 1. Intuitivement, nous nous attendrions à ce qu'un signal dispersé sur une vaste aire n'ait qu'un effet mineur sur beaucoup de neurones, et qu'effectivement le cerveau soit doté de tels signaux modulateurs qui n'agissent pas sur des neurones en particulier, mais modifient des attributs globaux comme la vigilance.

La dernière raison d'ignorer le biofeedback tient à l'idée que se font bon nombre de scientifiques du fonctionnement de chacun des neurones. Un neurone possède des milliers, voire des dizaines de milliers de synapses. Certaines se trouvent loin du corps cellulaire, d'autres tout près. Les synapses à proximité de corps cellulaire exercent une forte influence sur l'excitation d'une cellule. Une douzaine de synapses actives proches du corps cellulaire peuvent l'inciter à déclencher un potentiel, c'est-à-dire une décharge électrique. C'est bien connu. Toutefois, la grande majorité des synapses ne se trouve pas à proximité du corps des cellules. Elles sont dispersées loin et largement sur la structure en arborescence des dendrites des cellules. Parce que ces synapses sont éloignées des corps cellulaires, les scientifiques ont eu tendance à croire qu'un potentiel parvenant à l'une d'elles aurait un effet faible ou trop peu perceptible pour permettre à un neurone de produire un potentiel. Le temps qu'il atteigne le corps cellulaire, l'effet sur une synapse distante serait dissipé.

En règle générale, l'information qui monte le long de la hiérarchie corticale est transférée au travers des synapses proches des corps cellulaires. L'information ascendante est ainsi plus assurée de passer de région en région. De même, en règle générale, le feedback descendant le long de la hiérarchie corticale le fait au travers de synapses loin du corps cellulaire. Les cellules dans les couches 2, 3 et 5 envoient des dendrites dans la couche 1 et y

forment un grand nombre de synapses. La couche 1 est une masse de synapses, mais elles sont toutes loin des corps cellulaires des couches 2, 3 et 5. De plus, toute cellule particulière, disons dans la couche 2, ne formera, le cas échéant, que peu de synapses avec quelque fibre de feedback particulière que ce soit. C'est pourquoi des scientifiques pourraient rejeter l'idée qu'un bref pattern, dans la couche 1, peut provoquer avec exactitude l'excitation d'un ensemble de cellules dans les couches 2, 3 et 5. Mais c'est précisément ce qu'exige la théorie que j'ai élaborée.

La résolution de ce dilemme tient au fait que les neurones ne se comportent différemment que dans le modèle classique. En fait, ces dernières années, un nombre grandissant de scientifiques ont émis l'hypothèse que les synapses présentes sur des dendrites minces et distantes pourraient jouer un rôle actif et très spécifique dans l'excitation des cellules. Dans ces modèles, ces synapses distantes se comportent différemment des synapses des dendrites épaisses, situées à proximité du corps cellulaire. Par exemple, si deux synapses étaient très proches l'une de l'autre sur une dendrite mince, elles agiraient en tant que « détecteur de coïncidence » : si les deux synapses recevaient un potentiel d'input dans un très bref laps de temps, elles pourraient exercer un effet fort sur la cellule, bien qu'elles soient éloignées du corps cellulaire. Ce dernier générerait alors un potentiel. Le comportement des dendrites d'un neurone est encore un mystère ; je n'en dirai donc pas davantage. Ce qui est important est que le modèle de mémoire-prédiction du cortex exige que les synapses loin du corps cellulaire soient capables de détecter des patterns spécifiques.

Rétrospectivement, il semble absurde d'affirmer que la plupart des milliers de synapses d'un neurone ne jouent qu'un rôle modulateur. Les énormes quantités de biofeedbacks et le nombre extrêmement élevé de synapses ont une raison d'être. A partir de là, nous pouvons dire qu'un neurone typique est capable d'apprendre des centaines de coïncidences précises à partir des fibres de feedback, lorsqu'elles créent des synapses sur des

dendrites minces. Cela signifie que chaque colonne de notre néocortex est hautement flexible, en termes de patterns de biofeedback susceptibles de les activer. Cela signifie que chaque fonctionnalité particulière peut être précisément associée à des milliers d'objets et de séquences différents. Mon modèle a besoin du biofeedback pour être rapide et précis. Les cellules doivent pouvoir s'exciter lorsqu'elles détectent n'importe quel nombre de coïncidences précises sur leurs dendrites distantes. Ces nouveaux modèles neuronaux l'autorisent.

## COMMENT LE CORTEX APPREND

Toutes les cellules de toutes les couches du cortex ont des synapses, et la plupart des synapses peuvent être modifiées par l'expérience. On peut affirmer que l'apprentissage et la mémorisation s'effectuent dans toutes les couches, dans toutes les colonnes et dans toutes les régions du cortex.

Précédemment dans ce livre, j'ai mentionné l'apprentissage hebbien, du nom du neuropsychologue canadien Donald Olding Hebb. Le principe est très simple : quand deux neurones sont excités simultanément, les synapses qui se trouvent entre eux sont renforcées. Nous savons à présent que Hebb avait fondamentalement raison. Bien sûr, rien dans la nature n'est jamais si simple, et dans le cerveau, c'est encore plus compliqué. Notre système nerveux exécute diverses variantes des règles d'apprentissage hebbien. Par exemple, certaines synapses modifient leur force en réponse à de faibles variations de la temporisation des signaux neuronaux ; certains changements synaptiques sont passagers, d'autres persistants. Hebb n'avait toutefois échafaudé qu'un cadre pour l'étude de l'apprentissage, et non une théorie finale, mais ce cadre s'avéra extrêmement utile.

Les principes d'apprentissage hebbien peuvent expliquer la plupart des comportements corticaux mentionnés dans ce chapitre. C'est dans les années 1970 qu'il fut démontré que les mémoires autoassociatives reposant sur le classique algorithme d'apprentissage

hebbien étaient capables d'apprendre des patterns spatiaux et des séquences de patterns. Le principal problème était que les mémoires ne pouvaient gérer correctement les variations. Selon la théorie proposée dans ce livre, le cortex a contourné cette limitation, d'une part en empilant les mémoires auto-associatives dans une hiérarchie, et d'autre part en recourant à une architecture colonnaire sophistiquée. Ce chapitre est presque entièrement consacré à la hiérarchie et à son fonctionnement, car c'est grâce à elle que le cortex est si puissant. C'est pourquoi, au lieu de nous plonger dans les fastidieux détails expliquant comment chaque cellule est capable d'apprendre ceci ou cela, je préfère révéler quelques principes généraux d'apprentissage dans le cadre de la hiérarchie.

Au moment de la naissance, le cortex ne sait quasiment rien. Il ne sait rien de votre langue, de votre culture, de votre habitation, de votre ville, de la musique, des gens qui vous entourent… Rien. Toutes ces informations, la structure du monde, doivent être apprises. Les deux composants de base de l'apprentissage sont les classifications de patterns et la construction de séquences. Ces deux composants complémentaires de la mémoire interagissent. Tandis qu'une région apprend des séquences, les inputs qu'elle envoie aux cellules de la couche 4, dans les régions corticales élevées, changent. De ce fait, les cellules de la couche 4 apprennent à former de nouvelles classifications, qui modifient les patterns rétroprojetés vers la couche 1, dans la région inférieure, ce qui affecte les séquences.

Le fondement pour la formation des séquences est le regroupement des patterns faisant partie d'un même objet. Une façon de le faire consiste à grouper les patterns dont la création se succède. Quand un enfant tient un jouet dans sa main et le fait lentement bouger, son cerveau peut fort bien supposer que l'image qui se forme sur la rétine est celle du même objet, moment après moment, et que par conséquent l'ensemble changeant de patterns peut être groupé. A un autre moment, une instruction venue de l'extérieur sera nécessaire pour aider à savoir quels patterns appartiennent à un

même groupe. Pour apprendre que les pommes et les bananes sont des fruits, mais que les carottes et le céleri n'en sont pas, il faut qu'un enseignant guide l'enfant pour grouper ces éléments en fruits et en légumes. De toutes manières, le cerveau élabore lentement les séquences de patterns qui appartiennent à un même ensemble. Mais tandis que les régions du cortex construisent des séquences, l'input vers la prochaine région change. Celui qui représentait le plus souvent des patterns individuels représente peu à peu des groupes de patterns. L'input vers une région passe des notes à la mélodie, des lettres aux mots, des nez aux visages, et ainsi de suite. Comme les inputs montant vers une région deviennent plus « orientés objets », la région supérieure du cortex peut à présent apprendre des séquences concernant ces objets d'ordre plus élevé. Là où auparavant une région élaborait des séquences de lettres, elle construit maintenant des séquences de mots. Le résultat inattendu de ce processus d'apprentissage est que, au cours de l'apprentissage répétitif, les représentations des objets descendent dans la hiérarchie corticale. Au cours des premières années de la vie, les mémorisations concernant le monde se forment d'abord dans les régions supérieures du cortex, mais au fur et à mesure de l'apprentissage, elles sont reformées en bas et dans les parties inférieures de la hiérarchie corticale. Ce n'est pas le cerveau qui les déplace ; il doit les réapprendre sans cesse (je n'avance pas que toutes les mémorisations commencent en haut du cortex. La véritable formation des mémoires est autrement plus complexe. Je crois que la classification des patterns dans la couche 4 commence en bas et monte. C'est la *mémoire de séquences* qui, à mon avis, se reforme toujours plus bas dans le cortex). Tandis que les représentations simples descendent, les régions supérieures deviennent capables d'apprendre des patterns plus compliqués et plus subtils.

Vous pouvez observer la création et le mouvement descendant d'une mémoire hiérarchique en regardant comment un enfant apprend. Prenons l'acquisition de la lecture. La première chose qu'il apprend est la reconnaissance de chacune des lettres.

C'est une tâche lente et laborieuse qui exige un effort soutenu. Ensuite, il apprend des mots simples. Là encore, c'est d'abord difficile, même pour des mots de trois lettres. L'enfant parvient à les lire l'une après l'autre et les épeler, mais il lui faut une certaine pratique avant que le mot lui-même soit reconnu comme tel. Après avoir appris des mots simples, il se confronte à des mots compliqués, à plusieurs syllabes. Il ânonne d'abord chacune d'elles, les fait se suivre comme il le faisait des lettres pour former des mots. Après des années de pratique, il acquiert une certaine aisance dans la lecture. Il en arrive au point où il ne voit plus véritablement chaque lettre séparée, mais reconnaît des mots entiers et, souvent, une phrase entière d'un seul coup d'œil. Ce n'est pas qu'il soit devenu plus rapide ; il reconnaît en fait les mots et les phrases sous forme d'entités. Quand nous lisons un mot, voyons-nous les lettres ? Oui et non. De toute évidence, la rétine perçoit les lettres, et par conséquent les régions en V1 les perçoivent aussi. La reconnaissance des lettres se produit relativement bas dans la hiérarchie corticale, disons en V2 ou V4. Au moment où le signal arrive en IT, les différentes lettres ne sont plus représentées. Ce qui avait d'abord exigé un effort de tout le cortex visuel — la reconnaissance de chacune des lettres — se produit à présent plus près de l'input sensoriel. Lorsque la mémoire d'objets simples comme des caractères descend dans la hiérarchie, les régions plus élevées acquièrent la capacité à apprendre des objets complexes comme des mots et des phrases.

L'apprentissage de la musique est un autre exemple. Vous devez d'abord vous concentrer sur chaque note. La pratique aidant, vous parvenez à reconnaître des séquences de notes communes, puis des phrases entières. Après beaucoup de pratique, c'est comme si vous ne voyiez plus la plupart des notes. La partition musicale n'est là que pour vous rappeler la structure globale du morceau ; les séquences détaillées ont été mémorisées plus bas. Ce type d'apprentissage se produit à la fois dans les aires motrices et dans les aires sensorielles.

Un jeune cerveau est plus lent à reconnaître des inputs et plus lent à déclencher des commandes motrices car les mémoires mises en œuvre pour ces tâches sont plus haut dans la hiérarchie corticale. L'information doit s'écouler tout en haut et tout en bas, parfois à plusieurs reprises afin de résoudre d'éventuels conflits. Il faut du temps aux signaux neuronaux pour monter et descendre tout au long de la hiérarchie corticale. Un jeune cerveau n'a pas encore formé les séquences complexes dans les régions supérieures, et ne peut donc pas reconnaître et restituer des patterns complexes. Un jeune cerveau ne peut pas comprendre la structure d'ordre élevée du monde. Comparé à celui d'un adulte, le langage d'un enfant est simple, ses chansons sont simples et ses interactions sociales le sont aussi.

Quand vous étudiez et réétudiez un ensemble particulier d'objets, votre cortex reforme les représentations mémorielles de ces objets en bas de la hiérarchie. Ceci libère les régions supérieures pour apprendre des relations plus complexes, plus subtiles. Selon la théorie, c'est cette capacité qui fait la différence entre un érudit et quelqu'un de fruste.

Dans mon travail qui consiste à concevoir des ordinateurs, des collègues sont surpris de voir avec quelle rapidité je parviens à examiner un produit et identifier les problèmes inhérents à sa conception. Avec vingt-cinq ans d'expérience, j'ai en tête un modèle supérieur à la moyenne pour tout ce qui concerne l'évaluation d'un matériel informatique mobile. De même, un parent expérimenté voit tout de suite ce qui ne va pas chez son enfant, alors que des parents moins aguerris ne sauront pas par quel bout prendre la situation. Un directeur d'entreprise expérimenté voit immédiatement les avantages et les inconvénients d'une organisation structurelle, alors qu'un cadre frais émoulu d'une école n'y comprend rien. Ils ont le même input, mais le modèle du novice n'est pas aussi sophistiqué. Dans tous ces cas et dans une infinité d'autres, nous commençons par apprendre les bases, la structure la plus simple. Au fil du temps, notre savoir descend dans la

hiérarchie corticale, laissant aux régions supérieures la possibilité d'apprendre des structures d'ordre plus élevé. C'est cette structure d'ordre élevé qui fait de nous des gens expérimentés. Le cerveau des gens doués et des génies est capable de saisir des structures de structures et des patterns de patterns qui échappent au commun des mortels. C'est par la pratique que nous devenons des experts dans un ou plusieurs domaines, mais le talent et le génie ont certainement une composante génétique.

## L'HIPPOCAMPE : AU-DESSUS DE TOUT

Trois grandes structures cérébrales se trouvent sous la feuille néocorticale et communiquent avec elle. Il s'agit des ganglions basaux, du cervelet et de l'hippocampe. Tous les trois existaient bien avant la formation du néocortex. A très gros traits, nous pourrions dire que les ganglions basaux formaient le système moteur primitif, que le cervelet a appris la coordination précise des relations entre les événements, et que l'hippocampe a stocké les mémoires des événements et des lieux spécifiques. Jusqu'à un certain point, le néocortex a subsumé ses fonctions originales. Par exemple, un être humain né avec un cervelet atrophié souffrira d'un déficit au niveau de la coordination de ses mouvements. Il lui faudra faire davantage d'efforts conscients pour se mouvoir, mais à part ce handicap, sa vie sera quasiment normale.

Nous savons que le néocortex régit toutes les séquences motrices complexes et peut contrôler directement nos membres. Cela ne signifie pas que les ganglions basaux sont sans importance, mais seulement que le néocortex a pris le contrôle d'une bonne partie de notre contrôle moteur. C'est pourquoi j'ai décrit toutes les fonctions globales du néocortex indépendamment des ganglions basaux et du cervelet. Certains scientifiques ne seront pas d'accord avec ce procédé, mais c'est celui que j'ai utilisé dans ce livre et dans mon travail.

Il en va tout autrement de l'hippocampe. Cette circonvolution inversée est l'une des régions du cerveau qui a été la plus

étudiée car elle est essentielle dans la formation de nouvelles mémoires. Si l'hippocampe était détruit, vous perdriez toute capacité à mémoriser. Sans l'hippocampe, vous pouvez certes toujours parler, marcher, voir, entendre et, pour un bref moment, paraître normal. Mais en réalité, vous seriez profondément handicapé car il vous serait impossible de vous rappeler de tout ce qui est nouveau. Vous vous souviendrez des gens que vous avez connus avant de perdre l'hippocampe, mais pas des nouvelles rencontres. Et même si vous allez régulièrement chez votre médecin, pour vous, ce sera toujours la première fois. Vous n'auriez plus aucune mémoire de ce qui s'est passé après la perte de l'hippocampe.

Pendant des années, j'avais négligé l'hippocampe car je ne voyais pas ce que son étude pouvait m'apporter. Il est bien sûr primordial pour l'apprentissage, bien qu'il ne soit pas le dépositaire suprême de tout ce que nous connaissons. C'est le néocortex qui joue ce rôle. La vision classique de l'hippocampe est que de nouvelles mémorisations s'y forment, et plus tard, après quelques jours, semaines ou mois, elles sont transférées dans le néocortex. Ceci n'avait pour moi aucun sens. Nous savons que la vue, l'ouïe et le toucher — des flux sensoriels — sont acheminés directement jusque dans les aires sensorielles du cortex sans transiter par l'hippocampe. Il me semblait que ces informations sensorielles devaient spontanément former de nouvelles mémorisations dans le cortex. Pourquoi aurions-nous besoin d'un hippocampe pour apprendre ? Comment une structure distincte comme l'hippocampe peut-elle interférer avec le cortex et y empêcher l'apprentissage, et n'y transférer les informations qu'ultérieurement ?

J'avais décidé de laisser l'hippocampe de côté en attendant qu'un jour son rôle m'apparaisse plus clairement. Ce jour arriva au moment où je m'attelais à l'écriture de ce livre. L'un de mes collègues du Redwood Neuroscience Institute, Bruno Olshausen, me fit remarquer que les connexions entre l'hippocampe et le néocortex laissent à penser que l'hippocampe est une région

hiérarchiquement située au-dessus du néocortex, et non une structure distincte. De ce point de vue, l'hippocampe est au sommet de la pyramide néocorticale. C'est le bloc tout en haut, à la Figure 5. Dans l'évolution des espèces, le néocortex est apparu en sandwich entre l'hippocampe et le reste du cerveau. Apparemment, cette vision de l'hippocampe tout en haut de la hiérarchie corticale était connue depuis quelque temps, mais je n'en avais pas conscience. J'avais discuté avec plusieurs spécialistes de l'hippocampe et je leur avais demandé de m'expliquer comment cette structure en forme de cheval de mer s'y prend pour transférer les mémorisations vers le cortex. Aucun ne le put. Et aucun ne mentionna que l'hippocampe est au sommet de la hiérarchie corticale, probablement parce que cette partie du cerveau occupe non seulement une position élevée, mais aussi parce qu'elle est reliée à beaucoup de parties anciennes du cerveau.

Cette nouvelle perspective me fit pourtant instantanément entrevoir une solution à mon problème.

Pensez à l'information acheminée des yeux, des oreilles et de la peau jusqu'au néocortex. Chaque région du néocortex essaye de comprendre sa signification. Chaque région essaye de comprendre l'input en termes de séquences qu'elle connaîtrait. Si elle comprend l'input, elle dit : « Je comprends ça, c'est juste une partie de l'objet que je vois actuellement. Je ne passerai pas sur les détails. » Si une région ne comprend pas l'input, elle le transmet plus haut dans la hiérarchie jusqu'à ce qu'une région supérieure y comprenne quelque chose. Toutefois, un pattern totalement nouveau grimpera sans cesse dans la hiérarchie. Chacune des régions qu'il rencontrera dans son ascension dira : « Je ne sais pas ce que c'est, je ne l'ai pas anticipé, pourquoi ne pas le confier à une instance supérieure ? » Il résulte de ce processus que, le sommet de la pyramide corticale atteint, vous vous retrouvez avec une information impossible à comprendre d'après l'expérience passée. Il ne vous reste que la partie de l'input véritablement nouvelle et inattendue.

Au cours d'une journée, nous sommes confrontés à beaucoup de nouveautés destinées au sommet de la pyramide : un article dans un journal, le nom de quelqu'un qui vous a été présenté ce matin, l'accident de la route que vous avez vu sur le chemin du retour… Ce sont ces événements encore inexpliqués ou non anticipés, ces matériaux nouveaux, qui entrent dans l'hippocampe et y sont stockés. Ces informations n'y demeurent pas définitivement : soit elles seront transférées dans le cortex, au niveau inférieur, soit elles seront finalement perdues.

J'ai remarqué qu'en prenant de l'âge, j'ai plus de mal à mémoriser ce qui est nouveau. Par exemple, mes enfants se souviennent des moindres détails de la plupart des pièces de théâtre auxquelles ils ont assisté l'année précédente. Moi, non. C'est peut-être parce que j'en ai vu tellement dans ma vie qu'il est rare que j'y décèle du nouveau. Les pièces nouvelles concordent avec la mémoire des pièces passées, de sorte que l'information n'est pas transmise à l'hippocampe. Pour mes enfants, chaque pièce est toute nouvelle et transite donc par l'hippocampe. Si ceci est vrai, nous pourrions en déduire que plus vous savez, moins vous vous souvenez.

Contrairement au néocortex, l'hippocampe possède une structure hétérogène dotée de plusieurs régions spécialisées. Il est impeccable pour l'unique tâche consistant à stocker rapidement tous les patterns qu'il voit. L'hippocampe occupe une position parfaite pour mémoriser ce qui est tout nouveau, en haut de la pyramide corticale. Il est aussi en parfaite position pour s'en souvenir, ce qui permet de stocker les nouveautés dans la hiérarchie corticale, et ce qui est un processus relativement lent. Vous pouvez vous remémorer instantanément un tout nouvel événement qui est dans l'hippocampe, mais vous ne vous souviendrez en permanence d'une information présente dans le cortex que si vous en avez fait et refait l'expérience, que ce soit dans la réalité ou en pensée.

## UNE VOIE ALTERNATIVE POUR MONTER DANS LA HIÉRARCHIE

Le cortex est doté d'une autre grande voie pour transmettre une information de région en région en montant dans la hiérarchie. Elle commence aux cellules de la couche 5, qui envoient des projections vers le thalamus (mais vers une autre partie que celle évoquée auparavant), puis du thalamus vers des régions supérieures du cortex. Chaque fois que deux régions du cortex sont directement connectées d'une manière hiérarchique, elles le sont aussi indirectement au travers du thalamus. Cette seconde voie, ou voie alternative, ne transmet l'information que vers le haut de la hiérarchie, pas vers le bas. Il existe, en montant dans la hiérarchie corticale, une voie directe entre deux régions et une autre voie indirecte passant par le thalamus.

La seconde voie possède deux modes opératoires déterminés par les cellules du thalamus. Dans l'un, la voie est le plus souvent close afin que l'information ne l'emprunte pas. Dans l'autre mode, l'information circule librement et fidèlement entre les régions. Deux scientifiques, Murray Sherman, de l'Université de l'Etat de New York à Stony Brook, et Ray Guillery, de l'Ecole de médecine de l'Université du Winsconsin ont décrit cette voie alternative et postulé qu'elle pourrait être aussi importante, voire plus, que la voie directe qui était jusqu'à présent celle exposée dans ce chapitre. Je nourris quelques spéculations sur ce que fait cette seconde voie.

Lisez le mot *imagination*. La plupart des gens reconnaissent ce mot au premier coup d'œil, dès la première fixation. Personne ne s'attache à la lettre *i* qui se trouve au milieu. Regardez maintenant le point sur ce *i*. Votre regard est fixé exactement au même endroit, mais dans le premier cas vous voyez le mot, dans le deuxième le caractère typographique, et dans le dernier cas le point. Fixez le *i* puis essayez de faire passer votre perception du mot à la lettre et de la lettre au point. Si vous éprouvez quelques difficultés, essayez de dire « point », « i » et « imagination » tandis

que vos yeux restent rivés sur le point. Dans chaque cas, la même information parvient en V1, et pourtant, quand elle arrive dans une région plus élevée comme IT, vous percevez d'autres choses, différents niveaux de détail. La région IT sait reconnaître chacun des trois objets. Elle identifie le point isolément, la lettre *i* et le mot entier du premier coup d'œil. Mais quand vous percevez le mot entier, V4, V2 et V1 se chargent des détails, de sorte que tout ce que IT reçoit à reconnaître, c'est le mot. Quand vous lisez, vous ne percevez habituellement pas chacune des lettres, mais rien que des mots et des phrases. Vous ne pouvez percevoir les lettres que si vous choisissez de le faire. Nous effectuons tout le temps ce genre de déplacement de l'attention, mais nous n'en sommes généralement pas conscients. Vous pouvez entendre une musique en bruit de fond et vous rendre à peine compte de la mélodie. Mais si vous décidez d'écouter, il vous sera possible d'isoler le chanteur ou la guitare basse. C'est toujours le même son qui entre dans la tête, mais la perception est à présent focalisée. Chaque fois que vous vous grattez la tête, ce geste produit un fort bruit interne dont vous n'êtes généralement pas conscient. Mais si vous vous concentrez sur ce son, vous le percevrez clairement. C'est un autre exemple d'input sensoriel normalement géré en bas de la hiérarchie corticale, mais qui peut être amené à des niveaux plus élevés si vous y prêtez attention.

