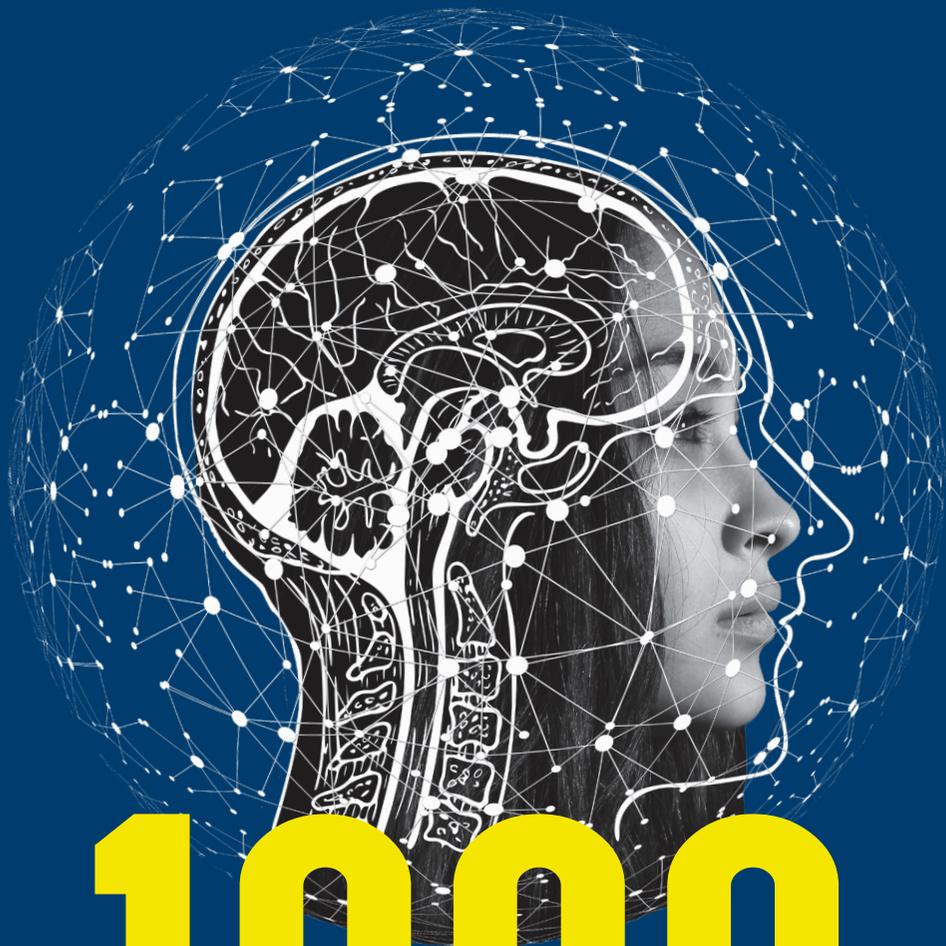


«Remarquable» – Henry Markram, Human Brain Project

JEFF HAWKINS



1000 CERVEAUX

Une nouvelle compréhension du cerveau
et de l'intelligence artificielle

1000
CERVEAUX

JEFF HAWKINS

1000 CERVEAUX

Une nouvelle compréhension du cerveau
et de l'intelligence artificielle

Préface de Mackenzie Weygandt Mathis

Avant-propos de Richard Dawkins

Traduit de l'anglais (États-Unis)
par Anatole Muchnik

quanto

Version originale: *A thousand Brains – A New Theory of Intelligence*

First published by Basic Books

ISBN : 978-1-5416-7581-0

Copyright © 2021 by Jeffrey C. Hawkins

Avant-propos : copyright © 2021 by Richard Dawkins

All rights reserved.

Traduction française : Anatole Muchnik

Éditorial : Sylvain Collette

Révision scientifique : Yohann Thenaisie

Maquette et mise en page : Kim Nanette

Illustration de couverture : Gerd Altmann, Pixabay

Première édition française 2023

© Presses polytechniques et universitaires romandes, Lausanne

Quanto est un label des Presses polytechniques et universitaires romandes

ISBN 978-2-88915- 520-0

Les Presses polytechniques et universitaires romandes bénéficient d'un soutien structurel de l'Office fédéral de la culture pour les années 2021-2024.

Tous droits réservés

Reproduction, même partielle, sous quelque forme ou sur quelque support que ce soit, interdite sans l'accord écrit de l'éditeur

Imprimé en France

Table des matières

Préface	7
Avant-propos	11
Première partie	
UNE NOUVELLE VISION DU CERVEAU	19
1 Cerveau ancien-cerveau nouveau	31
2 La grande idée de Vernon Mountcastle	41
3 Vous avez un modèle du monde dans la tête	49
4 Le cerveau livre ses secrets	59
5 Des cartes dans le cerveau	77
6 Concepts, langage et pensée de haut niveau	91
7 Théorie de l'intelligence des mille cerveaux	111
Deuxième partie	
L'INTELLIGENCE MACHINE	133
8 Pourquoi il n'y a pas de « I » dans l'IA	139
9 Quand les machines sont conscientes	157
10 L'avenir de l'intelligence machine	167
11 Les risques existentiels que pose l'intelligence machine	185
Troisième partie	
L'INTELLIGENCE HUMAINE	197
12 Fausses croyances	201
13 Les risques existentiels de l'intelligence humaine	215
14 La fusion du cerveau et de la machine	229
15 La planification successorale de l'humanité	239
16 Les gènes ou le savoir	255
Dernières réflexions	275
Suggestions de lectures	281
Remerciements	289
Crédit des illustrations	293