Je présume que cette voie alternative traversant le thalamus est le mécanisme par lequel nous prêtons attention aux détails que nous ne remarquons d'ordinairement pas. Elle évite le groupement des séquences dans la couche 2 en acheminant les données brutes vers la prochaine région supérieure du cortex. Les biologistes ont montré que la voie alternative peut être transformée de deux façons. L'une l'est par un signal provenant de la région supérieure du cortex lui-même ; c'est cette technique que vous avez mise en œuvre quand je vous ai demandé de prêter attention à des détails que vous n'auriez normalement pas remarqués, comme le point sur le *i* ou le son lorsque vous vous grattez la tête. La seconde façon

d'activer la voie est un signal fort, inattendu, provenant d'une région inférieure. Si l'input dans la voie alternative est suffisamment puissant, il envoie un signal de réveil vers la région supérieure, laquelle peut à son tour se lier à la voie. Par exemple, si je vous montre un visage et si je vous demande ce que c'est, vous répondrez « un visage ». Si je vous montre le même visage mais avec une marque bizarre sur le nez, vous reconnaîtrez d'abord le visage, mais aussitôt, les niveaux inférieurs de la vision remarqueront que quelque chose ne va pas. Cette erreur force l'ouverture d'une voie attentionnelle. Les détails emprunteront à présent la voie alternative, évitant le groupement qui se produit normalement, et votre attention sera attirée vers la marque bizarre. C'est cette dernière que vous voyez à présent et pas uniquement le visage. Si elle est suffisamment insolite, la marque accaparera toute votre attention. De cette manière, les événements inhabituels sont rapidement signalés à notre attention. Voilà pourquoi nous ne pouvons éviter de nous focaliser sur des difformités et autres patterns inhabituels, car notre cerveau le fait spontanément. Toutefois, les erreurs ne sont souvent pas suffisamment fortes pour ouvrir la voie alternative. C'est pourquoi nous ne remarquons parfois pas une faute d'orthographe au cours de la lecture.

## POUR CONCLURE

Pour découvrir et établir une nouvelle structure scientifique, il est indispensable de rechercher les concepts les plus simples, capables d'unifier et d'expliquer une grande quantité de faits disparates. Une simplification excessive est la conséquence inévitable de cette démarche. Des détails importants risquent d'être ignorés et des faits mal interprétés. Si la structure est admise, des affinements seront inévitablement proposés ainsi que des corrections là où une hypothèse initiale s'est fourvoyée, n'a pas été approfondie ou est erronée.

Dans ce chapitre, j'ai présenté un grand nombre de spéculations sur le fonctionnement du néocortex. Je pense qu'une partie

de ces idées se révélera erronée et que l'ensemble de ces idées sera révisé. Beaucoup de détails n'ont pas même été mentionnés. Le cerveau est un organe très complexe ; les neurobiologistes qui liront ce livre verront que je n'en ai exposé ici qu'une représentation très rudimentaire. Je pense cependant que, dans son ensemble, la structure tient la route. Tout ce que je peux espérer, c'est que l'idée centrale sera préservée tandis que les détails évolueront au gré de nouvelles données et de nouvelles découvertes.

Enfin, peut-être êtes-vous réticent à l'idée que d'un vaste mais simple système de mémorisation puisse résulter tout ce que l'être humain est capable d'entreprendre. Ne serions-nous, vous et moi, rien d'autre qu'un système de mémorisation hiérarchique ? Nos vies, nos croyances et nos ambitions seraient-elles engrangées dans des milliers de milliards de minuscules synapses ? J'ai commencé à écrire des programmes informatiques en 1984. J'en avais développé auparavant, très courts, mais cette année, c'était la première fois que je programmais un ordinateur à l'aide d'une interface graphique, et la première fois que je travaillais sur de vastes applications sophistiquées. J'avais écrit des programmes pour un système d'exploitation créé par Grid Systems. Avec ses fenêtres, ses polices multiples et ses menus, il était en avance sur son temps.

Un jour, je me suis rendu compte qu'un projet était quasiment impossible à mener à son terme. En tant que programmeur, j'écrivais une ligne de code à la fois. Je groupais ces lignes en blocs nommés sous-routines. Les sous-routines étaient groupées en modules et ces modules réunis pour former une application. Le projet de tableur sur lequel je travaillais comportait un si grand nombre de sous-routines et de modules que plus personne ne pouvait en comprendre la globalité. Le programme était devenu trop compliqué. Pourtant, une seule ligne de code ne fait que bien peu de choses. Placer un seul pixel à l'écran en exige plusieurs. Afficher le tableur sur la totalité de l'écran oblige l'ordinateur à exécuter des millions d'instructions réparties sur des centaines de sous-routines. Les

sous-routines en appellent d'autres d'une manière répétitive et récursive. Tout ceci était si compliqué qu'il me sembla impossible de prévoir tout ce qui se passerait lorsque le programme serait exécuté. Il me parut hautement improbable que, le moment venu, il serait capable de tracer rapidement les différents éléments du tableur. L'apparence extérieure était celle d'un tableur avec ses tableaux de chiffres, ses étiquettes, ses textes et ses graphiques. Et pourtant, il se comporta comme un tableur. Je savais qu'à l'intérieur de l'ordinateur le processeur exécutait une seule instruction après une autre. Il me parut incroyable que l'ordinateur parvienne à trouver son chemin dans le dédale de modules et de sous-routines et puisse exécuter toutes ces instructions avec célérité. Si je n'en avais pas su davantage, j'aurais été certain que cette usine à gaz ne pourrait jamais fonctionner. Je me rendis compte que si quelqu'un m'avait exposé ce jour-là le concept d'ordinateur doté d'une interface graphique et d'une application de tableur, j'aurais rejeté son projet en affirmant qu'il était irréalisable. Je lui aurais affirmé que l'ordinateur mettrait une éternité pour effectuer le moindre calcul. Cela aurait été humiliant car il s'avère que le tableur fonctionne. C'est alors que je réalisai que l'idée que je me faisais de la vitesse du microprocesseur et de la puissance d'une conception hiérarchique était inadéquate.

On peut en tirer une leçon au sujet du néocortex. Il n'est pas constitué de composants hyper rapides et les règles qui le régissent ne sont pas très compliquées. Il est cependant doté d'une structure hiérarchique contenant des milliards de neurones et des trillions de synapses. S'il nous est difficile d'imaginer comment un système de mémorisation logiquement simple mais numériquement énorme est à l'origine de la connaissance, des langues, des cultures, des arts, de ce livre ainsi que des sciences et de la technologie, je pense que c'est parce que l'idée que nous nous faisons de la capacité du cortex et de la puissance de sa structure hiérarchique est inadéquate. Le néocortex fonctionne. Il n'y a là rien de magique. Nous parvenons à comprendre ses mécanismes.

Et comme pour l'ordinateur, nous parviendrons finalement à créer des machines intelligentes fonctionnant sur les mêmes principes.

# 7

# CONSCIENCE ET CRÉATIVITÉ

**Quand** j'expose ma théorie du cerveau, le public saisit en général rapidement l'importance de la notion de prédiction dans la multitude des activités humaines. Les participants posent de nombreuses questions : D'où provient la créativité ? Que sont la conscience et l'imagination ? Comment faisons-nous la distinction entre la réalité et les fausses croyances ? Bien que ces sujets n'aient pas été au premier plan de mes motivations lorsque j'ai commencé à étudier le cerveau, ils n'en intéressent pas moins beaucoup de monde. Je ne prétends pas être un spécialiste dans ces domaines, mais le cadre de mémoire-prédiction de l'intelligence peut fournir quelques réponses et idées dignes d'intérêt. Dans ce chapitre, je réponds aux questions les plus fréquemment posées.

## LES ANIMAUX SONT-ILS INTELLIGENTS ?

Un rat est-il intelligent ? Un chat est-il intelligent ? A quel moment, dans l'évolution, l'intelligence est-elle apparue ? J'aime bien ces questions car les réponses sont surprenantes.

Tout ce que j'ai écrit au sujet du néocortex et de son fonctionnement repose sur un postulat très élémentaire, à savoir que le monde est structuré et qu'il est par conséquent prédictible. Il existe une foule de patterns autour de nous : les visages ont des yeux, les yeux ont des pupilles, le feu est chaud, la gravité fait tomber les objets, les portes sont ouvertes ou fermées, et ainsi de suite. Le monde n'est ni aléatoire ni homogène. La mémoire, la prédiction et le comportement n'auraient pas de sens si le monde n'était pas structuré. Tout comportement, que ce soit celui d'un être humain, d'un escargot, d'un organisme unicellulaire ou d'un arbre, est un moyen d'exploitation de la structure du monde au bénéfice de la reproduction.

Imaginons un être unicellulaire vivant dans une mare. Il est doté d'un flagelle qui lui permet de nager. A la surface de la cellule, des molécules détectent la présence de nutriments. La concentration de nutriments n'étant pas la même dans toute la mare, il se produit un changement progressif de la valeur des nutriments perçus, d'un côté de la cellule à un autre. Lorsqu'elle nage dans la mare, la cellule parvient à détecter cette variation superficielle. C'est là une forme simple de structure dans l'univers d'un animal unicellulaire. La cellule exploite la conscience qu'elle a de la chimie locale en nageant vers les lieux où la concentration nutritionnelle est la plus élevée. Nous pourrions dire que cet organisme simple fait une prédiction : il prédit qu'en nageant dans une direction donnée il trouvera plus de nutriments. Une mémoire est-elle impliquée dans cette prédiction ? Oui. Cette mémoire est l'ADN (acide désoxyribonucléique) de l'organisme. L'animalcule unicellulaire n'a pas appris, au cours de sa vie, à exploiter la variation superficielle. En réalité, l'apprentissage s'est effectué au cours de l'évolution de l'espèce, et il est à présent stocké dans son ADN. Si la structure du monde changeait subitement, cet animalcule unicellulaire ne pourrait pas apprendre à s'adapter. Il ne pourrait pas modifier son ADN ni le comportement qui en résulte. Pour cette espèce, l'apprentissage ne peut

s'effectuer qu'au cours d'un processus évolutionnaire, sur de nombreuses générations.

Cet organisme unicellulaire est-il intelligent ? Selon l'idée commune que nous nous faisons de l'intelligence, non. Mais cet être se trouve à la frange lointaine d'un continuum d'espèces qui recourent à la mémoire et à la prédiction pour mieux se reproduire ; de ce point de vue plus académique, la réponse est oui. Il ne s'agit pas d'affirmer que certaines espèces sont intelligentes et que d'autres ne le sont pas. Tous les êtres vivants utilisent la mémoire et la prédiction. Il existe seulement une continuité au niveau des méthodes et de la sophistication au niveau de cette utilisation.

Les plantes aussi ont recours à la mémoire et à la prédiction pour exploiter la structure du monde. Un arbre procède à une prédiction lorsqu'il plonge ses racines dans le sol et étend sa ramure vers le ciel. L'arbre prédit où il trouvera de l'eau et des minéraux, en se fiant à l'expérience accumulée par ses ancêtres. Bien sûr, un arbre ne pense pas. Son comportement est un tropisme. Mais les espèces exploitent la structure du monde de la même manière que l'organisme unicellulaire. Chaque espèce végétale possède un ensemble de comportements qui exploitent des parties légèrement différentes de la structure du monde.

Finalement, les plantes ont développé des systèmes de communication essentiellement fondés sur la lente libération de signaux chimiques. Quand un insecte endommage une partie d'un arbre, le système vasculaire de ce dernier émet des produits chimiques, ce qui déclenche un mécanisme de défense à base de toxines. Grâce à de tels systèmes de communication, l'arbre peut manifester un comportement un peu plus complexe. Les neurones ont probablement évolué en tant que moyen permettant de communiquer une information plus rapidement que le système vasculaire d'un végétal. Vous pourriez considérer un neurone comme une cellule équipée de ses propres appendices vasculaires. A un certain point, au lieu de déplacer lentement des produits

chimiques le long de ces appendices, le neurone a commencé à utiliser des potentiels électrochimiques, qui voyagent beaucoup plus vite. Au début, les transmissions synaptiques rapides et les systèmes nerveux simples n'étaient guère impliqués dans les apprentissages. Il s'agissait simplement de signaler plus vite.

Mais par la suite, quelque chose d'intéressant se produisit dans le cours de l'évolution. Des connexions entre des neurones devinrent modifiables. Un neurone put envoyer ou ne pas envoyer un signal, en fonction de ce qui s'était passé récemment. Le comportement put désormais être modifié au cours de l'existence d'un organisme. Le système nerveux devint façonnable, et il en alla de même du comportement. Comme des mémorisations pouvaient être rapidement formées, l'animal put apprendre la structure du monde au cours de la durée de sa vie. Si le monde changeait brusquement — par l'arrivée d'un nouveau prédateur dans son environnement —, l'animal n'était pas tenu d'en rester à son comportement génétiquement défini et imposé, qui pouvait dès lors ne plus être approprié. Le système nerveux façonnable procura un formidable avantage évolutionnaire et suscita l'apparition de nouvelles espèces, qu'il s'agisse de poissons, d'escargots ou de l'être humain.

Comme nous l'avons vu au Chapitre 3, tous les mammifères possèdent un cerveau archaïque au-dessus duquel siège le néocortex. Ce tissu nerveux est apparu le plus récemment dans l'évolution. Mais avec sa structure hiérarchique, ses représentations invariantes et ses prédictions par analogie, le cortex permet aux mammifères d'exploiter bien plus efficacement la structure du monde que les animaux dépourvus de néocortex. Nos ancêtres « corticalement » bien dotés étaient de ce fait capables d'envisager comment fabriquer un filet et l'utiliser pour capturer du poisson. Les poissons ne sont pas capables d'apprendre que le filet signifie la mort, et moins encore de concevoir et fabriquer des outils pour le couper. Tous les mammifères, de la musaraigne au chat et aux humains, ont un néocortex. Tous sont intelligents à des degrés divers.

## QU'EST-CE QUI DIFFÉRENCIE L'INTELLIGENCE HUMAINE ?

Le cadre de mémoire-prédiction apporte deux réponses à cette question. La première est assez directe : notre néocortex est plus vaste que celui, disons, d'un singe ou d'un chien. En élargissant le néocortex jusqu'aux dimensions d'une nappe de table, notre cerveau s'est rendu capable d'apprendre un modèle plus élaboré du monde et procéder ainsi à des prédictions plus complexes. Nous voyons plus d'analogies, plus de structures parmi les structures que les autres mammifères. Lorsque nous recherchons un partenaire, nous ne nous attachons pas à de simples attributs comme la santé, mais nous nous renseignons auprès de son entourage et de ses parents, nous observons sa conduite et son langage, et nous jugeons ses qualités morales. Nous portons notre attention à ces attributs secondaires et tertiaires afin de prédire comment le partenaire potentiel se comportera dans le futur. Les opérateurs boursiers recherchent des structures dans les fluctuations financières. Les mathématiciens recherchent des structures dans les chiffres et les équations. Les astronomes recherchent des structures dans le mouvement des planètes et des étoiles. Notre néocortex surdimensionné nous permet de considérer notre foyer comme faisant partie d'une ville, qui fait partie d'un pays, qui fait partie d'une planète, qui n'est qu'une infime partie d'un vaste univers : des structures dans des structures. Aucun autre animal n'est capable de ruminer à cette profondeur. Je suis persuadé que ma chatte Keo n'a aucune idée du monde en dehors de son territoire.

La seconde différence entre l'intelligence humaine et celle des animaux est le langage. Des ouvrages entiers ont été écrits sur les propriétés supposées uniques du langage et comment nous l'avons développé. Toutefois, le langage s'insère parfaitement dans le cadre de mémoire-prédiction. Ecrits ou parlés, les mots ne sont que des patterns faisant partie du monde, au même titre que les mélodies, les automobiles et les maisons. La syntaxe et la sémantique d'une langue ne diffèrent pas de la structure hiérarchique des autres objets du quotidien. Et de la même manière,

nous associons le sifflet d'une locomotive avec l'image mentale d'un train, les mots prononcés avec la mémoire que nous avons de leurs contreparties physiques ou sémantiques. C'est par le langage qu'un être humain peut faire appel à la mémoire et créer de nouvelles juxtapositions d'objets mentaux dans l'esprit d'un autre humain. Le langage est pure analogie, et c'est par lui que nous pouvons faire en sorte que d'autres humains fassent l'expérience et apprennent des choses qu'ils n'ont peut-être jamais vues véritablement. Le développement du langage exige un vaste néocortex capable de gérer la structure imbriquée de la syntaxe et de la sémantique. Il exige aussi un cortex moteur pleinement développé ainsi qu'une musculature nous permettant d'émettre des sons sophistiqués, parfaitement articulés, et d'effectuer des gestes compliqués et précis. Le langage nous permet de saisir les patterns que nous avons appris au cours de notre vie et les transmettre à nos enfants et à notre entourage. Qu'il soit parlé, écrit ou incorporé aux traditions culturelles, le langage est le moyen par lequel nous transmettons de génération en génération ce que nous savons du monde. Aujourd'hui, les communications par le papier, le câble ou les faisceaux hertziens nous permettent de partager notre savoir avec des millions de personnes de par le monde. Les animaux dépourvus de langage ne transmettent que peu d'informations à leur progéniture. Un rat peut apprendre de nombreux patterns au cours de sa vie, mais il ne pourra pas transmettre des informations nouvelles et détaillées (« Ecoute-moi bien, face en poils, je vais t'expliquer comment mon père m'a appris à reconnaître la mort-aux-rats »).

L'intelligence peut être retracée sur trois époques, chacune ayant recours à la mémoire et à la prédiction. La première remonte à l'ère où les espèces se servaient de l'ADN comme support de la mémoire. Au cours de leur vie, les individus ne pouvaient ni apprendre ni s'adapter. Ils ne pouvaient transmettre à leur progéniture qu'une mémoire du monde inscrite dans l'ADN, au travers de leurs gènes.

La deuxième époque commença lorsque la nature inventa le système nerveux modifiable, capable de former rapidement des mémorisations. Au cours de sa vie, un animal put désormais apprendre ce qu'était la structure de son environnement et adapter son comportement. Mais il lui était toujours impossible de communiquer son savoir à ses rejetons autrement que par l'observation directe. La création et l'expansion du néocortex se sont produites lors de cette deuxième époque, mais ne l'ont pas défini.

La troisième et dernière époque est propre à l'être humain. Elle commence avec l'invention du langage et l'expansion du cortex. Nous autres humains sommes capables d'en apprendre beaucoup sur la structure du monde, tout au long de notre existence, et nous pouvons efficacement communiquer cette connaissance à autrui grâce au langage. Vous et moi participons actuellement à ce processus. J'ai consacré une bonne partie de ma vie à rechercher des structures dans le cerveau et savoir comment elles conduisent à la pensée et à l'intelligence. Par ce livre, je diffuse mon savoir auprès de vous. Bien sûr, je n'aurais jamais pu le faire si je n'avais eu moi-même accès au savoir accumulé par des centaines de savants, qui ont eux-mêmes appris auprès d'autres au cours des âges. J'ai eu la faculté de recueillir et d'assimiler ce que d'autres ont écrit au sujet de leurs propres pensées et observations.

Nous sommes devenues les créatures les plus aptes à s'adapter de la planète, et les seules capables de transmettre nos connaissances sur la planète entière, à destination de nos semblables. La population humaine connaît une évolution effrénée car nous pouvons apprendre, exploiter la structure du monde et communiquer tout cela aux autres humains. Nous pouvons prospérer n'importe où, que ce soit dans la forêt tropicale, dans le désert, dans la toundra glacée ou dans une jungle de béton. La combinaison d'un vaste cortex et du langage a conduit aux réussites exponentielles de notre espèce.

## QU'EST LA CRÉATIVITÉ ?

On me demande souvent ce qu'est la créativité, peut-être parce que beaucoup de gens pensent qu'elle est un effet de notre esprit que la machine ne saurait produire, et aussi parce qu'elle représente l'un des défis posés à la construction d'une machine intelligente. Alors, qu'est la créativité ? Nous avons déjà rencontré la réponse plusieurs fois dans cet ouvrage. La créativité n'est pas quelque chose qui se produit dans une région particulière du cortex. Elle n'a rien à voir avec les émotions ou l'équilibre, qui sont enracinés dans des structures ou des circuits situés hors du cortex. La créativité est plutôt une propriété inhérente à chaque région corticale. C'est un composant indispensable de la prédiction.

Comment cela peut-il être vrai ? La créativité n'est-elle pas quelque qualité extraordinaire exigeant une intelligence élevée et découlant d'un don ? Pas vraiment. La créativité peut être définie tout simplement par la capacité à faire des prédictions par analogie, quelque chose qui se produirait partout dans le cerveau et que vous faites continuellement quand vous êtes éveillé. La créativité est sans solution de continuité. Elle s'étend des actes quotidiens de la perception par les régions sensorielles du cortex (écouter une chanson interprétée sur une nouvelle clé) aux actes les plus difficiles, les plus rares du génie se manifestant aux niveaux les plus élevés du cortex (la composition d'une symphonie d'une manière complètement innovante). A un niveau fondamental, tous les actes de la perception sont semblables aux manifestations rares du génie. A ce détail près que les actes du quotidien sont si communs que nous ne les remarquons même pas.

Vous possédez actuellement une connaissance de base concernant la création des mémoires invariantes, concernant la manière dont nous les utilisons pour faire des prédictions, et aussi comment nous faisons des prédictions sur les événements futurs qui sont toujours quelque peu différents de ce que nous avons expérimenté par le passé. Rappelez-vous aussi que les mémoires invariantes sont des séquences d'événements. Nous procédons à des

prédictions en combinant le rappel de la mémoire invariante de ce qui doit se produire ensuite avec les détails se rapportant à ce moment à venir (souvenez-vous de l'histoire de la prédiction de l'heure à laquelle le train arrivera). La prédiction est l'application de séquences de mémoire invariante à de nouvelles situations. Par conséquent, toutes les prédictions corticales sont des prédictions par analogie. Nous prédisons le futur par analogie avec le passé.

Supposons que vous soyez invité à dîner dans un restaurant que vous ne connaissez pas, et que vous désiriez vous laver les mains. Même si vous n'avez jamais été dans ces lieux auparavant, votre cerveau prédit qu'il doit y avoir un endroit, quelque part dans le restaurant, équipé d'un lavabo prévu pour le lavage des mains. Comment le sait-il ? Parce que les autres restaurants dans lesquels vous êtes allé ont des toilettes et que, par analogie, cela doit aussi être le cas de celui-ci. De plus, vous savez où regarder et ce que vous cherchez. Vous prédisez qu'il doit y avoir une porte ou une enseigne avec un pictogramme dessus évoquant un homme ou une femme. Vous prédisez qu'elle doit se trouver à l'arrière de la salle, vers le bar ou dans un couloir, mais généralement hors de la vue des convives. Là encore, vous n'avez jamais été dans ce restaurant, mais par analogie avec d'autres établissements, vous êtes capable de trouver ce que vous cherchez. Vous ne regardez pas autour de vous au hasard. Vous recherchez les patterns qui vous permettront de trouver rapidement les toilettes. Ce type de comportement est un acte créatif : il prédit le futur par analogie avec le passé. Ce n'est normalement pas sous cette forme que nous imaginons la créativité, mais c'en est un petit peu.

J'ai récemment acheté un vibraphone. Nous avons un piano, mais je n'avais jamais joué du vibraphone auparavant. Le jour où je l'ai amené à la maison, j'ai pris une partition sur le piano, je l'ai placée sur un pupitre derrière le vibraphone et j'ai commencé à jouer des mélodies simples. Ma virtuosité n'était pas extraordinaire. Mais, fondamentalement, c'était un acte créatif. Réfléchissez à tout ce qu'il implique. L'instrument est très différent du

piano : le vibraphone est doté de lamelles dorées alors que le piano a des touches noires et blanches. Les lamelles dorées sont larges et leur longueur varie progressivement, tandis que les touches sont étroites et de deux tailles. Le vibraphone possède deux jeux de lamelles, mais sur le piano, les touches noires et blanches alternent. Sur un instrument, j'agite des baguettes, sur l'autre mes doigts. Je me tiens debout devant l'un, je m'assieds à l'autre. Les muscles et la gestuelle mis en œuvre pour jouer du vibraphone sont différents de ceux pour jouer du piano.

Comment me fut-il possible de jouer sur un instrument qui ne m'était pas familier ? La réponse est que mon cortex voyait une analogie entre les touches du piano et les lamelles du vibraphone. C'est cette similarité qui me permit de jouer un air. Ce n'est pas très différent que de chanter sur une autre clé. Dans les deux cas, vous savez ce que vous avez à faire par analogie avec ce qui a été appris dans le passé. Je suppose que pour vous aussi la similarité entre ces deux instruments peut paraître évidente, mais uniquement parce que votre cerveau y voit spontanément des analogies. Essayez de programmer un ordinateur pour trouver des similarités entre des objets comme un vibraphone et un piano, et vous constaterez combien c'est difficile. La prédiction par analogie — la créativité — est si profondément inscrite en nous-même qu'habituellement nous ne la remarquons pas.

Nous nous rendons cependant compte de la créativité lorsque notre système de mémoire-prédiction opère à un plus haut niveau d'abstraction, lorsqu'il fait des prédictions peu communes à l'aide d'analogies atypiques. Par exemple, la plupart d'entre nous reconnaîtront qu'un mathématicien qui démontre un théorème difficile fait preuve de créativité. Examinons de près ce qui se passe au niveau de son mental. Le mathématicien considère l'équation et se demande comment il va résoudre le problème. Si la réponse n'est pas évidente, il reformule l'équation. En l'écrivant sous une autre forme, il verra le problème sous un autre angle. Il regarde de nouveau l'équation et, soudain, une partie lui

semble familière. Il pense : « Je reconnais cette structure. Elle ressemble à celle d'une autre équation sur laquelle j'avais travaillé plusieurs années auparavant. » Il fait ensuite une prédiction par analogie : « Peut-être parviendrais-je à résoudre cette nouvelle équation à l'aide des mêmes techniques que j'avais utilisées autrefois. » Il est à présent capable de résoudre le problème par analogie avec une situation précédemment apprise. C'est un acte créatif.

Mon père était atteint d'un mystérieux trouble sanguin que son médecin ne parvenait pas à diagnostiquer, ce qui l'empêchait de proposer un traitement. Pour en savoir plus, il examina des résultats d'analyses faites sur plusieurs mois pour voir s'il pouvait y découvrir des formes caractéristiques, autrement dit des patterns (mon père avait imprimé des graphiques, ce qui permit au médecin de visualiser clairement les chiffres). Bien que les symptômes ne correspondissent pas exactement à ceux de maladies connues, il y découvrit quelques similarités. Le médecin finit par établir un traitement fondé sur une synthèse de ceux qu'il avait précédemment prescrits. Reconnaître des patterns de ce genre exige une excellente connaissance des maladies rares.

Les métaphores de Shakespeare sont des modèles de créativité : « L'amour est une fumée faite de la vapeur des soupirs », « La philosophie, ce doux lait de l'adversité », « Il y a des poignards dans les sourires »… C'est à cause de telles métaphores, évidentes quand vous les entendez mais difficiles à inventer, que Shakespeare est considéré comme un génie de la littérature. Pour les créer, il lui fallut découvrir une succession d'intelligentes analogies. Quand il écrit qu'« il y a des poignards dans les sourires », il ne parle ni d'armes blanches ni de physionomie. Les poignards sont assimilés à l'intention de faire du mal, les sourires assimilés à la fourberie. Deux analogies clairvoyantes en quelques mots seulement ! Du moins, c'est ainsi que je l'interprète. Les poètes ont un don pour corréler des mots ou des concepts apparemment étrangers d'une manière qui les éclaire d'un jour nouveau. Les

analogies inattendues sont pour eux un moyen de révéler des structures de plus haut niveau.

Les œuvres d'art hautement créatives sont appréciées car elles violent nos prédictions. Quand un acteur est utilisé à contre-emploi, quand le scénario dérape intelligemment, quand un effet spécial surprend, vous aimez cela parce que ces choix artistiques rompent la routine. La peinture, la musique, la poésie, le roman — toutes les formes d'expression artistique — tentent de briser les conventions et de contrevenir aux attentes du public. Il existe des tensions antagonistes dans ce qui fait le génie d'une œuvre. Nous voudrions que l'art nous soit familier tout en désirant qu'il soit unique et inattendu. Trop de familiarité affadit ou rend anodin ; trop d'inhabituel ébranle et rend difficile à apprécier. L'art brise les patterns attendus tout en en proposant d'autres. Prenons la musique classique. La musique la plus appréciée séduit d'abord à un niveau élémentaire : un bon rythme, une mélodie simple et de belles phrases musicales. N'importe qui peut la comprendre et l'apprécier. Elle est cependant quelque peu différente et imprévue ; plus vous l'écoutez, plus vous y découvrez des patterns dans des parties inattendues, comme la répétition d'harmonies inhabituelles ou des changements de clé. Il en est de même pour la littérature ou les classiques du cinéma. Plus vous les lisez ou les regardez, plus vous y découvrez des détails créatifs ainsi que la complexité de la structure.

Vous avez sans doute connu cette situation où vous regardez quelque chose en vous demandant où vous avez bien pu l'avoir vu. Vous êtes certain d'avoir déjà rencontré ce pattern auparavant, ailleurs… Vous avez pu ne pas avoir tenté de résoudre le problème, car c'est seulement une représentation invariante qui, dans votre cerveau, a été activée par une situation toute nouvelle. Vous avez vu une analogie entre deux événements qui ne sont normalement pas liés. Je pourrais reconnaître que promouvoir une idée scientifique est comparable à vendre une idée commerciale, ou qu'imposer une réforme politique est comme élever des

enfants. Si j'étais un poète, voilà, j'aurais une nouvelle métaphore. Si j'étais chercheur ou ingénieur, j'aurais trouvé une solution à un problème qui perdure. Etre créatif, c'est mêler et faire concorder les patterns de tout ce dont vous avez déjà fait l'expérience dans votre vie, ou que vous avez déjà connu. Cela revient à dire : « C'est un peu ça. » Le mécanisme neuronal qui permet cela est partout dans le cortex.

## CERTAINS SONT-ILS PLUS CRÉATIFS QUE D'AUTRES ?

Une question que j'entends très souvent, liée à la précédente, est : « Si tous les cerveaux sont par nature créatifs, pourquoi existe-t-il des différences dans la créativité ? » Le cadre de mémoire-prédiction propose deux réponses, l'une liée à la nature, l'autre à l'acquis.

En ce qui concerne l'acquis, tout le monde fait ses propres expériences de la vie. C'est pourquoi chacun de nous développe différents modèles de mémoire du monde dans son cortex, et procède à différentes analogies et prédictions. Si j'ai été exposé à la musique, je serai capable de chanter en diverses clés et de jouer des mélodies simples sur divers instruments. Si je n'ai jamais été exposé à la musique, je ne pourrai pas faire ces efforts prédictifs. Si j'ai étudié la physique, je serai capable d'expliquer le comportement des objets qui nous entourent par analogie avec les lois de la physique. Si j'ai grandi avec des chiens, je serai apte à déceler les analogies concernant les chiens et je prédirai mieux leur comportement. Certaines personnes sont plus créatives dans les contextes sociaux, ou dans les langues, ou les mathématiques ou la diplomatie, selon l'environnement dans lequel elles ont grandi. Nos prédictions, et par conséquent nos talents, sont fondées sur nos expériences.

Au Chapitre 6, j'ai décrit comment les mémorisations sont déplacées vers le bas dans la hiérarchie corticale. Plus vous êtes exposé à certains patterns, plus la mémoire de ces patterns est reformée à des niveaux inférieurs. Ceci permet d'apprendre les

relations entre les objets abstraits d'ordre élevé situés tout en haut de la hiérarchie. C'est l'essence même de la compétence. Un expert dans un domaine est quelqu'un qui, à travers la pratique et l'exposition répétées, est capable de reconnaître les patterns les plus subtils qui échappent au novice, comme la forme de l'aileron d'une voiture de la fin des années 1950, ou la taille de la tache sur le bec d'une mouette. Un expert reconnaît des patterns placés sur des patterns, hiérarchiquement parlant. Il existe au bout du compte une limite physique à ce que nous pouvons apprendre, qui découle de la taille du cortex. Mais notre cortex, à nous autres humains, est vaste comparé à celui d'autres espèces, et nous bénéficions d'une remarquable flexibilité envers ce que nous pouvons apprendre. Tout dépend de ce à quoi nous sommes exposés tout au long de notre existence.

En ce qui concerne la nature, chaque cerveau présente des variations physiques. Certaines différences sont assurément d'origine génétique, comme la taille des régions (d'un individu à un autre, des différences pouvant aller jusqu'à trois replis peuvent se manifester dans l'ensemble de l'aire V1) et la latéralité hémisphérique (le corps calleux qui relie les deux hémisphères cérébraux tend à être plus épais chez la femme que chez l'homme). Parmi les êtres humains, certains cerveaux ont vraisemblablement plus ou moins de cellules ou différents types de connexions. Il est peu probable que le génie créatif d'Albert Einstein soit purement le produit de l'environnement stimulant qu'était le bureau des brevets où il travaillait lorsqu'il était jeune homme. De récentes analyses de son cerveau — que l'on pensait perdu, mais qui avait été conservé dans un bocal de formol — ont révélé que son cerveau était quelque peu particulier. Il comprenait davantage de cellules nourricières, ou cellules gliales, par neurone que la moyenne. Il présentait des tracés de sillons inhabituels dans les lobes pariétaux, une région supposée importante pour les capacités mathématiques et le raisonnement spatial. Sa surface était aussi 15 % plus grande que la plupart des cerveaux. Nous ne

saurons peut-être jamais quelle était la source de la créativité et de l'intelligence d'Einstein, mais l'on peut sans risque affirmer que le facteur génétique n'y est pas étranger.

Quelles que soient les différences entre un cerveau génial et un cerveau moyen, nous sommes tous créatifs. Et c'est par la pratique et par l'étude que nous pouvons améliorer nos aptitudes et nos talents.

## PEUT-ON DÉVELOPPER LA CRÉATIVITÉ ?

Oui, sans aucun doute. Je me suis aperçu qu'il existe deux moyens de favoriser la découverte d'analogies fructueuses lorsque vous êtes confronté à un problème. Le premier, c'est de partir du principe qu'il existe toujours une solution. Les gens abandonnent facilement. Or, vous devez vous persuader que la solution n'attend que d'être découverte, et vous devez penser longuement au problème, tout le temps.

Deuxièmement, vous devez laisser votre esprit vagabonder, lui laisser le champ libre et lui accorder du temps pour découvrir la solution. La recherche de la solution consiste littéralement à trouver un pattern dans le monde, ou un pattern stocké dans le cortex, analogue au problème posé. Si vous butez sur le problème, le modèle de mémoire-prédiction suggère que vous testiez différents moyens afin d'augmenter les probabilités de découvrir une analogie avec une expérience passée. Si vous vous contentez de vous asseoir et de rester le regard perdu dans les nuages, vous n'irez pas très loin. Essayez de vous attacher à une partie du problème et abordez-le sous différents angles : littéralement et figurativement. Quand je joue au Scrabble, je change constamment l'ordre de mes pièces. J'espère toujours qu'avec un peu de chance les lettres formeront un mot, mais le seul fait de modifier leur succession peut me rappeler un mot ou évoquer une partie de mot qui pourrait contribuer à la solution. Si vous regardez un dessin dont la signification vous échappe, essayez de le regarder à l'envers, de modifier ses couleurs ou de changer d'angle. Par exemple, quand je m'interrogeais

sur le nombre de patterns différents en V1 qui pourraient former des représentations invariantes en IT, j'étais bloqué. J'ai donc pris le problème à rebours, en me demandant comment un pattern constant en IT pourrait conduire à différentes prédictions en V1. Prendre le problème dans l'autre sens se révéla aussitôt positif, ce qui m'amena à penser que V1 ne devait pas être considérée comme une seule région corticale.

Quand vous butez sur un problème, prenez un peu de recul. Faites autre chose puis reprenez le problème en le reformulant. En procédant ainsi suffisamment de fois, un déclic se produira à un moment ou à un autre. Cela peut être une question de jours ou de semaines, mais cela finira par arriver. Le but de l'opération est de découvrir une situation analogue quelque part dans votre expérience passée ou présente. Pour réussir, vous devez réfléchir souvent au problème, mais aussi vous adonner à d'autres activités afin que le cortex puisse se consacrer à la recherche d'un élément mémorisé analogue.

Voici un autre exemple illustrant la redéfinition d'un problème conduisant à une solution innovante. En 1994, mes collègues et moi essayions de trouver comment saisir du texte dans des ordinateurs de poche. Tous s'étaient focalisés sur un logiciel de reconnaissance de l'écriture manuscrite. Ils me disaient : « Tu écris sur du papier, donc tu dois pouvoir écrire de la même manière sur un écran. » Ceci fut malheureusement très difficile à réaliser. Voilà encore une activité dans laquelle un ordinateur n'excelle pas, alors que pour notre cerveau c'est très simple. La raison en est que le cerveau fait appel à la mémoire et exploite le contexte courant pour prédire ce qui sera écrit. Les mots et les lettres, méconnaissables par eux-mêmes, sont facilement reconnus dans leur contexte. Des patterns concordant avec ceux stockés dans l'ordinateur ne sont pas suffisants pour assurer la tâche. J'avais conçu plusieurs ordinateurs qui utilisaient la reconnaissance de l'écriture manuscrite traditionnelle, mais le résultat ne fut jamais vraiment probant.

Je m'étais battu pendant plusieurs années pour améliorer le logiciel de reconnaissance et j'étais dans une impasse. Un jour, je décidai de prendre du recul et d'aborder le problème différemment. Je me mis à rechercher des problèmes analogues. Je me dis : « Comment saisissons-nous du texte dans un ordinateur de bureau ? En le tapant sur un clavier. Comment savons-nous de quelle manière il faut taper sur un clavier ? Eh bien, à vrai dire, ce n'est pas facile. C'est une invention récente et il faut un certain temps pour apprendre. La dactylographie sur un clavier hérité des machines à écrire est ardue et n'a rien d'intuitif. Ce n'est pas du tout comme l'écriture. Pourtant, des millions de gens apprennent à taper. Pourquoi ? Parce que ça fonctionne. » Mes pensées se poursuivaient par analogies : « Peut-être puis-je envisager un système de saisie de texte qui ne serait pas forcément intuitif, qu'il faudrait apprendre, mais que les gens adopteraient parce qu'il fonctionne. »

C'est effectivement par ce processus mental que je m'en sortis. J'avais utilisé la dactylographie comme analogie pour trouver comment saisir du texte avec un stylet afin de l'afficher à l'écran. J'avais reconnu que les gens étaient disposés à apprendre une tâche difficile, la dactylographie, car c'était un moyen fiable et rapide pour entrer du texte dans une machine. Par conséquent, si nous pouvions mettre au point une nouvelle technique de saisie de texte à l'aide d'un stylet, qui serait aussi rapide et fiable, les gens l'adopteraient même si un apprentissage était nécessaire. C'est ainsi que je conçus un alphabet qui transcrirait fidèlement ce que vous écririez en texte informatique. Nous l'appelâmes Graffiti. Avec les systèmes traditionnels de reconnaissance de l'écriture manuscrite, quand l'ordinateur interprète mal votre écriture, vous ne savez pas pourquoi. En revanche, le système Graffiti produit toujours la lettre correcte, à moins que vous ne fassiez une erreur dans l'écriture. Notre cerveau déteste l'imprédictibilité, et c'est pour la même raison que beaucoup de gens détestent les systèmes traditionnels de reconnaissance de l'écriture manuscrite.

Nombreux sont ceux qui estimèrent que Graffiti était une idée extraordinairement stupide. Elle allait à l'encontre de tout ce qu'ils pensaient du fonctionnement supposé de l'ordinateur. Le mot d'ordre, à cette époque, était que l'ordinateur doit s'adapter à l'utilisateur, et non l'inverse. Mais j'étais persuadé que les gens accepteraient cette nouvelle technique de saisie de texte par analogie avec le clavier. Il s'avéra que Graffiti était une bonne solution qui fut appréciée. A ce jour, j'entends toujours dire que ce sont les ordinateurs qui doivent d'adapter aux utilisateurs. Ce n'est pas toujours vrai. Notre cerveau préfère des systèmes réguliers et prédictibles, et nous aimons acquérir de nouvelles aptitudes.

## LA CRÉATIVITÉ PEUT-ELLE M'INDUIRE EN ERREUR ? PUIS-JE ME TROMPER ?

Une analogie erronée est toujours risquée. L'histoire de la science abonde d'exemples de séduisantes analogies qui se sont révélées fausses. Par exemple, le célèbre astronome allemand Johannes Kepler s'était convaincu que les orbites des six planètes connues à cette époque étaient définies par des solides platoniciens. Les solides platoniciens sont les seules formes tridimensionnelles pouvant être construites entièrement à partir de polygones réguliers. Il en existe cinq : le tétraèdre (4 triangles équilatéraux), le cube (6 faces carrées), l'octaèdre (8 triangles équilatéraux), le dodécaèdre (12 pentagones) et l'icosaèdre (20 triangles équilatéraux). Ils ont été découverts par les Grecs anciens qui s'interrogeaient sur les relations entre les mathématiques et le cosmos.

Comme tous les érudits de la Renaissance, Kepler avait été profondément influencé par la pensée grecque. Il lui semblait que ce n'était pas par coïncidence qu'il existe cinq solides platoniciens et six planètes. Ainsi qu'il le décrivait dans son ouvrage *Mysterium cosmographicum* (*Le Mystère cosmographique*) publié en 1596 : « La Terre est le cercle qui mesure tout : circonscris-lui le dodécaèdre. Le cercle comprenant ce dernier sera Jupiter : à Jupiter circonscris le cube. Le cercle comprenant ce dernier sera Saturne : maintenant

inscris l'icosaèdre à la Terre. Le cercle inscrit dans celui-ci sera Vénus. A Vénus inscris l'octaèdre. Le cercle inscrit dans celui-ci sera Mercure. Tu tiens là la raison du nombre des planètes. » Il avait vu là une belle analogie, mais hélas complètement erronée.

Johannes Kepler avait imaginé les orbites des planètes sous la forme de solides platoniciens imbriqués, tous centrés sur le Soleil. Il avait pris pour base la sphère définie par l'orbite de Mercure et l'avait inscrite dans un octaèdre. Le sommet de l'octaèdre définissait une sphère plus grande correspondant à l'orbite de Vénus. Englobant l'orbite de Vénus, il plaça un icosaèdre dont le sommet extérieur atteignait l'orbite de la Terre. La progression se poursuivit pour les planètes extérieures : un dodécaèdre tracé autour de l'orbite terrestre donnait celle de Mars, un tétraèdre autour de l'orbite martienne donnait l'orbite de Jupiter et enfin, un cube autour de l'orbite jovienne donnait l'orbite de Saturne. Cette vision était belle et élégante. L'imprécision des données astronomiques de l'époque permettait de croire à ce système. Des années plus tard, Kepler se rendit compte de son erreur, lorsqu'il hérita des relevés astronomiques fort précis de son collègue décédé, l'astronome danois Tycho Brahe : il apprit que les orbites des planètes ne sont pas circulaires, mais elliptiques.

L'enthousiasme de Kepler servit d'avertissement aux savants, et aussi à tous les penseurs. Le cerveau est un organe qui échafaude des modèles et procède à des prédictions créatives, mais les modèles et les prédictions peuvent facilement être aussi trompeurs que valides. Notre cerveau recherche en permanence des patterns et cherche à faire des analogies. Si des corrélations correctes ne peuvent être trouvées, le cerveau est plus qu'heureux d'accepter des fausses. Les pseudosciences, les tendances dogmatiques, la foi et l'intolérance découlent souvent de fausses analogies.

## QU'EST LA CONSCIENCE ?

C'est une des questions que les neurobiologistes redoutent, à mon avis sans raison. Si des scientifiques comme Christof Koch

abordent volontiers la question de la conscience — la conscience psychologique, c'est-à-dire celle de l'activité psychique, et non la conscience morale dont il n'est pas question dans ces pages —, la plupart considèrent qu'elle relève de la philosophie, voire des pseudosciences. Je pense qu'elle mérite que l'on s'y intéresse, ne serait-ce que pour l'unique raison que beaucoup d'entre nous s'interrogent à son sujet. Je ne saurais apporter une réponse satisfaisante et complète, mais je pense que la mémoire et la prédiction peuvent y contribuer. Voici d'abord un aperçu de l'énigme telle qu'elle apparaît dans une conversation.

Il y a quelque temps, j'assistais à une conférence scientifique dans la bucolique région de Long Island Sound. Au coucher du soleil, une douzaine d'entre nous s'assirent au bord d'un quai, face à l'eau, leur verre de vin à la main, discutant avant le dîner et les sessions du soir. La conversation porta bientôt sur la conscience. Comme je le disais, les neurobiologistes parlent peu de cette question, mais la beauté du site aidant, et sans doute aussi le vin du pays, ce sujet fut abordé.

Une scientifique britannique parla longuement et finit par asséner : « Nous ne saurons bien sûr jamais ce qu'est la conscience. » Je n'étais pas d'accord : « La conscience n'est pas un problème primordial. Je pense qu'il s'agit tout simplement de ce que nous éprouvons du fait d'avoir un cortex. » Un silence tomba sur le groupe, puis plusieurs scientifiques tentèrent de me reprendre sur mon erreur : « Reconnaissez combien le monde est vivant et beau. Comment pouvez-vous nier que votre conscience perçoit le monde ? Vous devez admettre que vous éprouvez quelque chose de particulier. » Pour marquer le coup, je rétorquai : « Je ne vois pas de quoi vous parlez. Si j'en crois ce que vous dites de la conscience, je dois en conclure que je suis différent de vous. Je ne ressens pas ce que vous ressentez. Peut-être ne suis-je pas un être conscient, mais un zombi. » Aux Etats-Unis, les zombis sont souvent mis à contribution lorsque les philosophes dissertent sur la conscience. Dans les croyances antillaises, ce sont des revenants, mais dans le langage

courant, ce terme désigne un individu privé de conscience. Sa démarche est mécanique et il ne sait pas où il va.

La scientifique britannique me regarda : « Bien sûr que vous êtes conscient. »

« Non, je ne le pense pas. C'est peut-être ce qu'il vous semble, mais je ne suis pas un être humain conscient. Que cela ne vous tourmente pas, je me sens très bien ainsi. »

Elle répondit : « Ainsi donc, vous ne percevez pas toutes ces merveilles ? » Elle étendit son bras vers l'eau mordorée tandis que le soleil disparaissait à l'horizon, le ciel prenant des teintes roses irisées.

« Evidemment que je vois tout ça, et alors ? »

« Alors, comment expliquez-vous votre expérience subjective ? »

Je répondis : « Oui, je sais que je suis ici. J'ai en mémoire des événements comme ce coucher de soleil. Mais je n'ai pas l'impression qu'il se passe quelque chose de spécial. Alors, si vous, vous ressentez quelque chose de particulier, peut-être n'en suis-je tout bonnement pas conscient. » J'étais en train de l'épingler sur ce qui, à son avis, était si miraculeux et inexplicable dans la conscience. J'essayais de l'obliger à définir la conscience psychologique.

Nous continuâmes à ergoter sur ce genre d'arguments jusqu'au moment du dîner. Je ne pense pas avoir changé l'avis de quiconque sur ce qu'ils pensaient de l'existence et de la signification de la conscience, mais j'avais tenté de leur faire comprendre que pour la plupart des gens, la conscience est comme une sauce nappée par-dessus le cerveau physique. Vous avez un cerveau constitué de cellules et la conscience a été versée par-dessus : ce serait ça la condition humaine. De ce point de vue, la conscience serait une mystérieuse entité distincte du cerveau. C'est pourquoi ceux que nous appelons par dérision les zombis auraient un cerveau, mais pas de conscience. Ils ont toute la mécanique requise — les neurones, les synapses… —, mais il leur manque le coulis. Ils peuvent faire tout ce qu'un humain fait. Vus de l'extérieur, il serait difficile de les différencier des humains.

L'idée que la conscience serait une particularité provient des croyances anciennes dans l'élan vital, une force spéciale qui, pensait-on autrefois, animait tout être vivant. Cette force de vie expliquait à l'époque la différence entre les cailloux et les plantes, ou encore le métal et une jeune fille. Quelques personnes y croient encore. Nous en savons suffisamment aujourd'hui sur les différences entre l'inanimé et l'animé pour comprendre qu'une « sauce spéciale » est inutile. Nous en savons long sur l'ADN, le repliement des protéines, la transcription des gènes et le métabolisme. Bien que nous ne sachions pas tout des mécanismes de la vie, nous en savons suffisamment sur la biologie pour nous passer de l'aspect magique. De même, plus personne ne soutient qu'il faut recourir à la magie ou à un esprit surnaturel pour faire bouger les muscles. Nous avons dans notre corps des protéines repliables qui rétractent les longues molécules des tissus musculaires. La littérature à ce sujet ne manque pas.

Néanmoins, beaucoup persistent à croire que la conscience est différente et qu'elle ne saurait être expliquée par des termes biologiques réducteurs. Là encore, j'insiste sur le fait que je ne suis pas un spécialiste de la conscience. Je n'ai pas lu tout ce qu'en pensent les philosophes. Mais j'ai quelques idées sur ce que, à mon avis, les gens confondent allègrement dans ce débat. Je crois que la conscience est simplement ce que nous éprouvons du fait d'avoir un cortex. Mais je crois que nous pouvons aller plus loin. La conscience peut être scindée en deux grandes catégories. L'une est assimilable à la conscience de sa propre activité psychique, une notion facile à comprendre. L'autre est le *quale*, l'idée selon laquelle les effets subjectifs associés à la sensation seraient en réalité indépendants des inputs sensoriels. Le quale est la partie ardue de l'exposé qui suit.

Pour la plupart des gens, le mot *conscience* se rapporte à la première catégorie : « As-tu été conscient que tu m'as croisé sans me dire bonjour ? », « Tu étais conscient quand t'es tombé du lit la nuit dernière ? », « Tu n'es pas conscient quand tu dors ». Dans le

contexte qui nous intéresse, le mot « conscience » est synonyme de « formation de mémoire déclarative ». Les mémoires déclaratives sont les mémorisations dont vous pouvez vous souvenir et en parler avec quelqu'un d'autre. Elles peuvent être exprimées verbalement. Si vous me demandez où j'ai passé le dernier week-end, je peux vous le dire. C'est une mémoire déclarative. Si vous me demandez comment on garde l'équilibre sur une bicyclette, je vous dirai que c'est en tenant le guidon et en pédalant, mais je ne saurais expliquer exactement ce qui se passe. L'équilibre à vélo dépendant plutôt d'une activité neuronale dans le cerveau archaïque, ce n'est pas une mémoire déclarative.

Je vous propose une petite expérience de la pensée qui montre en quoi la notion courante de la conscience est identique à la formation de mémoires déclaratives. Rappelez-vous que la mémoire est censée résider dans les modifications physiques des synapses et des neurones qui les relient. Par conséquent, si j'avais un moyen d'annuler ces changements physiques, votre mémoire serait effacée. Imaginez que je dispose à présent d'un commutateur capable de ramener votre cerveau à l'état physique exact qui était le sien à un moment du passé. Peu importe qu'il s'agisse de plusieurs heures ou de journées entières. J'actionne le commutateur de ma machine à remonter dans la mémoire et vos synapses et neurones sont réinitialisés à un état antérieur. Ce faisant, j'efface de votre mémoire tout ce qui s'est produit entre-temps.

Supposons que vous vivez le jour d'aujourd'hui et que vous vous réveillez demain. Mais juste au moment où vous ouvrez l'œil, je bascule le commutateur et j'efface les dernières vingt-quatre heures. Vous n'auriez absolument aucun souvenir de la journée précédente. Pour votre cerveau, hier n'aurait jamais existé. Je vous dirais que nous sommes mercredi et vous protesteriez : « Pas du tout, nous sommes mardi. J'en suis certain. Quelqu'un a changé le calendrier. Quoi qu'il en soit, c'est aujourd'hui mardi. A quoi jouez-vous ? » Mais tous ceux que vous auriez rencontrés mardi affirmeraient que vous avez été conscient

toute la journée. Ils vous ont vu, ont déjeuné avec vous et ont discuté avec vous. Vous ne vous en souvenez pas ? Vous direz que non, que cela n'est jamais arrivé. Pour finir, je vous montre une vidéo où l'on vous voit déjeuner avec ces gens. Peu à peu, vous vous persuadez que ce jour a existé, bien que vous n'en ayez plus la mémoire. C'est comme si, pendant une journée, vous étiez un zombi, un être dépourvu de conscience psychologique. Et pourtant, ce mardi-là, vous étiez conscient. Votre conviction d'avoir été conscient n'a disparu que quand votre mémoire déclarative a été effacée.

La question du quale est plus difficile à cerner. Ce concept est souvent exprimé au travers de questions apparentées à la philosophie zen, comme « Pourquoi le rouge est-il rouge et le vert est-il vert ? Le rouge est-il pour moi ce qu'il est pour vous ? Pourquoi le rouge est-il émotionnellement connoté ? Il présente pour moi une indicible qualité de sentimentalisme, mais quelle sorte de sentimentalisme suscite-t-il en vous ? »

Je trouve de telles définitions difficiles à transposer dans le domaine de la neurobiologie. C'est pourquoi je vais reformuler la question. Pour moi, une interrogation équivalente, mais que je trouve aussi difficile à expliquer, est : « Pourquoi nos sens semblent-ils qualitativement différents ? Pourquoi la vue semble-t-elle différente de l'ouïe et l'ouïe différente du toucher ? Si le cortex est partout pareil, s'il utilise les mêmes procédés, s'il ne fait qu'exploiter des patterns, si aucun son ni aucune lumière ne parviennent dans le cerveau, mais uniquement des patterns, pourquoi la vision semble-t-elle si différente de l'audition ? » Il m'est difficile de décrire en quoi la vue diffère de l'ouïe, mais il va de soi que c'est le cas. Je présume qu'il en est de même pour vous. Pourtant, un axone qui représente le son et un autre qui représente la lumière sont, pour toutes sortes de raisons pratiques, identiques. La clarté et le bruit ne sont pas véhiculés sur l'axone d'un neurone sensoriel.

Le cerveau des personnes atteintes de synesthésie — le peintre Wassily Kandinsky fut un cas célèbre — brouille la distinction

entre les sens : certains sons ou certaines textures se traduisent par une couleur. Ce syndrome nous apprend que l'aspect qualitatif d'un sens n'est pas immuable. Suite à certaines modifications physiques, le cerveau peut transmettre un aspect qualitatif de la vision vers un input auditif.

Quelle est donc l'explication des qualia ? Deux réponses me viennent à l'esprit, dont aucune ne me satisfait complètement. Selon l'une, bien que l'ouïe, le toucher et la vue œuvrent de façon semblable dans le néocortex, ils sont gérés différemment sous le cortex. L'ouïe est liée à un ensemble de structures sous-corticales spécifiques à l'audition qui traitent les patterns auditifs avant qu'ils parviennent au cortex. Les patterns somatosensoriels voyagent à travers un ensemble d'aires sous-corticales uniquement consacrées aux sens somatiques. Il se peut qu'à l'instar des émotions les qualia — le pluriel de « quale » — ne soient pas médiatisés que par le cortex. S'ils sont liés de quelque manière avec les parties sous-corticales du cerveau dont le câblage est unique, s'ils sont peut-être liés à des centres émotionnels, cela pourrait expliquer pourquoi nous les percevons différemment, même si cela ne nous explique pas pourquoi une sensation apparentée au quale existe en premier lieu.

L'autre réponse est que la structure des inputs — les différences dans les patterns eux-mêmes — dicte l'expérience que vous éprouvez des aspects qualitatifs de l'information. La nature d'un pattern spatio-temporel dans le nerf auditif est différente de la nature d'un pattern spatio-temporel dans le nerf optique. Le nerf optique est constitué de millions de fibres et transporte des informations spatiales. Le nerf auditif ne compte que trente mille fibres et transporte des informations temporelles. Ces différences peuvent être liées à ce que nous appelons les qualia.

Nous pouvons être certains que, quelle que soit la manière dont la conscience est définie, la mémoire et la prédiction jouent un rôle crucial dans sa création.

Les notions d'esprit et d'âme sont liées à la conscience.

Quand j'étais enfant, je me demandais ce qu'il serait advenu de mon Moi si j'étais né dans le corps d'un autre enfant, dans un autre pays, comme si mon Moi était en quelque sorte indépendant de mon corps. Cette sensation d'un esprit différent du physique est courante ; c'est une conséquence naturelle du fonctionnement du néocortex. Notre néocortex élabore un modèle du monde dans sa mémoire hiérarchique. Les pensées sont ce qui est produit lorsque ce modèle fonctionne par lui-même ; se souvenir conduit à des prédictions, qui agissent comme inputs sensoriels, qui conduisent à de nouveaux appels à la mémoire, et ainsi de suite. Nos pensées les plus contemplatives ne sont pas régies ni même liées au monde réel ; ce sont de pures créations de notre modèle. Nous fermons les yeux et recherchons la quiétude afin que nos pensées ne soient interrompues par aucun input sensoriel. Bien sûr, notre modèle a originellement été créé par l'exposition au monde réel par le truchement de nos sens, mais quand nous planifions et pensons le monde, nous le faisons *via* le modèle cortical, et non par le monde lui-même.

Pour le cortex, notre corps n'est qu'une partie du monde extérieur. Rappelez-vous que le cerveau est une boîte obscure et silencieuse. Il ne connaît le monde que grâce aux patterns présents sur les fibres nerveuses sensorielles. Du point de vue du cerveau considéré comme un mécanisme à patterns, le cerveau ne différencie pas votre corps du reste du monde. Il n'y a pas de distinction spéciale entre la limite du corps et le commencement de l'extérieur. Mais le cortex n'a pas la capacité de modéliser le cerveau lui-même car il n'y a pas de sens à l'intérieur de l'encéphale. Voilà pourquoi nos pensées semblent être indépendantes de notre corps, pourquoi nous avons l'impression que notre esprit et notre âme sont dissociés du corps. Le cortex élabore un modèle de notre corps, mais il ne peut en élaborer un pour le cerveau. Vos pensées, localisées dans le cerveau, sont physiquement séparées du corps et du reste du monde. L'esprit est indépendant du corps, pas du cerveau.

Cette différenciation se révèle nettement lors d'un traumatisme ou d'une maladie. Quand une personne perd un membre, le modèle de ce membre dans le cerveau perdure néanmoins, produisant un effet de « membre fantôme » qui semble appartenir encore au corps. Par ailleurs, un traumatisme cérébral peut entraîner la perte du modèle du bras, bien que ce dernier soit intact. Dans ce cas, le patient peut souffrir du syndrome du « membre étranger ». Il éprouve alors l'inconfortable et parfois insupportable sensation que son bras ne lui appartient pas, qu'il est contrôlé par quelqu'un d'autre, et dans les cas graves, le patient en arrive à exiger l'amputation. Quand le cerveau est intact alors que notre corps est malade, nous avons l'impression qu'un esprit sain est emprisonné dans un corps moribond, alors qu'en fait il s'agit d'un cerveau sain dans un corps moribond. Il est naturel de croire que notre esprit survivra à la destruction du corps, mais quand le cerveau disparaît, l'esprit en fait autant. Cette vérité est évidente lorsque le cerveau se meurt avant le corps. Les patients atteints de la maladie d'Alzheimer, ou ceux dont le cerveau a subi de sévères dommages, perdent leur esprit même si leur corps demeure sain.

## QU'EST L'IMAGINATION ?

Conceptuellement, l'imagination est plutôt simple. Des patterns parviennent dans chaque aire corticale provenant soit des sens, soit de régions inférieures de la hiérarchie corticale. Chaque aire corticale produit des prédictions qui sont renvoyées vers le bas de la hiérarchie. Pour imaginer quelque chose, vous laissez tout bonnement les prédictions faire un demi-tour dans le cortex et devenir des inputs. Sans rien faire physiquement, vous pouvez suivre les conséquences de vos prédictions : « Si ceci se produit, il arrivera cela, puis ça », et ainsi de suite. C'est ainsi que nous procédons lorsque nous préparons une réunion de travail, lorsque nous jouons aux échecs, lorsque nous nous préparons pour une épreuve sportive, et pour mille autres choses.

Aux échecs, vous imaginez le déplacement du cavalier à une certaine position puis vous imaginez la nouvelle configuration des pièces après ce mouvement. Cette image à l'esprit, vous prédisez ce que fera l'adversaire et comment seront disposées les pièces ensuite. Puis vous prédisez le coup que vous jouerez, et ainsi de suite. Vous parcourez diverses étapes et leurs conséquences. Vous prenez enfin votre décision selon la séquence d'événements imaginés, que le déplacement initial de la pièce fût bon ou non. Certains athlètes, comme les descendeurs en ski alpin, améliorent leurs performances en répétant sans cesse la course mentalement. Les yeux fermés, ils se représentent chaque virage, chaque obstacle et même leur montée sur le podium, ce qui augmente leurs chances de réussir. « Imaginer » n'est qu'un autre mot pour « planifier ». C'est là que la capacité prédictive de notre cerveau est profitable. Elle nous permet de connaître les conséquences de nos actions avant de les entreprendre.

Imaginer exige un mécanisme neuronal capable de transformer une prédiction en input. Au Chapitre 6, j'avais conjecturé que c'est dans les cellules de la couche 6 que les prédictions précises se produisent. Les cellules qui s'y trouvent effectuent des projections vers les niveaux inférieurs de la hiérarchie, mais projettent aussi en arrière, vers les cellules d'entrée de la couche 4. Il en résulte que les outputs d'une région peuvent devenir leurs propres inputs. Comme je l'ai mentionné précédemment, Stephen Grossberg, qui a une longue expérience de la modélisation corticale, appelle cela le « feedback replié ». Quand vous fermez les yeux et imaginez un hippopotame, l'aire visuelle de votre cortex devient active, exactement comme si vous regardiez véritablement un hippopotame. Vous voyez ce que vous imaginez.

## QU'EST LA RÉALITÉ ?

Les gens me demandent, avec une inquiétude teintée d'étonnement : « Vous voulez dire que notre cerveau crée un modèle du

monde ? Et que ce modèle du monde pourrait être plus important que la réalité vraie ? »

Je leur réponds alors : « Oui, jusqu'à un certain point, c'est ça. »

« Mais le monde n'existe-t-il pas au-dehors du cerveau ? »

Bien sûr. Les gens sont réels, les arbres aussi, ma chatte est réelle, les situations sociales qui sont les vôtres le sont aussi. Mais votre compréhension du monde et vos réactions sont fondées sur des prédictions issues de votre modèle interne. A tout moment, vous ne pouvez capter directement par vos sens qu'une minuscule partie du monde. Cette partie infime impose les mémorisations qui seront invoquées, mais ce n'est pas suffisant en soi pour échafauder l'ensemble de votre actuelle perception. Par exemple, je suis actuellement en train de taper du texte dans mon bureau, et quelqu'un frappe à la porte d'entrée. Je sais que ma mère vient d'arriver et je l'imagine en bas des escaliers, bien que je ne l'aie en réalité ni vue ni entendue. Rien dans mon input sensoriel n'était spécialement en relation avec ma mère. C'est mon modèle mnémonique du monde qui a prédit qu'elle est là, par analogie avec une expérience passée. La majeure partie de ce que vous percevez ne provient pas de vos sens ; c'est le produit de votre modèle mnémonique interne.

La question « Qu'est la réalité ? » dépend donc largement de la précision avec laquelle votre modèle cortical reflète la vraie nature du monde.

De nombreux aspects du monde qui nous entoure sont si constants que presque tous les humains en possèdent le même modèle interne. Bébé, vous avez appris que la lumière qui éclaire un objet rond projette une ombre d'une certaine forme, et que la forme de la plupart des objets peut être estimée d'après des indices fournis par le monde naturel. Vous avez appris qu'en lâchant une tasse du haut de la chaise de bébé la gravité la fait immanquablement tomber par terre. Vous avez appris à reconnaître des textures, des géométries, des couleurs, ainsi que le rythme des jours

et des nuits. Ces propriétés physiques simples du monde sont invariablement apprises par tous les peuples.

Une grande partie de notre modèle du monde est cependant fondée sur les coutumes, la culture et sur ce que nos parents nous ont appris. Ces parties du modèle sont plus variables et peuvent même être complètement différentes d'un peuple à un autre. Un enfant élevé avec amour par des parents attentifs à ses besoins émotionnels parviendra probablement à l'âge adulte en prédisant que le monde est sûr et agréable. Les enfants maltraités par un parent ou par les deux seront enclins à imaginer un futur dangereux et cruel, dans lequel nul n'est digne de confiance, quelle que soit la manière dont il sera traité par la suite. Une grande part de la psychologie est fondée sur les conséquences des premières années de l'existence, sur la qualité de l'affection et des soins prodigués, car c'est à ce moment que le cerveau construit son modèle du monde.

Notre culture aussi participe à la formation du modèle du monde. Par exemple, des études ont démontré que l'espace et les objets sont perçus différemment par les Occidentaux et par les Asiatiques. Les Asiatiques sont plus attentifs à l'espace qui sépare les objets tandis que les Occidentaux s'attachent davantage aux objets eux-mêmes, une différence qui se manifeste au niveau de l'esthétique et de la résolution des problèmes. Des recherches ont montré que dans des cultures fondées sur l'honneur — celles de tribus afghanes ou de communautés du sud des Etats-Unis —, la violence est plus facilement acceptée comme naturelle. Différentes croyances religieuses inculquées très tôt dans la vie peuvent conduire à des modèles moraux complètement différents, y compris en ce qui concerne le traitement réservé aux hommes et aux femmes, voire la valeur de la vie elle-même. Il va de soi que ces divers modèles du monde ne sauraient être universellement acceptés comme légitimes, même s'ils peuvent le paraître pour tel ou tel individu. Le raisonnement moral, tant le bien que le mal, est appris.

Notre culture ainsi que l'expérience accumulée par la famille nous enseignent des stéréotypes, qui sont hélas un aspect inévita-

ble de la vie. Tout au long de ce livre, vous pourriez remplacer le mot *stéréotype* par *mémoire invariante* ou par *représentation invariante* sans altérer substantiellement le propos. La prédiction par analogie s'apparente beaucoup à un jugement stéréotypé. Or, les stéréotypes négatifs entraînent de terribles conséquences sociales. Si ma théorie de l'intelligence est exacte, nous ne pourrons pas éradiquer la propension des gens à penser par stéréotypes, car c'est par stéréotypes que le cerveau fonctionne. C'est une fonction qui lui est inhérente.

La seule manière de circonscrire les méfaits des stéréotypes est d'enseigner à nos enfants comment reconnaître les faux stéréotypes, développer l'empathie et le scepticisme. Il nous appartient de promouvoir l'aptitude à la pensée critique, qui s'ajoutera à l'enseignement des valeurs les plus élevées. Le scepticisme, qui est au cœur de toute méthodologie scientifique, est la seule manière de faire la part des faits et de la fiction.

---

J'espère être parvenu à vous convaincre que l'intelligence n'est qu'un produit du cerveau. Ce n'est pas un élément manipulable ou qui coexiste avec les cellules cérébrales. Les neurones ne sont que des cellules. Aucune force mystique ne fait se comporter une cellule nerveuse ou un ensemble de cellules nerveuses autrement qu'elles doivent naturellement le faire. Ce fait établi, nous pouvons à présent porter notre attention sur la manière dont les capacités des cellules du cerveau peuvent être implémentées afin qu'elles se souviennent et prédisent, bref, développer un algorithme cortical qui sera gravé dans le silicium.

# 8

# L'AVENIR DE L'INTELLIGENCE

**Il est difficile** de prédire l'usage ultime d'une nouvelle technologie. Comme nous l'avons découvert tout au long de ce livre, le cerveau procède à des prédictions par analogie avec le passé. Notre inclination naturelle est donc d'imaginer qu'une nouvelle technologie sera utilisée pour effectuer les mêmes tâches que celle qui l'a précédée. Nous pensons que nous utiliserons un nouvel outil pour faire la même chose, mais plus vite, plus efficacement ou à moindre coût.

Les exemples abondent. Les Indiens d'Amérique avaient appelé le train « le cheval de fer » et l'automobile « le chariot sans chevaux ». Pendant des décennies, le téléphone fut considéré que comme une variante du télégraphe qui ne devait être utilisé que pour transmettre des nouvelles importantes ou en cas d'urgence. Ce n'est que dans les années 1920 que les gens commencèrent à s'en servir pour converser. La photographie fut d'abord une nouvelle forme de portrait. Quant au cinéma, il était considéré comme une variante du théâtre; c'est pourquoi, pendant la majeure partie du XX$^{e}$ siècle, un rideau fut tiré sur l'écran.

Pourtant, l'usage ultime de la plupart des nouvelles technologies est souvent imprévu et beaucoup plus éloigné que ce que notre imagination avait d'abord envisagé. Le téléphone a évolué en un réseau sans fil permettant à quiconque de communiquer avec autrui par la voix, le texte ou des images jusqu'à l'autre bout du monde. Le transistor a été inventé en 1947 par les Bell Labs. Il fut évident pour tous que c'était un progrès décisif dans l'électronique, mais son premier usage ne fut qu'une amélioration de ce qui existait déjà : le transistor remplaça les tubes à vide. Ceci permit de fabriquer des postes de radio et des ordinateurs plus petits et plus fiables, ce qui était important et séduisant à cette époque, mais la grande différence se limitait à la taille et à la fiabilité du matériel. Les applications les plus révolutionnaires du transistor ne furent découvertes que plus tard. Une période d'innovations progressives fut nécessaire avant que quelqu'un conçoive le circuit intégré, le microprocesseur, le processeur de signal numérique et le composant de mémoire. Le microprocesseur fut développé dès 1970 avec en vue l'idée du calculateur de bureau. Là encore, les premières applications ne firent que se substituer à des technologies existantes. Le calculateur électronique devait remplacer les machines à calculer mécaniques. Les microprocesseurs étaient clairement destinés à succéder aux solénoïdes utilisés par certains appareils de contrôle, comme les commutateurs des feux de signalisation. C'était des années avant que la véritable puissance du microprocesseur puisse se manifester. Personne à l'époque n'aurait pu prévoir l'ordinateur personnel moderne, le téléphone mobile, l'Internet, le GPS et toutes les autres technologies de l'information aujourd'hui très banales.

Il serait donc vain de croire que nous pourrions prédire dès à présent les applications révolutionnaires de systèmes de mémoire basés sur le cerveau. Je m'attends à ce que ces machines intelligentes améliorent l'existence à bien des égards. C'est certain. Mais prédire l'avenir d'une technologie à plus de quelques années est illusoire. Vous n'en aurez pour preuve que la lecture des prévi-

sions absurdes des futurologues. Dans les années 1950, il était dit qu'en l'an 2000 des réacteurs atomiques seraient installés dans le sous-sol de nos maisons et que nous passerions nos vacances sur la Lune. Mais du moment que nous gardons ces élucubrations à l'esprit, sachant que nous risquons de les répéter, nous pouvons nous aventurer à spéculer sur ce que seront les machines intelligentes. Nous en tirerons quelques grossières mais utiles conclusions sur ce que nous réserve le futur.

Voici quelques questions intrigantes : peut-on fabriquer des machines intelligentes et si oui, de quoi auront-elles l'air ? Ressembleront-elles plutôt aux robots humanoïdes des films de science-fiction, au boîtier noir, beige ou coloré d'un micro-ordinateur ou à quelque chose d'autre ? Comment nous en serviront-nous ? Est-ce une technologie dangereuse qui peut nous exposer à des risques ou menacer nos libertés individuelles ? Quelles sont les applications évidentes pour des machines intelligentes, et y a-t-il un moyen de savoir ce que seront ces fantastiques applications ? Quel sera en définitive l'impact de ces machines intelligentes sur notre façon de vivre ?

## POUVONS-NOUS CONSTRUIRE DES MACHINES INTELLIGENTES ?

Oui, nous pouvons construire des machines intelligentes, mais elles ne seront peut-être pas ce que vous imaginez. Bien qu'adaptées à leur fonction, je ne pense pas que nous construirons des machines intelligentes agissant comme des êtres humains, ni même qui interagiraient avec nous à la manière d'humains.

L'une des idées que nous nous faisons des machines intelligentes nous vient des films et des livres de science-fiction : elles sont attachantes ou malveillantes, ce sont parfois des robots qui conversent avec nous de nos sentiments, de nos idées ou des événements, ou qui fomentent d'interminables complots. Un siècle de science-fiction a formé les gens à considérer que robots et androïdes sont une part inévitable et désirable de notre avenir.

Des générations entières ont grandi avec les images de Robbie, le robot du film *Planète interdite*, les robots R2D2 et C3PO de *La Guerre des étoiles*, le lieutenant-commandant Data de *Star Trek*. Même le robot HAL, dans *2001 : l'odyssée de l'espace*, bien que dépourvu de corps, était très proche d'un humain, conçu pour être autant un compagnon qu'un copilote programmé guidant les humains dans leur long voyage interplanétaire. Les applications robotiques limitées, comme les voitures intelligentes, les mini sous-marins autonomes pour les plongées profondes, ainsi que les aspirateurs et tondeuses à gazon autoguidés, sont faisables et se banaliseront. Mais les androïdes et les robots comme le commandant Data et C3PO resteront pour longtemps encore des personnages de fiction, et cela pour deux bonnes raisons.

D'abord, l'intelligence n'est pas créée que dans le néocortex, mais aussi par le système émotionnel du cerveau archaïque et par la complexité du corps humain. Pour être un humain, il faut toute la machinerie biologique, et pas uniquement un cortex. Converser comme un être humain sur n'importe quel sujet — et satisfaire ainsi au test de Turing — exigerait d'une machine intelligente qu'elle jouisse d'une grande partie de l'expérience et des émotions d'un être humain, et que sa vie s'apparente à celle d'un humain. Les machines intelligentes auront l'équivalent d'un cortex et un ensemble de sens, mais le restant sera facultatif. Il pourrait être divertissant de regarder une machine intelligente traîner dans les parages son corps anthropomorphe, mais elle n'aurait pas du tout un esprit d'humain, à moins que nous ne la dotions d'un système émotionnel fondé sur celui d'un humain et d'une expérience elle aussi fondée sur celle d'un humain. Ce serait extrêmement difficile à faire et, à mon avis, parfaitement futile.

Deuxièmement, étant donné le coût et le travail qui seraient nécessaires pour construire et entretenir des robots humanoïdes, il serait difficile d'en justifier l'intérêt pratique. Un robot majordome serait plus onéreux et moins efficace que son homologue humain. Bien que le robot puisse être « intelligent », il n'offrirait

pas la qualité du relationnel et la compréhension aisée dont fait preuve un employé humain, qui les a d'office du seul fait d'appartenir au genre humain.

La machine à vapeur et le calculateur électronique furent tous deux pressentis pour devenir des robots, mais rien ne fut réalisé. Quand nous évoquons la création de machines intelligentes, beaucoup de gens trouvent naturel d'imaginer aujourd'hui encore des robots anthropomorphes, et là encore, il est peu probable que cela se passe ainsi. Le robot est un concept né de la révolution industrielle et affiné par la fiction. Nous ne devons pas nous en inspirer pour développer des machines authentiquement intelligentes.

A quoi ressembleront alors les machines intelligentes si elles ne sont pas des robots marchant et parlant ? L'évolution a découvert qu'en affublant nos sens d'un système de mémoire hiérarchisée, la mémoire modéliserait le monde et serait capable de prédire. En empruntant ce concept à la nature, nous devrions pouvoir construire des machines intelligentes. Commençons avec un assortiment de sens destinés à extraire des patterns du monde. Notre machine posséderait sans doute un ensemble sensoriel qui différerait de celui de l'être humain, et pourrait même s'exercer dans un monde différent du nôtre (nous y reviendrons d'ici peu). C'est pourquoi, n'imaginez pas qu'elle aura deux yeux et deux oreilles. Liez ensuite à ces sens un système de mémoire hiérarchisée fonctionnant sur le même principe que le cortex. Il faudra ensuite éduquer ce système mnémonique un peu comme nous enseignons aux enfants. Par des apprentissages répétés, notre machine intelligente élaborera peu à peu un modèle de *son* monde, tel qu'il est perçu pas *ses* sens. Il n'y aura nul besoin de lui programmer des règles, des bases de connaissances, des faits et tout autre concept de haut niveau qui sont le fléau de l'intelligence artificielle. La machine intelligente doit apprendre par l'observation de son environnement, avec au besoin des inputs provenant d'un instructeur. Après avoir créé un modèle de son monde, notre machine intelligente pourra ensuite découvrir des

analogies d'après ses expériences passées, prédire des événements, proposer des solutions à des problèmes tout nouveaux et mettre ses connaissances à notre disposition.

En pratique, notre machine intelligente pourra être embarquée à l'intérieur d'avions ou de voitures, ou être prosaïquement enclose dans un rack situé dans un local informatique. Contrairement aux êtres humains, dont le cerveau doit accompagner le corps, le système mnémonique d'une machine intelligente peut être stocké loin des capteurs (ou de son « corps », si elle en avait un). Par exemple, les capteurs d'un système de sécurité intelligent pourraient être dispersés dans toute l'usine ou dans toute une ville, mais le système de mémoire hiérarchisée serait installé dans le soubassement d'un immeuble : l'incarnation physique d'une machine intelligente pourra prendre beaucoup de formes.

Il n'y a aucune raison pour qu'une machine intelligente doive ressembler à un humain, agir ou percevoir comme lui. Ce qui la rend intelligente est son aptitude à comprendre le monde et interagir avec lui au travers d'un modèle mnémonique hiérarchisé, et sa capacité à appréhender son monde comme nous le faisons pour le nôtre. Comme nous le verrons, ses pensées et ses actions peuvent être complètement différentes de ce que fait un humain, ce qui ne la rend pas moins intelligente. L'intelligence se mesure à la capacité de prédiction que manifeste une mémoire hiérarchique, et non par un comportement anthropomorphe.

---

Portons notre attention au plus grand défi technologique auquel nous serons confrontés lors de la construction de machines intelligentes : la création de la mémoire. Nous devrons fabriquer de vastes systèmes mnémoniques hiérarchiquement organisés, fonctionnant comme le cortex. Ce défi se posera au niveau de la capacité et de la connectivité.

La *capacité* est le premier problème. Admettons que le cortex possède trente-deux trillions de synapses. Si nous ne représen-

tons chacune d'elles que sur deux bits (ce qui permet d'attribuer une parmi quatre valeurs à une synapse) et si chaque octet compte deux bits (un octet représente alors quatre synapses), la mémoire sera d'environ huit trillions d'octets, soit huit téraoctets. Sachant que la capacité du disque dur d'un micro-ordinateur moyen est d'environ cent milliards d'octets, il faudrait chaîner environ quatre-vingts disques durs pour héberger la même quantité de mémoire qu'un cortex humain (ne vous attachez pas aux chiffres exacts, car tout n'est ici que grossière estimation). Bref, la quantité de mémoire est assurément disponible. Nous ne sommes pas à un facteur de mille près, mais ce ne sera pas non plus le type de machine que vous pourriez glisser dans votre poche ou intégrer à un grille-pain. L'important est que la quantité de mémoire requise n'est pas un problème, ce qui n'était pas le cas voici dix ans (des disques durs d'un téraoctet sont d'ores et déjà commercialisés à un prix raisonnable). De plus, ce qui nous arrange est le fait que nous n'avons pas à recréer le cortex humain tout entier. Il en faut bien moins pour la plupart des applications.

Notre machine intelligente aura besoin de beaucoup de mémoire. Nous commencerons sans doute sa fabrication en recourant à des disques durs ou à des disques optonumériques, mais finalement, nous utiliserons aussi des composants de mémoire en silicium, comme les mémoires flash. Elles sont bon marché, consomment peu et sont solides. Disposer de composants dont la capacité est suffisante pour les machines intelligentes n'est qu'une question de temps. En fait, une mémoire intelligente offre un avantage sur la mémoire conventionnelle. Toute l'économie du secteur des semi-conducteurs est fondée sur le nombre de puces présentant des erreurs. Pour la plupart, le moindre défaut les rend inutilisables. Le taux de réussite de bonnes puces est appelé « yeld » ; il sert à déterminer si la conception d'un composant permet de le fabriquer et de le vendre en dégageant un profit. Les risques d'erreurs étant proportionnels aux

dimensions de la puce, la taille de la plupart d'entre elles n'excède pas celle d'un petit timbre-poste. La quantité de mémoire a été augmentée, non en fabriquant des puces plus grandes, mais en miniaturisant davantage ses composants élémentaires.

Les composants de mémoire intelligents seront intrinsèquement tolérants aux défauts. Rappelez-vous que dans le cerveau aucun composant ne contient de données indispensables. Des milliers de neurones disparaissent tous les jours et pourtant, la capacité mentale ne s'atténue que très lentement au cours de l'existence. Les composants de mémoire intelligents fonctionneront sur le même principe que le cortex ; ainsi, même si une partie des éléments de la mémoire se révèle défectueuse, le composant sera encore utilisable et commercialement viable. Plus probablement, la tolérance d'erreur d'une mémoire ressemblant à celle du cerveau permettra aux concepteurs de développer des composants significativement plus gros et plus denses que ceux d'aujourd'hui. Nous pourrions ainsi être en mesure de graver un cerveau dans le silicium bien plus rapidement que nous le pensons.

Le second écueil est la *connectivité*. Le cerveau réel comprend une grande quantité de matière grise sous-corticale. Comme nous l'avons signalé précédemment, elle est faite de millions d'axones qui s'étendent sous la mince enveloppe corticale, reliant entre eux les différents niveaux de la hiérarchie corticale. Une seule cellule du cortex peut ainsi être reliée à cinq ou dix mille autres cellules. Ce type de câblage massivement parallèle est difficile voire impossible à implémenter à l'aide des techniques traditionnelles de fabrication de composants en silicium. Les puces sont confectionnées en déposant quelques couches de métal alternant avec des couches d'isolant (ce procédé n'a rien de commun avec les couches du cortex). Les couches de métal contiennent les futurs câbles de la puce, et du fait que ces câbles ne peuvent se croiser sur une même couche, le nombre total de connexions câblées est limité. Cette configuration n'est pas exploitable pour en faire des systèmes mnémoniques inspirés du

cerveau, où des millions de connexions sont indispensables. La puce en silicium et la matière grise ne sont pas compatibles.

Beaucoup d'ingénierie et d'expérimentations seront nécessaires pour résoudre ce problème, mais nous en connaissons les bases et nous savons comment nous y prendre. Les câbles électriques acheminent les signaux beaucoup plus rapidement que les axones des neurones. Un seul câble d'une puce peut être partagé, et de ce fait utilisé pour plusieurs connexions, alors que dans le cerveau, chaque axone n'appartient qu'à un seul neurone.

Le réseau téléphonique est un bel exemple. Si nous devions tirer une ligne de chacun des téléphones de la Terre à tous les autres, la surface de la planète disparaîtrait sous un écheveau de fils de cuivre. En réalité, tous les téléphones se partagent un nombre relativement modeste de lignes à haute capacité. Cette technique fonctionne tant que la capacité de chaque ligne est nettement supérieure à celle requise pour transmettre une seule conversation. Le réseau téléphonique répond à cette exigence : un seul câble à fibres optiques est capable de véhiculer un million de conversations à la fois.

Dans le cerveau, des axones relient toutes les cellules qui échangent des informations entre elles, mais nous pouvons construire des machines intelligentes inspirées plutôt du réseau téléphonique, avec un partage des connexions. Des scientifiques ont pensé depuis des années résoudre le problème de la connectivité dans une puce simulant le cerveau. Bien que ce qui se passe dans le cortex demeure encore un mystère, les chercheurs savent que nous finirons un jour par résoudre le puzzle, et que nous serons ensuite confrontés au problème de la connectivité. Il est ici inutile de passer en revue les différentes approches. Il nous suffit de savoir que la connectivité sera peut-être l'obstacle technique majeur qui nous attend, mais nous devrions en venir à bout.

Les défis techniques relevés, plus aucun problème de fond ne nous empêchera de construire des machines authentiquement intelligentes. Bien sûr, une foule d'autres problèmes annexes

devront être réglés pour que ces systèmes soient compacts, bon marché, économes en énergie, mais rien de tout cela n'est un obstacle. Il nous a fallu cinquante ans pour passer d'un ordinateur de trente tonnes occupant 167 $m^2$ — l'ENIAC — à l'ordinateur de poche. Mais comme nous partons d'une position technologiquement plus élevée, la transition vers les machines intelligentes devrait être beaucoup plus rapide.

## DEVONS-NOUS CONSTRUIRE DES MACHINES INTELLIGENTES ?

Tout au long du XXI$^{e}$ siècle, les machines intelligentes devraient passer du domaine de la science-fiction à celui de la réalité. Mais auparavant, nous devrons aborder quelques questions éthiques, notamment nous demander si un possible danger pourrait l'emporter sur les avantages.

La perspective de créer des machines capables de penser et d'agir par elles-mêmes est depuis longtemps une source d'inquiétude. C'est compréhensible. A leur apparition, les connaissances et technologies nouvelles intimident toujours. Notre créativité nous laisse imaginer les terribles moyens par lesquels de nouvelles technologies pourraient permettre de prendre le contrôle de notre corps, de nous rendre inutiles, voire anéantir la valeur de la vie humaine. Mais l'Histoire révèle que ces sombres pressentiments ne surviennent jamais comme prévu. Au début de la révolution industrielle, les gens ont craint la machine à vapeur puis l'électricité. Des machineries alimentées par leur propre énergie, capables de se mouvoir de manière complexe, paraissaient à la fois miraculeuses et quelque peu inquiétantes. Or, de nos jours, l'électricité et la combustion interne ne sont plus ni étranges ni inquiétantes. Elles font partie de notre environnement comme l'air et l'eau.

Quand la révolution de l'information commença, elle suscita la crainte de l'informatique. D'innombrables histoires de science-fiction décrivaient de puissants cerveaux électroniques ou des

réseaux d'ordinateurs prenant spontanément conscience d'eux-mêmes et qui se révoltaient contre leurs maîtres organiques. Maintenant que les ordinateurs se sont intégrés à notre quotidien, ces craintes nous paraissent absurdes. L'ordinateur installé chez vous ou l'Internet n'ont pas plus de chances d'être un jour doués de sentiments et de sensations qu'un tiroir-caisse.

Toute technologie peut bien sûr être exploitée en bien ou en mal, et certaines se prêtent plus que d'autres à être dévoyées ou calamiteuses. L'énergie atomique est dangereuse, qu'elle soit utilisée dans les ogives nucléaires ou par les centrales électriques, car un simple accident ou une erreur peuvent entraîner la mort ou contaminer des millions de personnes. Bien que l'énergie nucléaire soit précieuse, il existe d'autres solutions. Les techniques de fabrication des véhicules peuvent produire des chars d'assaut ou des avions de chasse, mais aussi des voitures particulières, des camions et des avions civils, dont l'utilisation maladroite ou malencontreuse peut également tuer ou blesser. Mais les véhicules sont indubitablement plus nécessaires à la vie moderne et moins dangereux que l'énergie nucléaire. Les dommages causés par une erreur de pilotage d'un avion de ligne sont assurément moindres que ceux causés par le largage d'une bombe atomique. Beaucoup de technologies n'offrent d'ailleurs que des bénéfices. Le téléphone en est un exemple. Sa capacité à rapprocher les gens compense largement ses effets négatifs. C'est la même chose pour l'électricité et les sciences de la santé. Je pense que les machines intelligentes feront partie des technologies les plus inoffensives, les plus bénéfiques que nous ayons jamais développées.

Des personnes comme Bill Joy, cofondateur de Sun Microsystems, craignent pourtant que nous parvenions à développer des robots intelligents qui pourraient échapper à notre contrôle, essaimer sur la Terre entière et la remodeler à leur convenance. Cette image me rappelle les balais de *L'Apprenti sorcier* dont les débris se régénèrent d'eux-mêmes, et qui travaillent avec zèle

jusqu'au désastre final. Dans la même veine, quelques optimistes chercheurs en intelligence artificielle profèrent quelques prophéties pour la génération en cours qui sont pour le moins troublantes. Par exemple, Ray Kurzweil imagine des nanorobots parcourant notre cerveau, enregistrant chaque synapse et chaque connexion, et transmettant toutes les informations à un superordinateur qui reconfigurerait l'être humain directement de l'intérieur. Il en extrairait une version logicielle de vous-même qui serait pratiquement immortelle. Ces deux prédictions au sujet des machines intelligentes, le scénario de la machine qui s'emballe et devient folle et celui du téléchargement de cerveau à ordinateur, semblent refaire sans cesse surface.

La construction de machines intelligentes n'est pas la même chose que la construction de machines autoréplicantes. Il n'y a de toute façon pas de lien logique entre ces deux concepts. Ni le cerveau ni l'ordinateur ne peuvent s'autorépliquer, et les systèmes mnémoniques fondés sur le cerveau n'y changeront rien. Bien que l'un des points forts des machines intelligentes soit notre capacité à les produire en masse, elles ne viseront pas à se répliquer à la manière des bactéries et des virus. L'autoréplication n'exige pas de l'intelligence, et l'intelligence n'exige pas l'autoréplication.

De plus, je doute sérieusement que nous puissions un jour dupliquer notre intelligence dans des machines. Il n'existe actuellement, pour autant que je sache, aucune technique réelle ou imaginaire permettant d'enregistrer les trillions de détails qui forment votre « vous ». Il faudrait pour cela enregistrer et recréer la totalité de votre système nerveux ainsi que votre corps, et pas uniquement le néocortex. Et il nous faudrait comprendre en plus comment tout cela fonctionne. Un jour certainement, nous arriverons à le faire, mais le défi posé s'étend bien au-delà de la compréhension du fonctionnement de notre cortex. Déterminer l'algorithme néocortical et l'inscrire dans des machines à partir de zéro est une chose, mais analyser les milliards de milliards de

détails opérationnels d'un cerveau vivant et les répliquer dans une machine en est une autre, complètement différente.

---

Outre l'autoréplication et la duplication de l'intelligence, les gens se posent d'autres questions au sujet des machines intelligentes. Pourraient-elles menacer de grandes parties de la population comme le fait l'arme nucléaire ? Leur avènement pourrait-il servir les desseins de petits groupes d'individus malintentionnés ? Ou encore, les machines pourraient-elles devenir malfaisantes et se retourner contre nous, comme les implacables personnages des films *Terminator* et *Matrix ?*

La réponse à ces questions est non. En tant que périphériques d'information, des systèmes mnémoniques inspirés du cerveau feront partie des technologies les plus utiles que nous ayons développées jusqu'à ce jour. A l'instar des automobiles et des ordinateurs, ce ne seront que des outils. Le seul fait d'être intelligents ne signifie pas qu'ils auront des aptitudes spéciales pour détruire le bien d'autrui ou manipuler les gens. Et de même que nous ne voudrions pas confier l'arsenal nucléaire mondial au bon vouloir d'une seule personne ou d'un seul ordinateur, nous devons nous garder de trop nous reposer sur les machines intelligentes car, comme toutes les technologies, elles sont faillibles.

Ceci nous amène à la question de la malveillance. Pour certaines personnes, être intelligent est fondamentalement la même chose qu'avoir une mentalité humaine. Ils craignent que les machines intelligentes se sentent mises en esclavage, parce qu'eux-mêmes, humains, ne le supportent pas. Ils craignent que les machines intelligentes tentent de contrôler le monde car, au cours de l'Histoire, les gens intelligents ont toujours tenté d'étendre leur hégémonie. Ces craintes restent toutefois de fausses analogies. Elles résultent de l'amalgame de l'intelligence — l'algorithme cortical — et des commandes émotionnelles issues du cerveau archaïque — les réactions instinctives. Des machines

intelligentes n'auront toutefois pas toutes ces facultés. Elles n'auront pas d'ambitions personnelles. Elles ne désireront pas la santé, la reconnaissance sociale ou des gratifications sensuelles. Elles n'éprouveront ni appétit, ni toxicodépendance, ni troubles de l'humeur. Les machines intelligentes ne manifesteront rien ressemblant à des émotions humaines, à moins que nous ne les concevions minutieusement à cette fin. Les plus puissantes applications des machines intelligentes concerneront des domaines dans lesquels l'intellect humain se heurte à des difficultés, ceux dans lesquels nos sens sont inappropriés, ou les activités qui sont trop ennuyeuses. En général, le contenu émotionnel de ces activités est négligeable.

Les machines intelligentes iront de systèmes simples, pour des applications très ciblées, jusqu'à de puissants systèmes intelligents supra-humains mais, à moins que nous ne fassions fausse route et choisissions de les faire ressembler à des humains, nous éviterons l'anthropomorphisme. Peut-être devrons-nous un jour imposer des restrictions à ce que les gens pourront faire de leurs machines intelligentes, mais ce n'est pas demain la veille. Et le moment venu, les questions éthiques paraîtront relativement faciles à résoudre comparées à celles qui, aujourd'hui, concernent les biotechnologies génétiques et le nucléaire.

## POURQUOI CONSTRUIRE DES MACHINES INTELLIGENTES ?

La vraie question est : que feront les machines intelligentes ?
On m'a souvent demandé de faire des exposés sur l'avenir de l'informatique mobile. Un organisateur de conférence me demande à ces occasions de dire à quoi ressembleront, dans cinq ou vingt ans, les ordinateurs de poche ou les téléphones mobiles. Ils veulent ma vision du futur. Je ne puis la leur donner. Un jour, je me mis à arpenter l'estrade coiffé d'un chapeau de magicien, une boule de cristal à la main. J'expliquai que nul ne peut voir l'avenir avec précision. La boule de cristal est une illusion et

quiconque prétend savoir exactement ce qui se produira dans les années à venir est certain de se tromper. Le mieux que nous puissions faire est de comprendre les grandes tendances. Quand vous avez saisi les grandes lignes d'une idée, vous pouvez les suivre quelle que soit la direction qu'elles prennent, au fur et à mesure que les détails se révèlent.

Le plus célèbre exemple de tendance technologique avérée est la loi de Moore. Cofondateur de la société Intel, Gordon Moore avait prédit avec justesse que le nombre d'éléments électroniques susceptibles d'être intégrés sur une puce de silicium doublerait tous les dix-huit mois. Moore n'avait pas précisé si cette puce serait un composant de mémoire, un microprocesseur ou quoi que ce soit d'autre. Il n'avait pas précisé l'usage que l'on en ferait, ni si la puce serait enfermée dans un boîtier en plastique, en céramique ou collée à même un circuit imprimé. Il n'avait rien dit des divers procédés de fabrication. Il s'était tenu à la tendance la plus générale et avait vu juste.

A ce jour, nous ne pouvons prédire l'usage ultime des machines intelligentes. Il n'existe en effet aucun moyen d'obtenir des détails fiables. Si moi ou n'importe qui d'autre nous risquions à prédire en détail ce que ces machines feront, l'avenir prouverait inévitablement que nous avions tort. Mais nous pouvons toutefois faire mieux que hausser les épaules. Deux voies peuvent nous faire avancer. L'une consiste à envisager l'utilisation à très court terme des systèmes mnémoniques inspirés du cerveau ; c'est l'approche évidente, certes moins intéressante, à explorer en premier. La seconde approche consiste à réfléchir aux tendances à long terme qui, à l'instar de la loi de Moore, peuvent nous aider à imaginer les applications qui feront peut-être partie de notre futur.

Commençons par les applications à court terme. Ce sont celles qui sont évidentes, comme lorsque le transistor remplaça les tubes radio et le microprocesseur l'étage entier réservé à un ordinateur à lampes. Nous pouvons par-dessus le marché jeter un coup d'œil à certains domaines que l'intelligence artificielle avait

tenté d'aborder, mais sans trouver de solutions : la reconnaissance vocale, la vision et les automobiles intelligentes.

---

Si vous avez déjà eu l'occasion d'utiliser un logiciel de reconnaissance vocale pour entrer du texte dans un micro-ordinateur, vous vous êtes sans doute rendu compte combien il peut être bête. Comme la Chambre chinoise de John Searle, l'ordinateur n'a aucune cognition de ce qui est dit. Les quelquefois où j'ai essayé ce genre de produit, j'ai senti l'énervement monter en moi. Le moindre bruit dans la chambre, que ce soit un crayon qui tombe ou quelqu'un qui parle, faisait apparaître des mots supplémentaires à l'écran. Le taux d'erreur de la reconnaissance était très élevé. Souvent, les mots que le logiciel pensait avoir compris étaient dépourvus de sens : « Ce rat pelé de dira Marie que nous dit non dehors. » Un enfant se rendrait compte que c'est n'importe quoi, mais pas l'ordinateur. De même, ce que l'on appelle des interfaces à langage naturel a pendant des années été un sujet de recherches très prisé des informaticiens. Il s'agissait de pouvoir dire à l'ordinateur ou à d'autres équipements ce qu'ils doivent faire, en langage courant, et laisser la machine exécuter le travail. A un assistant numérique personnel, ou PDA, vous auriez pu demander de vive voix : « Déplace le match de basket de ma fille au dimanche à dix heures. » Ce genre d'ordre est impossible à traiter par l'intelligence artificielle traditionnelle. Même si l'ordinateur pouvait reconnaître chaque mot, pour exécuter la tâche, encore faudrait-il qu'il sache où votre fille va à l'école, que vous faites sans doute allusion au prochain dimanche, et même ce qu'est un match de basket, car le rendez-vous peut avoir été tout simplement libellé « Antony contre Meudon ». Ou alors, vous voudriez que l'ordinateur écoute et enregistre le flux audio d'une radio dès que le nom de tel ou tel morceau est prononcé, mais il se trouve que le présentateur se contente de décrire le prochain morceau sans mentionner le titre. Vous, et vous seul, pouvez deviner ce qui va suivre, mais pas un ordinateur.

Ces exemples et beaucoup d'autres applications exigent que la machine puisse comprendre le langage parlé. L'ordinateur ne peut exécuter ces tâches car il ne comprend pas ce qui est dit. Il fait correspondre des patterns audio à des modèles de mots mémorisés sans savoir ce que ces mots signifient. Imaginez que vous apprenez à reconnaître le son de chaque mot d'une langue étrangère, mais sans savoir ce que ces mots signifient, et que je vous demande ensuite de transcrire une conversation dans cette langue. La conversation se déroulant, vous n'auriez aucune idée de quoi il est question, mais vous essayeriez de saisir chaque mot séparément. Toutefois, comme les mots se chevauchent, interfèrent ou sont « mangés », et que des syllabes se perdent dans le bruit ambiant, vous éprouveriez beaucoup de difficultés à distinguer un mot d'un autre. C'est à ces obstacles que les logiciels de reconnaissance vocale sont aujourd'hui confrontés. Les ingénieurs ont découvert qu'il est possible d'améliorer quelque peu la précision du logiciel en recourant aux probabilités pour déceler les transitions d'un mot à un autre. Ils exploitent aussi les règles de grammaire pour décider entre deux homonymes ou homophones, comme « le maire est amer car sa mère est à la mer ». C'est une forme de prédiction très simple, mais les systèmes sont toujours aussi sots. Aujourd'hui, les logiciels de reconnaissance de la voix ne fonctionnent que dans des situations extrêmement contraintes, dans lesquelles le nombre de mots susceptibles d'être prononcés en un temps donné est très limité. Pourtant, les humains exécutent facilement un grand nombre de tâches reposant sur le langage, car le cortex comprend non seulement des mots, mais aussi des phrases et le contexte dans lequel elles sont dites. Nous anticipons les idées, les phrases et chaque mot. Notre modèle cortical du monde fait tout cela automatiquement.

Nous pouvons donc espérer que des systèmes mnémoniques inspirés du cortex transformeront la faillible reconnaissance vocale en compréhension vocale fiable. Au lieu de programmer les probabilités de transitions de mots à mots, une mémoire

hiérarchisée tiendra compte de l'accent, des mots, des phrases et des idées, et s'en servira pour interpréter ce qui est dit. A l'instar d'une personne, une telle machine intelligente pourra faire la distinction entre différents événements vocaux : par exemple une discussion entre vous et un ami dans une chambre, une conversation téléphonique, des commandes d'édition pour des tâches bureautiques… Il ne sera pas facile d'élaborer ces machines. Pour comprendre pleinement le langage humain, une machine devra acquérir de l'expérience et apprendre ce que font les humains. Par conséquent, même si des années s'écouleront avant que nous puissions construire une machine intelligente qui comprendrait le langage comme vous et moi, il nous sera possible, à plus court terme, d'améliorer les performances des systèmes de reconnaissance vocale grâce à des mémoires pseudo-corticales.

La vision est un autre domaine d'applications où l'intelligence artificielle n'a pas réussi, mais que des systèmes véritablement intelligents devraient être capables d'exploiter. Il n'existe aujourd'hui aucune machine capable de regarder une scène réelle — le paysage qui s'étend devant vous, ou une séquence filmée — et de décrire ce qu'elle voit. Il existe quelques applications qui ont fait leurs preuves dans des domaines très précis, comme aligner des puces sur un circuit imprimé ou comparer les traits d'un visage avec ceux stockés dans une base de données, mais il est actuellement impossible à un ordinateur d'identifier une variété d'objets ou d'analyser une scène de manière générale. Il ne vous est pas du tout difficile de regarder autour de vous, dans une pièce, pour trouver une chaise vide, mais ne demandez pas à un ordinateur de le faire. Observez l'image transmise par une caméra de sécurité : pouvez-vous faire la différence entre quelqu'un qui frappe à la porte en tenant un bouquet de fleurs et quelqu'un qui s'attaque à la porte avec une pince-monseigneur ? Bien sûr que oui. Mais cette distinction est loin encore des capacités des actuels logiciels. C'est pourquoi des agents de sécurité sont chargés de surveiller les écrans vingt-quatre heures sur

vingt-quatre et de repérer tout ce qui est suspect. Il est cependant difficile pour un humain de conserver une vigilance maximale tandis qu'une machine intelligente pourrait inlassablement s'acquitter de cette tâche.

Abordons à présent les moyens de transport. Les voitures sont devenues sophistiquées. Certaines sont équipées d'un système GPS (*Global Positioning System*) qui calcule la route d'un endroit à un autre, de capteurs qui allument les phares en entrant dans un tunnel, d'accéléromètres qui déploient les airbags et de capteurs de proximité qui signalent un obstacle en se garant. Des prototypes de véhicules capables de se déplacer de façon autonome sur des autoroutes spécialement équipées, ou lorsque les conditions sont idéales, ont été mis au point, mais ils ne sont pas encore commercialisés. Toutefois, conduire en toute sécurité et efficacement sur n'importe quel type de route et dans toutes les conditions de trafic exige plus que quelques capteurs et des circuits rétroactifs. En tant que bon conducteur vous avez une bonne intelligence du trafic, vous tenez compte des autres conducteurs, du comportement de leur véhicule, de la signalisation et d'une foule d'autres paramètres. Vous devez être capable de réagir efficacement à toute situation inattendue et remarquer si d'autres conducteurs ont une conduite à risque. Vous devez surveiller les feux clignotants des autres voitures, anticiper leur changement de couloir ou, si le clignotant dure longtemps, en déduire que le conducteur ignore qu'il fonctionne et n'a pas l'intention de changer de couloir. Vous devez reconnaître qu'une volute de fumée loin devant signifie qu'un accident vient peut-être de se produire et que vous devez ralentir. Si un ballon traverse la route devant vous, vous devez aussitôt penser que l'enfant qui lui court après risque de se précipiter à travers la chaussée, et donc freiner fortement.

Supposons que vous désiriez construire une voiture véritablement intelligente. Votre première tâche sera de choisir un ensemble de capteurs lui permettant de s'informer sur l'environnement. Ce sera peut-être une caméra, voire deux, l'une à l'avant, l'autre à

l'arrière, et un microphone pour écouter et analyser les bruits et les paroles. Peut-être installerez-vous un radar ou des capteurs à ultrasons qui détermineront avec exactitude la distance et la vitesse des autres automobiles, en plein jour ou dans l'obscurité. Le point important est que nous n'avons pas à nous limiter aux sens de l'être humain. L'algorithme cortical est flexible, et pour peu que le système de mémoire hiérarchisée ait été bien conçu, il devrait fonctionner quel que soit le type de capteurs installés. Théoriquement, cette voiture devrait appréhender mieux que nous le trafic environnant car ses sens artificiels seraient choisis en fonction de cette tâche. Les capteurs seraient liés à un système de mémoire hiérarchisée suffisamment dimensionné. Les concepteurs d'un véhicule intelligent meubleraient sa mémoire en exposant le système à des conditions de conduite réelles afin qu'il apprenne à élaborer un modèle du monde comme le font les humains, mais dans un domaine plus limité (le véhicule doit tout connaître de la route, mais peut se passer de la connaissance des silos à grains et des avions dans le ciel). La mémoire de la voiture apprendrait la structure hiérarchique du trafic et des routes afin que le système puisse comprendre et anticiper ce qui passe en termes de mouvements de véhicules, de signalisation routière, d'obstacles et d'intersections. Les ingénieurs pourraient concevoir le système de mémoire afin qu'il pilote réellement la voiture, ou se contente de surveiller ce qui se passe pendant que vous conduisez. Il pourrait prodiguer des conseils à la manière d'un instructeur invisible. La mémoire étant bien entraînée, la voiture peut comprendre et gérer n'importe quelle situation. Les ingénieurs auront alors le choix d'enregistrer la mémoire une fois pour toutes, afin de la copier dans toutes les voitures quittant les chaînes d'assemblage, ou de faire en sorte que la mémoire puisse continuer à apprendre après la vente du véhicule. De plus, ce qui est possible pour un ordinateur mais pas pour un humain, la mémoire pourrait être reprogrammée avec une nouvelle version tenant compte de nouvelles conditions ou règles de conduite.

Je ne dis pas qu'il est sûr et certain que nous fabriquerons des voitures intelligentes ou des machines qui comprendront le langage et sauront regarder. Mais ce sont d'excellents exemples du genre d'équipement que nous pourrions étudier et développer, et qu'il est possible de construire.

---

Personnellement, les applications évidentes des machines intelligentes m'intéressent moins. Pour moi, le véritable bénéfice et l'engouement pour une nouvelle technologie consistent à en trouver des usages inconcevables auparavant. De quelle manière les machines intelligentes nous surprendront-elles et quelles fabuleuses capacités verront le jour ? Je suis persuadé qu'à l'instar du transistor et du microprocesseur, les mémoires hiérarchiques transformeront nos vies d'une façon incroyable, pour notre plus grand bien. Comment cela ? Une manière d'entrevoir l'avenir des machines intelligentes est de penser aux aspects de la technologie qui s'y adaptent bien. Autrement dit, des éléments qui seront de plus en plus bon marché, de plus en plus rapides, de plus en plus réduits. Tout ce qui croît à une vitesse exponentielle dépasse rapidement notre imagination et tend à jouer un rôle primordial dans les évolutions les plus radicales des technologies futures.

Citons parmi les technologies qui ont connu une croissance exponentielle : les composants à base de silicium, le disque dur, les techniques de séquençage de l'ADN et la fibre optique. Ces technologies qui s'adaptent rapidement ont été à la base de nombreux nouveaux produits et de nouveaux secteurs industriels. D'une manière différente, les logiciels sont eux aussi adaptables. Une fois écrit, un programme peut être copié à l'infini pour un coût dérisoire.

En revanche, certaines technologies comme les batteries, les moteurs et la robotique traditionnelle ne s'adaptent que peu. En dépit de nombreux efforts et de perfectionnements constants, un bras robot fabriqué aujourd'hui n'est pas significativement

meilleur qu'un autre construit il y a quelques années. Les développements de la robotique sont progressifs et modestes ; ils n'ont rien de commun avec la courbe exponentielle de la conception des puces ou la prolifération des logiciels. Un bras robot qui coûtait un million de dollars en 1985 ne peut pas être fabriqué aujourd'hui pour à peine dix dollars tout en étant mille fois plus puissant. De même, les batteries dont nous disposons aujourd'hui ne sont pas beaucoup mieux que celles d'il y a dix ans. Vous objecterez peut-être qu'elles sont deux ou trois fois mieux, mais ce n'est pas mille ou dix mille fois ; en ce domaine, les progrès se font à petits pas. Si la capacité des batteries avait augmenté dans les mêmes proportions que la capacité des disques durs, les téléphones mobiles et autres appareillages électroniques n'auraient jamais à être rechargés, et des voitures électriques ultralégères parcourant mille ou deux mille kilomètres avec une seule charge pulluleraient sur nos routes.

Il nous incombe de découvrir les aspects des systèmes mnémoniques artificiels qui surpasseront considérablement nos cerveaux biologiques. Ce sont eux qui nous indiqueront ce que deviendra finalement cette technologie. Je vois quatre domaines dans lesquels ils dépasseront nos capacités : la vitesse, la capacité, la réplicabilité et les systèmes sensoriels.

## *La rapidité*

Tandis que le temps de réaction des neurones est de l'ordre de la milliseconde, les composants sur silicium réagissent dans l'ordre de la nanoseconde, et cette rapidité augmente. C'est une différence d'un million de fois, c'est-à-dire un ordre de grandeur de six. La différence de vitesse entre l'intelligence fondée sur l'organique et celle fondée sur l'électronique est lourde de conséquences. Les machines intelligentes seront capables de penser un million de fois plus vite que le cerveau humain. Un tel esprit artificiel pourrait assimiler des bibliothèques entières de livres, ou étudier une quantité faramineuse de données — des tâches qu'il

nous faudrait des années pour mener à bien — en quelques minutes, tout en retirant exactement la même compréhension de cette lecture ultrarapide. Il n'y a là rien de magique. Le cerveau humain a évolué avec deux contraintes de temps. L'une est la rapidité de réaction des cellules, l'autre la vitesse à laquelle le monde change. Il n'y aurait aucun avantage pour un cerveau biologique à penser un million de fois plus vite alors que le monde qui l'entoure est fondamentalement lent. Mais rien, dans l'algorithme cortical, ne dit qu'il doit toujours fonctionner lentement. Si une machine intelligente conversait ou interagissait avec un humain, elle serait obligée de ralentir pour fonctionner au rythme de l'intelligence humaine. Quand je lis un livre, tourner les pages ralentit la lecture. Mais si le livre est électronique, ce geste n'a plus lieu d'être et je lis un peu plus vite. Deux machines intelligentes peuvent tenir une conversation un million de fois plus rapide que si c'étaient des humains. Imaginez les progrès que permettrait une machine intelligente résolvant des problèmes mathématiques ou scientifiques un million de fois plus vite que les humains. En dix secondes, vous obtiendriez le résultat qui vous aurait demandé un mois de cogitations. Une telle intelligence fulgurante, jamais fatiguée, jamais lassée, se rendra utile par des moyens que nous ne pouvons pas encore imaginer.

## *La capacité*

En dépit de l'impressionnante capacité mnémonique du cortex humain, il est possible de créer des machines intelligentes qui la surpassent considérablement. La taille de notre cerveau est limitée par de sévères facteurs biologiques, notamment la taille de la boîte crânienne du fœtus par rapport au diamètre intérieur du bassin de la mère, le coût métabolique élevé du fonctionnement du cerveau (l'encéphale ne représente qu'environ 2 % du poids du corps mais consomme environ 20 % de l'oxygène que nous inspirons) et la lenteur des neurones. En revanche, nous pouvons construire des systèmes mnémoniques intelligents de n'importe

quelle taille, et contrairement au processus sinueux et aveugle de l'évolution des espèces, nous pouvons planifier longtemps à l'avance les intentions spécifiques et les détails de conception. D'ici à quelques décennies, la capacité du néocortex humain se révélera sans doute relativement modeste, comparée à celle des machines.

Lorsque nous construirons des machines intelligentes, nous pourrons accroître leur capacité de mémoire de plusieurs façons. L'augmentation de la profondeur de la hiérarchie conduira à une compréhension plus profonde, c'est-à-dire une capacité de déceler des patterns d'ordre plus élevé. Augmenter la capacité à l'intérieur des régions permet à la machine de mémoriser davantage de détails, ou de percevoir avec une plus grande acuité, de la même manière qu'un aveugle a affiné ses sens du toucher et de l'audition. L'ajout de nouveaux sens et hiérarchies sensorielles permettra à l'appareil d'élaborer de meilleurs modèles du monde, comme je l'expliquerai d'ici peu.

Il serait intéressant de voir s'il existe une limite supérieure à la taille d'un système de mémoire intelligent et d'en connaître les dimensions. Il n'est pas exclu que celle d'un appareil puisse devenir trop encombrée de données pour être utilisable, ou alors que le système tombe en panne lorsqu'une certaine limite théorique est proche. Le cerveau humain est peut-être déjà proche de cette taille maximale théorique, mais je pense que c'est peu probable. A l'échelle de l'évolution, il n'est devenu surdimensionné que très récemment, et rien n'indique qu'il existe une taille maximale définitive. Quel que soit le pic de capacité qui puisse se révéler pour un système de mémoire intelligent, le cerveau humain n'a certainement pas atteint le sien. Il en est sans doute encore loin.

Une façon d'envisager ce que pourraient être ces systèmes est d'étudier les limites connues des performances humaines. Albert Einstein était incontestablement d'une extrême intelligence, mais son cerveau n'était jamais qu'un cerveau. Nous pouvons supposer que son extraordinaire intelligence découlait largement de diffé-

rences physiques entre son cerveau et celui de la moyenne des gens. Ce qui fait la rareté d'un tel personnage est le fait que le génome ne produit pas souvent des encéphales de cette qualité. Cependant, le jour où nous graverons des cerveaux dans le silicium, nous pourrons en créer autant que nous voudrons. Ils auront tous les hautes capacités intellectuelles d'un Einstein, et peut-être même plus. A l'autre extrême, certaines pathologies mentales, comme le syndrome d'Asperger — qui est probablement une forme d'autisme —, nous donnent un aperçu d'autres dimensions possibles de l'intelligence. Ces personnes révèlent des capacités remarquables, comme une mémoire quasiment photographique ou l'exécution instantanée de calculs extrêmement compliqués (opérations sur de très grands nombres, détermination de nombres premiers, extraction de racines, calculs sur des dates éloignées…). Bien qu'atypique, leur cerveau n'en est pas moins un cerveau qui exploite l'algorithme cortical. Si un cerveau atypique est capable de manifester d'incroyables capacités mnémoniques, nous devrions théoriquement pouvoir les intégrer à un cerveau artificiel. Non seulement ces deux extrêmes des capacités cérébrales révèlent ce que l'on peut espérer recréer, mais ils indiquent aussi les domaines dans lesquels il sera possible de surpasser les performances humaines les plus élevées.

## *La réplicabilité*

Chaque nouveau cerveau organique doit se développer et s'astreindre à un apprentissage à partir de zéro, un processus qui, pour l'être humain, exige des dizaines d'années. Chaque humain doit découvrir par lui-même les bases de la coordination de ses membres et des groupes de muscles, de l'équilibre et de la marche, et apprendre les propriétés générales d'une multitude d'objets, d'animaux et d'autres gens. Il doit apprendre le nom des choses et la structure du langage, ainsi que les règles de la famille et de la société. Ces bases maîtrisées, des années de scolarité l'attendent. Au cours de sa vie, chaque individu est confronté à

un même ensemble d'apprentissages, et même si c'est plus dur pour certains que pour d'autres, chacun construit son propre modèle du monde dans son cortex.

Les machines intelligentes n'auront pas à subir ce long apprentissage, car les composants électroniques et autres supports de stockage peuvent être dupliqués à l'infini et leur contenu facilement transféré. A cet égard, les machines intelligentes pourraient se répliquer comme les logiciels. Une fois que le prototype d'un système a été mis au point et entraîné de façon satisfaisante, il peut être recopié à volonté. Il peut falloir des années pour concevoir le composant et la configuration matérielle, procéder aux apprentissages, perfectionner le système empiriquement, tout cela pour améliorer le système mnémonique d'une voiture intelligente, mais ces phases terminées, il pourra être produit en masse. Comme je l'ai mentionné précédemment, nous pourrions faire en sorte que chaque copie puisse continuer à apprendre, ou empêcher cette aptitude. Pour certaines applications, nous permettrions aux machines intelligentes de ne fonctionner que selon des schémas testés et connus. Une fois qu'une voiture intelligente sait aller partout où elle doit, vous ne voudriez pas qu'elle prenne la mauvaise habitude de se fier à de fausses analogies qui la détourneraient d'un trajet ? Ou encore, nous pourrions faire en sorte que tous les véhicules d'une même fabrication se comportent de la même manière. Mais pour d'autres applications, il faudra que le système mnémonique fondé sur le cerveau puisse continuer à apprendre sans cesse. Par exemple, une machine intelligente conçue pour découvrir des démonstrations mathématiques devra être capable d'apprendre par l'expérience, d'appliquer des acquis anciens à des problèmes nouveaux, et d'être globalement flexible et ouverte.

Il devrait être possible de partager des composants d'éléments appris tout comme nous partageons des éléments de logiciels. Une machine intelligente, d'une conception particulière, devrait pouvoir être reprogrammée grâce à un nouveau jeu de

connexions conduisant à un comportement différent ; ce serait comme si je pouvais télécharger un nouveau jeu de connexions dans votre cerveau et vous faire ainsi immédiatement passer de francophone à anglophone, ou de professeur de sciences politiques à musicologue. Les gens pourraient passer d'une compétence à une autre et reprendre le travail d'autrui. Supposons que j'aie développé et entraîné une machine équipée d'un système de vision supérieur, et que quelqu'un d'autre ait développé une machine aux capacités auditives supérieures. Si la conception était bien faite, il serait possible de combiner le meilleur des deux systèmes sans reprendre chaque apprentissage depuis le début. Cette sorte de partage et de permutation des compétences est absolument impossible pour des humains. Le secteur de la construction de machines intelligentes pourrait évoluer sur la même voie que l'industrie informatique, avec des groupes de techniciens entraînant les machines intelligentes afin qu'elles acquièrent des aptitudes et des connaissances spécialisées, et qui ensuite vendraient et échangeraient des configurations de mémoire. Reprogrammer une machine intelligente ne serait guère différent du changement d'une cartouche de jeu vidéo ou de l'installation d'un nouveau logiciel.

## *Les systèmes sensoriels*

Les humains disposent de quelques sens. Ils sont profondément enracinés dans nos gènes, dans notre corps et dans le câblage sous-cortical de notre cerveau. Nous ne pouvons pas les modifier. Nous avons parfois recours à la technologie pour améliorer nos sens, comme les lunettes de vision nocturne, le radar ou le télescope spatial Hubble. Ces instruments de haute technologie sont de brillants moyens de conversion de données, mais pas des nouveaux modes de perception. L'information que nous ne pouvons percevoir, ils la traduisent en affichages visuels ou auditifs que nous pouvons interpréter. Mais là encore, c'est à la fantastique flexibilité de notre cerveau que nous devons de pouvoir regarder

l'écran d'un radar et comprendre ce qu'il représente. De nombreuses espèces animales sont dotées de sens véritablement différents, comme l'écholocation des chauves-souris et des dauphins, l'aptitude des abeilles à voir la lumière polarisée et l'ultraviolet, et la sensibilité aux champs électriques de certains poissons.

Les machines intelligentes pourraient non seulement percevoir le monde par n'importe quel sens existant dans la nature, mais aussi par des sens nouveaux entièrement conçus par l'homme. Le sonar, le radar et la vision infrarouge sont des exemples évidents de sens non humains dont les machines intelligentes seront assurément dotées. Mais ce n'est qu'un début.

Bien plus intéressante, et de loin, est la manière dont les machines intelligentes pourraient s'ouvrir à des mondes de sensations étrangères, à des expériences authentiquement différentes. Comme nous l'avons vu, l'algorithme néocortical est fondamentalement concerné par l'identification des patterns dans le monde environnant. Il n'a aucune préférence quant à leur origine physique. Aussi longtemps que les inputs vers le cortex ne seront pas aléatoires, présenteront une certaine richesse ou une structure statistique, un système intelligent s'en servira pour former des mémoires invariantes et des prédictions. Il n'y a aucune raison pour que ces patterns d'entrée soient analogues aux sens des animaux, ou même découlent du monde réel. C'est, je pense, dans le domaine des sens différents que réside l'utilisation révolutionnaire des machines intelligentes.

Par exemple, nous pourrions concevoir un système sensoriel recouvrant le globe entier. Imaginez un maillage de capteurs météorologiques distants de cinquante ou cent kilomètres les uns des autres, répartis sur tout un continent. Ces capteurs seraient analogues aux cellules de la rétine. D'énormes éléments climatiques comme les tempêtes et les fronts se déplacent et changent sans cesse. En reliant le réseau de capteurs à une vaste mémoire de type corticale, nous pourrions permettre au système d'apprendre à prévoir le temps de la même manière que vous et moi

apprenons à reconnaître des objets et prédire comment ils évolueront. Le système discernerait des patterns météorologiques locaux et régionaux, des patterns qui existent depuis des décennies, des années ou des heures. En rapprochant ces capteurs dans certaines régions, nous créerions l'équivalent de la fovéa, permettant au système de météorologie intelligent de prévoir les microclimats. Notre cerveau météorologique parviendrait à penser aux systèmes météorologiques globaux et à les comprendre, tout comme vous pensez à des gens et les comprenez. Les météorologues travaillent actuellement à partir de maillages. Ils collectent des enregistrements provenant de divers lieux et qu'ils introduisent dans des supercalculateurs pour simuler le climat et prévoir le temps. Mais cette approche, fondamentalement différente de celle d'une machine intelligente, est comparable à un ordinateur jouant aux échecs — bête et dépourvu de cognition —, alors que notre machine météorologique intelligente serait comparable à la manière de jouer aux échecs d'un humain : en réfléchissant et en comprenant. La machine météorologique intelligente découvrirait des patterns qui échappent aux humains. Ce n'est qu'en 1960 que le phénomène climatique appelé El Niño a été découvert. Notre machine météorologique pourrait trouver d'autres patterns comme El Niño, ou encore apprendre comment prédire les tornades ou les moussons bien mieux que les humains. Présenter un grand nombre de données climatiques sous une forme facilement déchiffrable par l'homme est difficile ; en revanche, le cerveau météorologue pourrait détecter et traduire directement les phénomènes météorologiques.

D'autres systèmes sensoriels largement répartis pourraient permettre de construire des machines intelligentes comprenant et prédisant les migrations animales, les changements démographiques et la propagation des épidémies. Supposons que des capteurs soient installés sur l'ensemble du réseau de distribution électrique d'un pays. La machine intelligente reliée à ces capteurs observerait les variations de la consommation électrique de la

même manière que l'on peut observer le flux du trafic routier ou les mouvements de foule dans un aéroport. Par des expositions répétées, les humains parviennent à prédire ces patterns (demandez à un régulateur du trafic routier ou à un vigile de sécurité d'un aéroport). De même, la surveillance intelligente du maillage, sur le réseau électrique, sera plus apte qu'un humain à prévoir une demande d'énergie, ou encore des situations périlleuses comme un dysfonctionnement risquant de provoquer une panne de courant. Il devrait être possible de combiner des capteurs météorologiques et des capteurs démographiques afin d'anticiper les troubles politiques, les famines ou les épidémies. Agissant comme un diplomate extrêmement fin, une machine intelligente pourrait jouer un rôle dans l'atténuation des conflits et des souffrances humaines. Peut-être pensez-vous que ces machines devraient ressentir des émotions pour être sensibles aux patterns impliquant le comportement humain, mais je ne le pense pas. Nous ne sommes pas nés avec un bagage culturel, un bagage religieux et un ensemble de valeurs ; tout cela, nous l'apprenons. De même que je peux apprendre à comprendre les gens dont les motivations sont différentes des miennes, une machine intelligente peut comprendre les motivations et les émotions humaines, même si par elle-même elle en est dépourvue.

Nous pourrions créer des sens capables d'échantillonner de microscopiques entités. Il est théoriquement possible que des capteurs représentent des patterns de cellules ou de macromolécules. Par exemple, un des grands défis actuels est de comprendre comment la forme d'une protéine peut être prédite d'après la séquence d'acides aminés dont elle est composée. Savoir comment les protéines se replient et interagissent accélérerait la recherche pharmaceutique et par voie de conséquence le traitement de nombreuses maladies. Afin de prédire le comportement de ces molécules complexes, les ingénieurs et les scientifiques ont créé des modèles tridimensionnels de protéines, mais la tâche s'est révélée difficile. Une machine hyper intelligente, équipée de sens spécialement dévelop-

pés pour la chimie organique, serait en mesure de trouver une réponse. Si cela vous semble tiré par les cheveux, rappelez-vous que vous ne seriez pas surpris si un humain découvrait la solution. Notre inaptitude à aborder cette question pourrait être principalement liée à un trop grand décalage entre les sens humains et la nature du phénomène physique à comprendre. Des machines intelligentes pourront être dotées de sens personnalisés et d'une mémoire plus vaste que celle des humains, leur permettant de résoudre des problèmes qui ne sont pas à notre portée.

Equipées des sens adéquats et au prix d'une petite restructuration de la mémoire corticale, nos machines intelligentes pourraient vivre et penser dans les mondes virtuels propres aux mathématiques et à la physique. Par exemple, beaucoup de découvertes en mathématiques et en sciences exigent une compréhension du comportement des objets dans des mondes à plus de trois dimensions. De nombreux théoriciens qui ont étudié la nature de l'espace sidéral pensent que l'univers comporte une dizaine de dimensions, voire davantage. Les humains ont beaucoup de difficultés à conceptualiser des problèmes mathématiques à quatre dimensions ou plus. Une machine intelligente, judicieusement conçue, pourrait appréhender ces espaces hautement multidimensionnels avec la même aisance que nous nous représentons un espace tridimensionnel, et par conséquent prédire comment ils se comportent.

Enfin, il serait possible d'unifier un ensemble de systèmes intelligents au sein d'une hiérarchie globale, tout comme le cortex unifie l'ouïe, le toucher et la vue en remontant dans la hiérarchie corticale. Un tel système apprendrait automatiquement comment modéliser et prédire les patterns de pensée dans des populations de machines intelligentes. Grâce à des supports de communication répartie comme l'Internet, chacune des machines pourrait se trouver en un lieu différent, quelque part dans le monde. Les hiérarchies de grande taille apprennent des patterns plus profonds et discernent des analogies plus complexes.

Toute cette prospective annonce des machines intelligentes qui surpasseront de différentes manières nos aptitudes, dans des proportions considérables. Elles apprendront et penseront un million de fois plus vite que nous, se souviendront d'énormes quantités d'informations détaillées ou distingueront des patterns incroyablement abstraits. Leurs sens seront plus sensibles que les nôtres, ou alors ils seront répartis, ou encore ils détecteront des phénomènes infimes. Elles penseront en trois, quatre ou plus encore de dimensions. Aucune de ces intéressantes possibilités ne repose sur l'imitation des caractéristiques de l'être humain et de ses actions, et elles n'exigeront aucune robotique complexe.

Nous pouvons à présent constater pleinement comment le test de Turing, en assimilant l'intelligence au comportement humain, avait limité notre vision de ce qui est possible. En comprenant d'abord ce qu'est l'intelligence, nous pourrons construire des machines autrement plus utiles qu'en nous contentant de copier le comportement humain. Nos machines intelligentes seront de fantastiques outils qui contribueront à accroître notre connaissance de l'univers.

---

Dans combien de temps tout cela deviendra-t-il réalité ? Fabriquerons-nous des machines intelligentes dans cinquante ans, dans vingt ans ou dans cinq ans ? Un adage, dans les milieux high-tech, dit que le changement est plus long que prévu dans le court terme, mais se produit plus rapidement que prévu dans le long terme. Je l'ai bien souvent constaté. Quelqu'un tient une conférence, annonce une nouvelle technologie et affirme que dans quatre ans tout le monde l'aura adoptée. Et bien sûr, il se trompe. Les quatre ans en deviennent huit et le public commence à croire qu'il ne la verra jamais. Mais juste à ce moment, alors que l'idée semblait s'être enlisée dans une impasse, elle reprend du poil de la bête et devient sensationnelle. Quelque chose de semblable pourrait se

produire pour les machines intelligentes. Les progrès sembleront d'abord lents, puis tout décollera rapidement.

Lors de conférences consacrées aux neurosciences, j'aime bien faire le tour de la salle et demander à tout le monde combien de temps s'écoulera, à leur avis, avant que nous n'obtenions une théorie du cortex. Quelques personnes, moins de 5 %, disent « jamais » ou « nous en avons déjà une » (des réponses surprenantes quand on sait ce que ces gens font dans la vie). Un autre groupe de 5 % dit « dans cinq à dix ans ». La moitié de ceux qui restent pense qu'il faudra attendre dix à cinquante ans, ou espère du moins la voir de leur vivant. Le dernier groupe dit « de cinquante à deux cents ans », en tout cas pas de leur vivant. Je me range du côté des optimistes. Nous avons vécu la période « lente » depuis des décennies, de sorte que pour beaucoup de gens les progrès dans la neurobiologie théorique et les machines intelligentes avaient complètement marqué le pas. Si j'en juge des progrès effectués ces trente dernières années, il est tout à fait naturel d'en conclure que nous ne sommes pas près d'obtenir une réponse. Mais je pense que nous sommes à un tournant et que ce domaine va bientôt décoller.

Il est possible d'accélérer le futur, de rapprocher le tournant plus près du présent. Un des buts de ce livre est de vous convaincre qu'avec un cadre théorique correct nous ferons des progrès rapides dans la compréhension du cortex, qu'en prenant le cadre de mémoire-prédiction pour guide, nous finirons par déchiffrer, dans ses moindres détails, le fonctionnement du cerveau et comment nous pensons. C'est la connaissance dont nous avons besoin pour construire les machines intelligentes. Si ce modèle est bon, les progrès seront rapides.

Donc, bien que j'hésite à prédire le moment où les machines intelligentes seront une réalité, je pense que si suffisamment de gens s'attaquent aujourd'hui à la résolution des problèmes, nous devrions pouvoir créer des prototypes opérationnels et des simulations du cortex d'ici à quelques années. J'espère que d'ici à dix

ans les machines intelligentes seront l'un des domaines de la science et de la technologie les plus actifs. Je ne veux pas être plus précis car je sais combien il est facile de sous-estimer le temps qu'il faut pour que quelque chose d'important voie le jour. Pourquoi suis-je cependant si optimiste quant à la rapidité des progrès dans la compréhension du cerveau et la construction de machines intelligentes ? Ma confiance provient en grande partie du temps que j'ai déjà consacré au problème de l'intelligence. Quand je suis tombé amoureux des cerveaux, en 1979, je pensais que la résolution du puzzle qu'est l'intelligence pourrait se faire de mon vivant. Les années passant, j'ai attentivement observé le déclin de l'intelligence artificielle, l'ascension et la chute des réseaux neuronaux, et j'ai vécu « la Décennie du Cerveau » dans les années 1990. J'ai vu comment l'attitude envers la biologie théorique, et la neurobiologie en particulier, a évolué. J'ai vu comment les notions de prédiction, de représentation hiérarchique et de temps sont entrées dans le vocabulaire des neurosciences. J'ai vu les progrès de ma propre compréhension et ceux de mes collègues. Voici dix-huit ans que je me suis enthousiasmé pour le rôle de la prédiction, et de certaines manières, je l'ai toujours vérifié depuis. Parce que je suis immergé depuis deux décennies dans les neurosciences et l'informatique, mon cerveau a peut-être élaboré un modèle perfectionné discernant comment les changements technologiques et scientifiques se produisent, et ce modèle prédit des progrès rapides. Nous sommes à présent au tournant.

# ÉPILOGUE

**L'astronome** Carl Sagan avait coutume de dire que comprendre une chose ne diminue ni son merveilleux ni son mystère. Beaucoup de gens craignent que les révélations scientifiques entaillent un contrat conclu avec le merveilleux, comme si la connaissance ôtait le goût et la saveur des choses. Mais Sagan avait raison. En vérité, la compréhension nous met plus à l'aise vis-à-vis de notre rôle dans l'univers, et par là même l'univers devient encore plus attrayant et mystérieux. N'être qu'un grain infinitésimal dans l'infinitude du cosmos, mais vivant, conscient, intelligent et créatif, est de loin beaucoup plus passionnant que subsister sur une Terre plate, limitée, au centre d'un univers étriqué. Comprendre comment notre cerveau fonctionne ne diminue en rien le merveilleux et le mystère de l'univers, de nos vies, de notre futur. Notre étonnement ira en s'approfondissant lorsque nous appliquerons cette connaissance à la compréhension de nous-même, à la construction de machines intelligentes et à l'acquisition d'un savoir encore plus grand.

C'est pourquoi chercher à comprendre le cerveau afin de construire des machines intelligentes mérite d'être tenté. C'est la prochaine étape logique pour l'humanité.

Grâce à ce livre, j'espère avoir incité les jeunes ingénieurs et scientifiques à étudier le cortex, à adopter le cadre de mémoire-prédiction et à vouloir construire des machines intelligentes. Parvenue à son sommet, l'intelligence artificielle fut une belle aventure. Elle eut ses journaux, ses programmes d'étude, ses livres, ses modèles commerciaux et ses créateurs d'entreprise. Les réseaux neuronaux avaient eux aussi bénéficié d'un fantastique engouement lorsque ce domaine connut un épanouissement dans les années 1980. Hélas, pour envisager la construction de machines intelligentes, le cadre scientifique sous-jacent à l'intelligence artificielle et aux réseaux neuronaux n'était pas le bon.

Je propose aujourd'hui une voie nouvelle, plus riche de promesses. Que vous soyez lycéen ou étudiant, si ce livre vous a encouragé à travailler dans ces technologies — c'est-à-dire la mise au point des premières machines véritablement intelligentes, et par conséquent le démarrage d'un secteur d'activité innovant —, je vous engage à vous lancer dans l'aventure. Faites que tout cela arrive. L'une des clés du succès est de se lancer sans hésitation dans un domaine nouveau, même s'il n'est pas totalement certain qu'il aboutira à quelque chose. Le choix du moment est important : trop tôt, vous devrez vous battre, et si vous attendez que les incertitudes se lèvent, il sera trop tard. Je crois sincèrement que le moment est venu de travailler à la conception et à l'élaboration de systèmes de mémoire hiérarchisée inspirés du cortex. Ce champ est appelé à un développement immense, tant scientifiquement que commercialement. Les Intel et les Microsoft d'un secteur d'activité fondé sur les mémoires hiérarchiques naîtront dans les dix années à venir. Entreprendre à cette échelle peut être financièrement risqué et intellectuellement ardu, mais cela vaut la peine d'essayer. J'espère que vous vous joindrez à moi, et à tous les autres qui ont relevé le défi, pour participer à la création de l'une des technologies les plus fabuleuses que le monde aura connues.

# ANNEXE : LES PRÉDICTIONS TESTABLES

**Toute** théorie doit conduire à des prédictions qui peuvent être testées, car les tests expérimentaux sont l'unique moyen sûr pour vérifier la validité d'une idée nouvelle. Fort heureusement, le cadre de mémoire-prédiction est fondé sur la biologie et conduit à plusieurs prédictions spécifiques et nouvelles susceptibles d'être testées. Dans cette annexe, je répertorie des prédictions qui peuvent infirmer et/ou confirmer les hypothèses émises dans ce livre. Ce matériel est un peu plus avancé que celui du Chapitre 6, mais sa lecture n'est pas du tout requise pour comprendre le reste de l'ouvrage. Plusieurs de ces prédictions ne peuvent être effectuées que sur des sujets humains ou animaux éveillés, car ces tests impliquent l'attente et la prédiction d'un début de stimulus. Les prédictions ne sont pas présentées par ordre d'importance.

**Prédiction 1**

Nous devons trouver des cellules dans toutes les aires du cortex, y compris dans le cortex sensoriel primaire, qui révèlent une activité accrue par anticipation d'un événement sensoriel, mais aussi par opposition en réaction à un événement sensoriel.

Par exemple, le service de Tony Zador, au Cold Spring Harbor Laboratory, a découvert des cellules dans le cortex auditif primaire du rat qui s'excitent exactement au moment où l'animal s'attend à entendre un bruit, même si le bruit ne se produit pas (correspondance privée). Ceci devrait être une propriété générale du cortex. Nous devrions trouver une semblable activité anticipatoire dans le cortex visuel et dans le cortex somatosensoriel. Les cellules qui s'excitent par anticipation à un input sensoriel sont la définition même de la prédiction, une prémisse de base du cadre de mémoire-prédiction.

**Prédiction 2**

Plus une prédiction peut être spatialement spécifique, plus nous trouverons de cellules à proximité du cortex sensoriel primaire qui deviennent actives par anticipation à un événement.

Si un singe a été entraîné en lui présentant des séquences de patterns visuels afin qu'il puisse anticiper un pattern visuel particulier à un moment précis, des cellules devraient manifester une activité accrue précisément au moment où le pattern anticipé est attendu (réaffirmation de la prédiction 1). Si le singe a appris à s'attendre à un visage, nous devrions nous attendre à trouver des cellules anticipatoires dans les aires de reconnaissance des visages, mais pas dans les aires visuelles inférieures. Toutefois, si le singe fixe une cible et s'il a appris à attendre l'apparition d'un pattern particulier à un endroit précis dans son champ de vision, nous devrions trouver des cellules anticipatoires en V1 ou à proximité de V1. L'activité représentant la prédiction descend aussi loin qu'elle le peut dans la hiérarchie corticale, selon la spécificité de la prédiction. Parfois, elle peut continuer jusqu'aux aires sensoriel-

les primaires, et à d'autres moments elle s'arrête dans des régions supérieures. Des résultats analogues devraient exister pour d'autres modalités sensorielles.

**Prédiction 3**

Les cellules qui manifestent une activité accrue par anticipation à un input sensoriel devraient de préférence se trouver dans les couches corticales 2, 3 et 6, et la prédiction devrait cesser de descendre dans la hiérarchie au niveau des couches 2 et 3.

Les prédictions qui descendent dans la hiérarchie corticale le font à travers les cellules des couches 2 et 3, qui projettent ensuite vers la couche 6. Ces cellules de la couche 6 projettent largement sur la couche 1, dans la région en dessous de la hiérarchie, activant un autre ensemble de cellules de la couche 2 et de la couche 3, et ainsi de suite. Par conséquent, c'est dans les cellules de ces couches 2, 3 et 6 que nous devrions trouver une activité anticipatoire. Rappelez-vous que les cellules actives des couches 2 et 3 représentent un ensemble possible de colonnes actives ; ce sont des prédictions possibles. Les cellules actives, dans la couche 6, représentent un petit nombre de colonnes ; ce sont des prédictions spécifiques à une région du cortex. Lorsqu'une prédiction descend dans le cortex, l'activité s'arrête finalement aux couches 2 et 3. Supposons par exemple qu'un rat ait appris à anticiper un son parmi deux sons. En se fondant sur un signe extérieur, il sait quand il va entendre l'un des sons, mais ne peut prédire lequel. Dans ce scénario, nous devrions nous attendre à détecter une activité anticipatoire dans les couches 2 ou 3, dans les colonnes qui représentent chacun des deux sons. Il ne devrait y avoir aucune activité dans la couche 6 de la même région, car l'animal ne peut pas prédire quel son spécifique il va entendre. Lors d'un autre essai, si l'animal parvient à prédire exactement le son, nous devrions détecter une activité dans la couche 6, dans les colonnes qui réagissent à ce son spécifique.

Nous ne pouvons complètement exclure la possibilité de trouver des cellules anticipatoires dans les couches 4 et 5. Par exemple, il est probable qu'il y ait plusieurs classes de cellules dans ces couches dont la fonction nous est inconnue. Par conséquent, cette prédiction est relativement faible, mais je pense qu'elle mérite d'être mentionnée.

**Prédiction 4**

Une classe de cellules des couches 2 et 3 devrait de préférence recevoir un input des cellules de la couche 6, dans les régions corticales plus élevées.

Une partie du modèle de mémoire-prédiction stipule que les séquences de patterns apprises qui se produisent ensemble développent une représentation invariante temporairement constante, que j'appelle un « nom ». Je suppose que ce nom est un ensemble de cellules des couches 2 ou 3 à travers une région du cortex, dans différentes colonnes. L'ensemble de cellules reste actif aussi longtemps que des événements, membres de la séquence, se produisent (par exemple, un ensemble de cellules restant actif aussi longtemps qu'une note d'une mélodie est entendue). Cet ensemble de cellules représentant le nom de la séquence est activé par le feedback des cellules de la couche 6, dans les régions supérieures du cortex. Je suggère que ces cellules « nom » sont des cellules de la couche 2 à cause de leur proximité avec la couche 1. Mais il pourrait s'agir de n'importe quelle classe de cellules dans les couches 2 et 3, qui ont des dendrites dans la couche 1. Pour que le système de nommage fonctionne, les dendrites apicales de ces cellules « nom » doivent former des synapses de préférence avec les axones de la couche 1 provenant de la couche 6 des régions supérieures. Elles doivent éviter de former des synapses avec les axones de la couche 1 provenant du thalamus. La théorie laisse entendre que nous devrions trouver une classe de cellules, dans les couches 2 et 3, ayant des dendrites apicales dans la couche 1, qui ont une forte prédilection à former des synapses avec les axones des cellules de la couche 6, dans la région

au-dessus. D'autres cellules avec des synapses de la couche 1 ne devraient pas manifester cette préférence. C'est une puissante et, pour autant que je sache, toute nouvelle prédiction.

Selon une prédiction corollaire, nous devrions trouver une autre classe de cellules dans les couches 2 ou 3 dont les dendrites apicales forment des synapses, de préférence avec des axones provenant de régions non spécifiques du thalamus. Ces cellules prédisent les prochains éléments d'une séquence.

**Prédiction 5**

Un ensemble de cellules « nom » décrit dans la prédiction 4 doit rester actif pendant les séquences apprises.

Un ensemble de cellules qui reste actif au cours d'une séquence apprise est la définition d'un « nom » pour une séquence prédictible. C'est pourquoi nous devrions trouver des ensembles de cellules qui restent actives même si l'activité des cellules dans le reste de la colonne (les cellules des couches 4, 5 et 6) change. Nous ne pouvons hélas pas dire à quoi ressemblera l'activité des cellules « nom ». Par exemple, l'activité constante d'un pattern de nom peut être aussi simple qu'un potentiel opérant à l'unisson sur un ensemble de cellules « nom ». Par conséquent, ce groupe de cellules actives peut être difficile à détecter.

**Prédiction 6**

Une autre classe de cellules dans les couches 2 ou 3 (différente des cellules « nom » évoquées dans les prédictions 4 et 5) devrait être active en réponse à un input non anticipé, mais inactive en réponse à un input anticipé.

L'idée derrière cette prédiction est que les événements non anticipés doivent être passés en haut, dans la hiérarchie corticale, mais si l'événement est anticipé, nous ne le passerons pas en haut précisément parce qu'il a été prédit localement. Par conséquent, il devrait y avoir une classe de cellules, dans les couches 2 ou 3, différente de la classe « nom » décrite dans les

prédictions 4 et 5, qui manifeste une activité lorsqu'un événement non anticipé se produit, mais n'en manifeste pas si l'événement a été anticipé. Les axones de ces cellules doivent projeter vers des régions plus hautes du cortex. Je propose un mécanisme pour obtenir le changement d'activité. Une telle cellule pourrait être inhibée *via* un interneurone activé par une cellule « nom », mais à ce point, il n'existe aucun moyen de faire une prédiction fiable du mécanisme. Tout ce que nous pouvons dire est que quelques cellules doivent manifester cette activité différentielle. Voilà encore une forte et, pour autant que je sache, toute nouvelle prédiction.

**Prédiction 7**

D'après la prédiction 6, les événements non anticipés doivent se propager vers le haut de la hiérarchie. Plus l'événement est nouveau, plus l'input non anticipé doit monter haut. Des événements complètement nouveaux devraient atteindre l'hippocampe.

Les patterns très bien appris sont prédits plus bas dans la hiérarchie, et réciproquement, plus un input est nouveau, plus il devrait se propager haut dans la hiérarchie. Il devrait être possible d'élaborer une expérience permettant de rendre cette différence. Par exemple, un humain écoute une mélodie qui ne lui est pas familière, mais qui est simple. Si le sujet entend une note qui, bien qu'inattendue, s'accorde avec le style de la musique, la note inattendue devrait provoquer des changements d'activité dans le cortex auditif, vers quelque niveau plus haut de la hiérarchie corticale. Toutefois, si au lieu d'entendre une note qui s'accorde avec le style de la musique le sujet entend un son complètement incongru, comme un bruit de casse, nous pensons que le changement d'activité dû à ce son montera plus haut dans la hiérarchie corticale. Le résultat devrait être permuté si le sujet s'est attendu au bruit de casse et qu'à la place il a entendu la note. Il devrait être possible de tester cette prédic-

tion sur des êtres humains à l'aide de l'IRMf (imagerie par résonance magnétique fonctionnelle).

**Prédiction 8**

Une compréhension subite doit produire une cascade précise d'activités prédictives descendant le long de la hiérarchie corticale.

Le moment où l'on s'écrie « bon sang, mais c'est bien sûr ! », lorsqu'un curieux pattern sensoriel est finalement compris — comme la reconnaissance du dalmatien de la Figure 12 —, commence lorsqu'une région du cortex tente de faire concorder une nouvelle mémorisation avec ses inputs. Si la concordance correspond à la région locale, les prédictions sont passées vers le bas dans la hiérarchie corticale, en rapide succession vers toutes les régions inférieures. Si c'est là une interprétation correcte du stimulus, chaque région de la hiérarchie va choisir une prédiction correcte en une rapide succession. Le même effet devrait se produire lorsque vous regardez une image à deux interprétations, comme la silhouette qui peut être celle d'un vase ou de deux profils de visages se faisant face, ou encore un cube de Necker (une image montrant des cubes qui semblent orientés dans un sens ou dans l'autre selon la subjectivité du regard). Chaque fois que la perception d'une de ces illusions optiques change, nous devrions constater la propagation de nouvelles prédictions descendant dans la hiérarchie. Aux niveaux les plus bas, disons V1, une colonne représentant un segment de ligne de l'image devrait rester active quelle que soit la perception de l'illusion optique (les yeux n'ayant pas bougé). Toutefois, nous devrions constater, dans cette colonne, que l'état actif ou inactif de certaines cellules change. C'est-à-dire que la même caractéristique de bas niveau existe dans chaque image, mais différentes cellules d'une colonne peuvent être actives selon l'interprétation. Ce qui est important est qu'une propagation de prédictions devrait descendre en bas de la hiérarchie corticale lorsqu'une perception de haut niveau change.

Une propagation de prédictions semblable devrait se produire à chaque saccade oculaire effectuée sur un objet visuel appris.

**Prédiction 9**

Le cadre de mémoire-prédiction requiert que des neurones pyramidaux puissent détecter les coïncidences précises d'un input synaptique sur des dendrites minces.

Pendant des années, on a cru que les neurones pouvaient être de simples intégrateurs accumulant les inputs de toutes leurs synapses afin de déterminer si un neurone doit émettre un potentiel. Pour la neurobiologie d'aujourd'hui, il y a beaucoup plus d'incertitudes quant au comportement des neurones. Certains soutiennent encore l'idée que les neurones sont de simples intégrateurs, et beaucoup de modèles de réseaux neuronaux sont constitués de neurones fonctionnant de cette manière. Il existe aussi un grand nombre de modèles de neurones qui présument que ces derniers se comportent comme si chaque partie dendritique opérait indépendamment. Le modèle de mémoire-prédiction exige que les neurones soient capables de détecter les coïncidences de seulement quelques synapses qui seraient actives au cours d'un laps de temps réduit. Le modèle pourrait fonctionner même avec une seule synapse potentialisée, qui serait suffisante pour entraîner l'excitation d'une cellule, mais plus probablement, deux synapses actives ou plus devraient se trouver à proximité d'une dendrite mince. De ce fait, un neurone ayant des milliers de synapses pourrait apprendre à être excité par beaucoup de patterns d'entrée différents, précis et distincts. Ce n'est pas une idée nouvelle, et des évidences inclinent à la soutenir. C'est toutefois un abandon radical du modèle standard préconisé depuis de nombreuses années. S'il était démontré que les neurones ne s'excitent pas selon des patterns d'entrée précis et clairsemés, il serait difficile de préserver l'intégrité de la théorie de mémoire-prédiction. Les synapses sur les dendrites épaisses,

sur ou à proximité des corps cellulaires n'ont pas besoin de fonctionner ainsi, seulement les nombreuses synapses sur les dendrites minces.

**Prédiction 10**

Avec l'apprentissage, les représentations descendent le long de la hiérarchie.

Je soutiens que par un apprentissage répété le cortex réapprend les séquences dans les régions hiérarchiquement plus basses. Ceci découle naturellement de la manière dont la mémoire des séquences de patterns change le pattern d'entrée transmis aux prochaines régions supérieures du cortex. Ce processus entraîne deux conséquences. L'une est que nous trouverons des cellules qui répondent à un stimulus complexe plus bas dans le cortex après un entraînement intensif, et plus haut dans le cortex après un entraînement minimal. Chez un humain, par exemple, je m'attendrais à trouver des cellules qui réagissent à des caractères typographiques dans une région comme IT, après un apprentissage pour reconnaître chacune des lettres. Mais après lui avoir appris à lire des mots entiers, je m'attendrais à trouver des cellules qui réagissent aux lettres dans différentes parties de V4, en plus de IT. Des résultats identiques devraient être obtenus avec d'autres espèces, d'autres régions et d'autres stimuli. Une autre conséquence de ce processus d'apprentissage est que les endroits où les souvenirs se produisent et ceux où les erreurs sont détectées devraient se déplacer. C'est-à-dire que les sensations des patterns très bien appris devraient se propager sur une distance moindre vers le haut de la hiérarchie. Ceci devrait être détectable par les techniques d'imagerie. Nous devrions aussi pouvoir détecter une modification du temps de réaction à certains stimuli, car les inputs n'auraient pas à voyager très loin dans le cortex pour être reconnus et rappelés.

**Prédiction 11**

Les représentations invariantes devraient être trouvées dans toutes les aires corticales.

Il est bien connu qu'il existe des cellules qui réagissent à des inputs hautement sélectifs dont les détails sont invariants. Des cellules réagissant aux visages, aux mains ou à Bill Clinton ont été observées. Le modèle de mémoire-prédiction prédit que toutes les régions du cortex doivent former des représentations invariantes. Ces dernières doivent refléter toutes les modalités sensorielles en dessous d'une région du cortex. Par exemple, si j'avais une cellule « bill-clintonienne » dans le cortex auditif, elle serait excitée chaque fois que j'entends le nom de Bill Clinton. Je m'attendrais ensuite à trouver des cellules dans les aires associatives qui reçoivent à la fois l'input visuel et auditif, et qui réagissent soit à la vue, soit à l'audition des mots « Bill Clinton ». Nous trouverions des représentations invariantes dans toutes les modalités sensorielles et même dans le cortex moteur. Là, les cellules représenteraient des séquences motrices complexes. Plus nous remonterions dans la hiérarchie motrice, plus les représentations seraient complexes et invariantes (des études récentes ont permis de découvrir des cellules qui activent de complexes mouvements de la main à la bouche chez le singe). Ce ne sont pas des prédictions toutes nouvelles. La plupart des chercheurs admettent l'idée générale selon laquelle les représentations invariantes sont formées dans de nombreux endroits un peu partout dans le cortex. Toutefois, bien que j'aie présenté ceci comme un fait, la démonstration n'en a jamais été faite. Le modèle de mémoire-prédiction prédit que nous trouverons de telles cellules dans toutes les parties du cortex.

Les prédictions qui précèdent sont quelques moyens par lesquels le modèle évoqué dans ce livre peut être testé. Je suis sûr qu'il en existe d'autres. Il n'est cependant pas possible de prouver qu'une théorie est correcte, mais seulement qu'elle est erronée.

Donc, même si toutes les prédictions proposées dans cette annexe étaient avérées, cela ne serait pas la preuve que l'hypothèse de la mémoire-prédiction est correcte, mais ce serait toutefois une évidence militant puissamment en faveur de cette théorie. L'inverse est vrai aussi. Si certaines des prédictions se révélaient erronées, cela n'invaliderait pas forcément la totalité de la thèse. Pour certaines des prédictions, il existe d'autres façons d'obtenir les comportements requis. Par exemple, des noms de séquences peuvent être créés par d'autres moyens. Cette annexe a pour seul but de montrer que le modèle conduit à plusieurs prédictions, et peut de ce fait être testé. Concevoir des expériences est stimulant et demanderait bien plus d'exposés qu'il est approprié d'en faire figurer dans cet ouvrage. Ce serait génial si nous pouvions trouver des moyens de tester cette théorie par des techniques d'imagerie comme l'IRM. Il existe beaucoup de laboratoires d'imagerie et ces expériences peuvent être effectuées relativement vite, comparé à l'enregistrement direct depuis les cellules.

# BIBLIOGRAPHIE

**La plupart** des ouvrages et publications scientifiques se terminent par d'interminables bibliographies qui servent autant à présenter les sources des différentes contributions qu'à aider le lecteur à en savoir plus. Comme ce livre s'adresse à une diversité de lecteurs, y compris ceux qui n'ont pas de notions de neurobiologie, j'ai évité de le rédiger dans un style académique. Cette bibliographie est principalement destinée à aider le lecteur néophyte désireux d'approfondir ses connaissances. Je n'ai pas dressé la liste de toutes les publications sur le sujet, ni tenté de citer tous les chercheurs qui ont fait des découvertes fondamentales en ce domaine. J'ai préféré choisir des titres qui, je crois, intéresseront le lecteur. Il trouvera parmi eux quelques ouvrages qui s'adressent plutôt aux spécialistes. Le Web est aussi une bonne source d'informations. Une bibliographie plus détaillée peut être consultée sur le site de la version originale de ce livre : **www.OnIntelligence.org**.

Vous ne trouverez hélas que quelques références aux théories d'ensemble du cerveau car, à l'heure où j'écrivais le prologue, très peu d'ouvrages avaient été consacrés à ce sujet, et moins encore aux propositions exposées dans ce livre.

#### Histoire de l'intelligence artificielle et des réseaux neuronaux

Baumgartner, Peter and Sabine Payr, eds. *Speaking Minds : Interviews with Twenty Eminent Cognitive Scientists* (Princeton, N.J. : Princeton University Press, 1995).

Ce livre contient d'intéressantes conversations avec bon nombre des plus éminents théoriciens dans les domaines de l'intelligence artificielle, des réseaux neuronaux et des sciences cognitives. C'est un résumé agréable à lire de l'histoire récente de l'intelligence et de ses théories.

Dreyfus, Hubert L. *Intelligence artificielle. Mythes et limites* (Paris, Flammarion, 1984).

Une critique corrosive de l'intelligence artificielle, publiée d'abord sous le titre *What Computers Can't Do* (ce que les ordinateurs ne peuvent pas faire) et rééditée des années après sous un autre titre. C'est une histoire approfondie de l'intelligence artificielle écrite par l'un de ses plus durs critiques.

Anderson, James A. and Edward Rosenfeld, eds. *Neurocomputing, Foundations of Research* (Cambridge, Ass. : MIT Press, 1988).

Cet imposant ouvrage est une compilation annotée d'importantes publications sur la théorie des réseaux neuronaux et du cerveau s'étendant des années 1890 à 1987, présentées dans l'ordre chronologique. Il contient des articles de W. S. McCulloch et W. Pitts, Donald Hebbs, Steve Grossberg et beaucoup d'autres. Chaque article est précédé d'une introduction. C'est un moyen agréable d'accéder aux plus importantes publications dans ce domaine.

Searle, J. R. « Esprits, cerveaux et programmes ». Contribution parue dans *Vues de l'esprit. Fantaisies et réflexions sur l'être et l'âme* (Paris, InterEditions, 1987).

On y trouve la célèbre argumentation de la Chambre chinoise opposée à la computation en tant que modèle pour l'esprit. Vous trouverez sur le Web de nombreuses descriptions et discussions concernant les expériences sur la pensée proposées par John Searle.

Turing, A. M. « Computing Machinery and Intelligence », in *Mind*, vol. 59 (1950), pp. 433-60.

Présentation du célèbre test de Turing pour détecter la manifestation de l'intelligence. A ce sujet aussi, de nombreuses références et discussions peuvent être trouvées sur le Web.

Palm, Günther. Neural Assemblies : An Alternative Approach to Artificial Intelligence (New York : Springer Verlag, 1982).

Pour comprendre le fonctionnement du cortex et comment il stocke des séquences de patterns, il est préférable de s'être familiarisé avec la notion de mémoires auto-associatives. Bien qu'on ait écrit beaucoup sur ce sujet, je n'ai trouvé aucun ouvrage qui en présente un résumé limpide facilement accessible. Palm est l'un des pionniers dans ce domaine. Son ouvrage est difficile à trouver et pas facile à lire, mais il couvre toutes les bases des mémoires auto-associatives, y compris la mémoire de séquence.

**Néocortex et neurobiologie générale**

Les ouvrages qui suivent sont recommandés à tous ceux qui désirent en savoir plus sur la neurobiologie et le néocortex.

Crick, Francis H. C. « Réflexions sur le cerveau », *Pour la Science*, 25 (nov. 1979).

C'est la traduction de l'article paru dans *Scientic American* (vol. 241, sept. 1979) qui avait éveillé mon intérêt pour le cer-

veau. Bien qu'il date de plus d'un quart de siècle, il est toujours aussi exaltant.

Koch, Christof. *Quest for Consciousness : A Neurobiological Approach* (Denver, Colo. : Roberts and Co., 2004).

Ce sont plusieurs ouvrages généralistes consacrés au cerveau, et publiés chaque année. Celui de Christof Koch, sur la conscience, aborde aussi le cerveau, la neuroanatomie, la neurophysiologie et la conscience psychologique. C'est un excellent ouvrage d'initiation.

Mountcastle, Vernon B. *Perceptual Neuroscience : The Cerebral Cortex* (Cambridge, Mass. : Harvard University Press, 1998).

Un ouvrage remarquable, consacré à tout ce qui se rapporte au néocortex. Il est bien écrit, bien présenté et, bien que technique, agréable à lire. C'est une des meilleures introductions au néocortex.

Kandel, Eric R., James H. Schwartz, Thomas M. Jessel, eds. *Principles of Neural Science*, 4th ed. (New York : McGraw-Hill, 2000).

Cette encyclopédie en un seul volume fait le tour de tout ce qui a trait à la neurologie. Elle fournit des explications détaillées sur toutes les parties du système nerveux, y compris les neurones, les organes sensoriels et les neurotransmetteurs.

Shepherd, Gordon M., ed. *The Synaptic Organization of the Brain*, 5th ed. (New York : Oxford University Press, 2004).

Ce livre m'a bien aidé, bien que j'aie préféré les éditions précédentes rédigées par un seul auteur. Il contient des données techniques sur toutes les parties du cerveau, notamment les synapses. C'est un ouvrage de référence.

Koch, Christof and Joel L. Davis, eds. *Large-scale Neuronal Theories of the Brain* (Cambridge, Mass. : MIT Press, 1994).

Il existe peu de littérature sur les théories globales du cerveau. Ce livre est une compilation d'articles sur les théories neuronales. Il donne un aperçu des diverses approches pour comprendre le fonctionnement global du cerveau. Vous trouverez tout au long de ce livre les prémisses du cadre de mémoire-prédiction.

Braitenberg, Valentino and Almut Schüz. *Cortex : Statistics and Geometry of Neuronal Connectivity*, 2nd ed. (New York : Springer Verlag, 1998).

Ce livre décrit les propriétés statistiques du cerveau de la souris. Je sais que ce n'est pas très enthousiasmant, mais c'est un livre original et utile. Il raconte le cortex en chiffres.

#### Quelques articles sur la neurobiologie

Les articles qui suivent sont les sources originales de certains des importants concepts décrits dans ce livre. La plupart ne peuvent être trouvés qu'en bibliothèque ou sur le Web.

Mountcastle, Vernon B. « An Organizing Principle for Cerebral Function : The Unit Model and the Distributed System », in Gerald M. Edelman and Vernon B. Mountcastle, eds., *The Mindful Brain* (Cambridge, Mass. : MIT Press, 1978).

C'est dans cet article que j'ai découvert les hypothèses de Mountcastle sur le fonctionnement du cortex à partir d'un principe commun. Mountcastle soutient aussi que la colonne corticale est l'unité de base de la computation. Ces idées sont à la fois le fondement et la source d'inspiration de la théorie proposée dans ce livre.

Creutzfeldt, Otto D. « Generality of the Functional Structure of the Neocortex », *Naturwissenschaften*, vol. 64 (1977), pp. 507-17.

C'est après avoir écrit *Intelligence* que j'ai pris connaissance de cette publication qui, à l'instar des écrits de Vernon Mountcastle, milite en faveur d'un algorithme cortical. Elle parut un peu avant celle de Mountcastle et la complète parfaitement.

Felleman, D. J. and D. C. Van Essen. « Distributed Hierarchical Processing in the Primate Cerebral Cortex », *Cerebral Cortex*, vol. 1 (january/february 1991), pp. 1-47.

C'est l'article, devenu un classique, qui décrit l'organisation hiérarchique du cortex visuel. Le cadre de mémoire-prédiction est fondé sur l'hypothèse que c'est non seulement le système visuel mais la totalité du néocortex qui est hiérarchiquement structurée.

Sherman, S. M. and R. W. Guillery. « The Role of the Thalamus in the Flow of Information to the Cortex », *Philosophical Transactions of the Royal Society of London*, vol. 357, no. 1428 (2002), pp. 1696-708.

L'article fournit une vue d'ensemble de l'organisation thalamique et expose l'hypothèse de Sherman-Guillery selon laquelle le thalamus sert à répartir le flux d'information entre les aires corticales. Cette idée est développée au Chapitre 6, dans la section intitulée « Une voie alternative pour monter dans la hiérarchie ».

Rao, R. P. and D. H. Ballard. « Predictive Coding in the Visual Cortex : A Functional Interpretation of Some Extra-Classical Receotive-field Effects », *Nature Neuscience*, vol. 2, no. 1 (1999), pp. 79-87.

J'ai ajouté cet article comme exemple d'une recherche récente qui traite de la prédiction et des hiérarchies. La publication de Rao et Ballard présente un modèle de biofeedback dans les hiérarchies corticales, dans lequel les neurones des aires supérieures tentent de prédire des patterns d'activité dans les aires inférieures.

Guillery, R. W. « Branching Thalamic Afferents Link Action and Perception », *Journal of Neurophysiology*, vol. 90 (2003), pp. 539-48.

Young M. P. « The Organization of Neural Systems in the Primate Cerebral Cortex », *Proceedings of the Royal Society : Biological Sciences*, vol. 252 (1993), pp. 13-18.

Ces deux articles bien écrits démontrent que de toute évidence le comportement moteur et la perception sensorielle sont intimement liés et font partie du même processus. Guillery soutient que les aires corticales sensorielles jouent un rôle dans le comportement moteur, et Young montre que le cortex moteur et le cortex somatosensoriel sont si étroitement liés qu'ils devraient être considérés comme un seul et même système. Ces idées sont brièvement évoquées au Chapitre 6.

# REMERCIEMENTS

**Chaque** fois que quelqu'un me demande « De quoi vivez-vous ? », je ne sais jamais que répondre. A vrai dire, je ne fais pas grand-chose. Mais je me suis entouré de gens qui semblent en faire beaucoup. Ma contribution consiste à les pousser un tout petit peu de temps en temps, et au besoin, leur indiquer une nouvelle voie. Le succès que j'ai rencontré dans ma carrière, je le dois avant tout au travail acharné et à l'intelligence de mes collègues.

J'ai eu le privilège de rencontrer beaucoup de scientifiques. Presque tous m'ont enseigné quelque chose, et de ce fait, presque tous ont contribué à développer l'idée qui sous-tend ce livre. Je les remercie tous, bien que je ne puisse en mentionner ici que quelques-uns. Bruno Olshausen, qui travaille à la fois au Redwood Neuroscience Institute (RNI) et à l'université de Californie, à Davis, est une encyclopédie vivante de la neurobiologie. Il a sans relâche signalé des faiblesses et toujours suggéré comment les rectifier, ce qui est sans doute l'une des attitudes les plus précieuses. Bill Softky, lui aussi du RNI, fut le

premier à m'apprendre ce qu'est la réduction temporelle, dans la hiérarchie corticale, ainsi que les propriétés des dendrites minces. Rick Granger, de l'université de Californie à Irvine, m'a fait mieux connaître la mémoire de séquence et le rôle que le thalamus peut jouer. Bob Knight, de l'université de Californie à Berkeley, et Christof Koch, du California Institute of Technology, ont joué un rôle prépondérant dans la formation du Redwood Neuroscience Institute et dans beaucoup d'autres domaines scientifiques. Toute l'équipe du RNI m'a stimulé et incité à affiner mes idées. De nombreuses hypothèses de ce livre sont le résultat direct des rencontres et des discussions au RNI. Merci à tous.

Donna Dubinsky et Ed Colligan sont mes partenaires en affaires depuis une douzaine d'années. C'est grâce à leur dur travail et à leur assistance que j'ai réussi à être un créateur d'entreprises tout en travaillant à temps partiel sur la théorie du cerveau, un arrangement qui n'est pas courant. Donna avait pour habitude de dire que l'un de ses objectifs consistait à faire tourner la boutique afin que je puisse libérer du temps pour ma théorie du cerveau. Sans Donna et Ed, ce livre n'aurait pas existé.

Je n'aurais jamais pu écrire *Intelligence* sans de sérieuses aides. Jim Levine, mon agent, a cru dans ce livre avant même que je sache ce que j'y mettrais. N'écrivez jamais un livre sans un agent comme Jim. Il m'a présenté à Sandra Blakeslee, mon coauteur. Je tenais à ce que ce livre soit accessible à un large public; à cet égard, Sandra fut précieuse. Si des pages sont encore ardues à lire, j'en suis le seul responsable. Matthew Blakeslee, le fils de Sandra, qui écrit aussi des ouvrages scientifiques, a fourni plusieurs des exemples de ce livre et suggéré le terme de *cadre de mémoire-prédiction*. J'ai beaucoup apprécié de pouvoir travailler avec le personnel des éditions Henry Holt. Je tiens à remercier tout particulièrement John Sterling, président d'Henry Holt et éditeur. Je ne l'ai rencontré qu'une seule fois et nous avons conversé à plusieurs reprises au téléphone. Il ne lui en pas fallut davantage pour savoir exactement comment le livre devait être structuré. Il

comprit immédiatement les problèmes auxquels je serais confronté en proposant une théorie de l'intelligence, et suggéra la façon dont le livre devait être rédigé et positionné.

Je veux aussi remercier mes filles Anne et Kate pour ne s'être jamais plaintes pendant que leur papa passait de nombreux week-ends rivé au clavier de l'ordinateur. Et je tiens aussi à remercier mon épouse Janet. Vivre avec moi n'est pas toujours facile. Je l'aime plus que les encéphales.

# INDEX


**A**

ADN 208
Aire 130
  associative 59, 136, 139
  de Broca 57
  IT 134
  sensorielle 59
  somatosensorielle 59
  TM 59
  V1 133, 143
  V4 59
  visuelle 59
Ame 96, 231
Animaux 207
Antéropropagation 135
Apprentissage 192
  hebbien 192
Asimov (Isaac) 43
Asperger (syndrome d'~) 263
Auto-associatif 40, 90, 169
Autoréplicante (machine) 250
Aveugle 68
Axone 61, 70, 166, 171
  myéline 169

**B**

Bach y Rita (Paul) 76
Bayes (Thomas) 108
Bekesy (Georg von) 73
Bit 92
Blocks World 26
Brahe (Tycho) 225
Braille 68
Broca (aire de ~) 57

**C**

Cadre de mémoire-prédiction 124, 282
Caméra de sécurité 256
Capacité 261
Cellule
  gliale 220
  inhibitrice 173
  nom 278, 279
Cerveau 53
  archaïque 117
  limbique 117
  primitif 117
  reptilien 117

Cervelet 54, 197
Chambre chinoise 28, 124
Champ réceptif 133
Classification 158
    de forme 160
    de patterns 160
Clément (Jean-Baptiste) 110
Clinton (Bill) 130, 284
Cochlée 73
Comportement 30, 116
Comportementalisme 25
Connectivité 246
Connexionniste 34
Conscience 225
Copernic (Nicolas) 44
Corps calleux 220
Cortex 55, 116
    apprentissage 192
    colonnes 163
    couches 55, 162
    embryon 164
    fonctionnement 127
    région 162
CPU 22
Créativité 214
Crick (Francis) 18
Cube de Necker 281

**D**

Darwin (Charles) 44, 64
Dauphin 123
Deep Blue 27
Dendrite 61, 166

**E**

Echecs 27, 47, 234
Einstein (Albert) 63, 220, 262
El Niño 267
Elan vital 228
Ellenby (John) 32
ENIAC 248
Ensemble 150
Erdös (Paul) 14
Espace (perception de l'~) 236
Esprit 231
Essen (David van) 59
Ethique 248

**F**

Feedback 35, 135, 189
    replié 183, 234
Feedforward 135
Felleman (David) 59
Fixation 132
Flux 186
Fonctionnalisme 47
Formes (théorie des ~) 96
Fried (Itzhak) 130

**G**

Ganglions basaux 197
Garland (Judy) 98
Goldberg (Rube) 49
Graffiti 39, 223
Grid Systems 31
GridTask 31
Grossberg (Stephen) 108, 183, 234
Guillery (Ray) 201
Gyrus fusiforme 57

**H**

Hebb (Donald O.) 192
Hebbien (apprentissage) 192
Hiérarchie 58, 130, 139, 143, 146, 148
    flux 186
Hippocampe 55, 197
Hoff (Ted) 19

**I**

Imagerie fonctionnelle 42
Imagination 183, 233
Input 25
Intel 17
Intelligence artificielle 21

IRM 66
IRMf 281

**J**
Joy (Bill) 249

**K**
Kandinsky (Wassily) 230
Kasparov (Gary) 27
Keller (Helen) 77
Kepler (Johannes) 224
Kluge 49
Koch (Christof) 130, 225
Kreiman (Gabriel) 130
Kurzweil (Ray) 250

**L**
Labyrinthe 119
Langage 211
Langue 76
Lincoln (Abraham) 154
Loi de Moore 253

**M**
Machine de Turing 22
Mackay (D. M.) 108
Madeleine de Proust 91
Matière grise 169, 246
McCulloch (Warren) 24
Mémoire 81
  auto-associative 40, 90, 169
  capacité 244
  déclarative 229
  intelligente 245
Météorologie 266
Microprocesseur 240
MIT 20
Modèle 12, 79, 114, 148, 235
Monde 74, 161
  des blocs 26
  modèle 12, 79, 114, 148, 235
  virtuel 269
Moore (Gordon) 19, 253
Mountcastle (Vernon) 63, 165
Muet 68
Mumford (David) 108
Myéline 169

**N**
Nanorobot 250
Néocortex 53, 117
Nestor 38
NetTalk 37
Neurone 56, 61, 89, 190
  portes logiques 24
  pyramidal 62
Nom 152, 153, 175, 176, 278, 279

**O**
Octet 92
Olshausen (Bruno) 198
Ouïe 73
Output 25

**P**
Paléocortex 53
Papille 112
Parallélisme 82
Pattern 33, 36, 71, 132
  changeant 176
  classification 160
  constant 153, 176
  d'entrée 75
  spatial 71, 132
  temporel 41, 71, 132
Pitts (Walter) 24
Plantes 209
Platon 95
  théorie des Formes 96
Porte faussée 105
Porte logique 24
Potentiel d'action 70, 131
Prédictibilité 151
Prédiction 104, 106, 165, 174, 214

invariante 180
probabiliste 110
Proprioceptif (système ~) 74
Proust (Marcel) 91

**Q**
Quale 228, 230

**R**
Rao (Rajesh) 108
Rapidité 260
Réalité 151, 234
Reconnaissance vocale 254
Redwood Neuroscience Institute 10
Région corticale 64, 136, 143, 145, 152, 162
classifications 160
fonctionnement 172
Règle des cent étapes 82
Relativité 63
Remplissage 111
Réplicabilité 263
Représentation invariante 86, 92, 130, 284
Réseau bayésien 108
Rétroaction 35
Rétropropagation 36, 135
RNI 10
Robot 241
Robotique 259

**S**
Saccade 113, 132
Sagan (Carl) 273
Scissure centrale 122
Searle (John) 28
Segment 143, 148
Sens 265
Séquence 150
formation 158
nom 152, 153, 175, 176
Shakespeare (William) 217
Sherman (Murray) 201
Solide platonicien 224
Stéréotype 236
Substitution sensorielle 76
Synapse 61, 89, 174, 190
Synesthésie 230
Système
expert 26
sensoriel 265

**T**
Tache aveugle 112
Test de Turing 23, 270
Thalamus 55, 171, 201
Théorie des Formes 96
Transistor 37, 240
Transports 257
Turing
Alan 22, 124
machine 22
test 23, 270

**V**
Vue 71
saccades 72

**W**
Weihenmayer (Erik) 76

**Y**
Yeld 245

**Z**
Zador (Tony) 276

# À PROPOS DES AUTEURS

JEFF HAWKINS est l'un des créateurs d'entreprises et concepteur d'ordinateurs les plus en vue de la Silicon Valley. Fondateur de Palm Computing et de Handspring, il est aussi à l'origine du Redwood Neuroscience Institute, chargé de promouvoir la recherche sur la mémoire et la cognition. Il est membre du comité scientifique du Cold Spring Harbor Laboratory. Il vit dans le nord de la Californie.

SANDRA BLAKESLEE écrit des articles scientifiques et médicaux pour *The New York Times* depuis plus de trente ans. Elle est coauteur, avec V. S. Ramachandran et Judith Wallerstein, de *Phantoms in the Brain*, un best-seller consacré à la psychologie et au mariage. Elle vit à Santa Fe, dans le Nouveau-Mexique.