Première partie

**UNE NOUVELLE
VISION
DU CERVEAU**

Les cellules dans votre tête sont en train de lire ces lignes. Songez un peu à quel point c'est remarquable. Une cellule est une chose simple. Une cellule seule ne peut ni lire, ni penser, ni vraiment faire quoi que ce soit. Mais aussitôt qu'on en réunit suffisamment pour composer un cerveau, elles ne sont pas seulement capables de lire un livre, mais de l'écrire. Elles conçoivent des bâtiments, inventent des technologies et déchiffrent les mystères de l'Univers. La façon par laquelle un cerveau constitué de simples cellules crée de l'intelligence est une question extrêmement intéressante, qui demeure mystérieuse.

Comprendre le fonctionnement du cerveau est l'un des grands défis qui se posent à l'humanité. C'est une quête qui a récemment suscité des dizaines d'initiatives nationales et internationales, comme le Human Brain Project européen ou l'initiative internationale BRAIN. Des dizaines de milliers de spécialistes des neurosciences représentant des dizaines de spécialités dans à peu près tous les pays du monde s'efforcent de comprendre le cerveau. S'ils étudient le cerveau de différents animaux et posent des questions de tout type, leur objectif ultime est de découvrir comment le cerveau humain engendre l'intelligence des humains.

Peut-être avez-vous trouvé surprenante mon affirmation que le cerveau humain demeure un mystère. Chaque année, des découvertes sont annoncées à son sujet, de nouveaux livres paraissent et les chercheurs de domaines connexes comme l'intelligence artificielle affirment que leur création approche l'intelligence de la souris, par exemple, ou du chat. Il serait facile d'en conclure que la recherche se fait une idée assez précise de la façon dont fonctionne le cerveau. Mais si vous leur posez la question, les chercheurs en neurosciences diront presque tous qu'on est encore dans les ténèbres. Malgré la quantité phénoménale de faits isolés découverts à propos du cerveau, nous ne comprenons que mal son fonctionnement global.

En 1979, déjà rendu célèbre par ses travaux sur l'ADN, Francis Crick a écrit un article sur l'état des sciences du cerveau intitulé « Thinking About the Brain » (Penser le cerveau). Il y décrit l'abondance de faits récoltés par les chercheurs sur le cerveau, mais n'en conclut pas moins que « malgré l'accumulation constante de connaissances détaillées, le fonctionnement du cerveau humain

demeure profondément mystérieux. » Et de poursuivre : « Ce qui manque de façon criante, c'est un cadre conceptuel général au sein duquel interpréter ces résultats. »

Crick avait constaté qu'après des décennies de récolte de données sur le cerveau, la communauté scientifique connaissait un grand nombre de faits. Mais nul n'avait trouvé comment assembler ces faits pour composer quelque chose de significatif. Le cerveau était un puzzle géant de milliers de pièces étalées sous nos yeux, mais dont on ne parvenait à rien tirer de cohérent. On était même loin d'imaginer à quoi pourrait ressembler une solution. Selon Crick, le mystère du cerveau n'était pas dû à quelque manque de données, mais à notre incapacité d'assembler les pièces en notre possession. Quarante années ont passé depuis cet article de Crick, et malgré les nombreuses découvertes importantes accomplies depuis, dont j'évoquerai certaines plus loin, son observation reste globalement vraie. La façon dont l'intelligence émerge des cellules dans notre tête demeure profondément mystérieuse. Et alors que s'accumulent les nouvelles pièces du puzzle, année après année, on a parfois le sentiment de s'éloigner de la compréhension du cerveau au lieu de s'en approcher.

J'ai lu cet article de Crick quand j'étais jeune, et il m'a profondément inspiré. Persuadé depuis que le mystère du cerveau serait résolu de mon vivant, je n'ai cessé de poursuivre cet objectif. Voici quinze ans que je dirige dans la Silicon Valley une équipe de chercheurs qui étudie une partie du cerveau nommée néocortex. Le néocortex occupe environ 70 % du volume d'un cerveau humain, il est responsable de tout ce qu'on associe à l'intelligence ; cela va de notre sens de la vision, du toucher et de l'ouïe à la pensée abstraite qui constitue les mathématiques ou la philosophie en passant par le langage sous toutes ses formes. Notre entreprise vise à comprendre le fonctionnement du néocortex de manière suffisamment détaillée pour expliquer la biologie du cerveau et bâtir des machines opérant selon les mêmes principes.

Début 2016, nos travaux se sont considérablement accélérés. Notre entendement a fait un bond. Nous avons saisi que nous, et avec nous l'ensemble de la recherche, étions passés à côté d'un

ingrédient essentiel. Cette prise de conscience nous a permis de voir comment assembler les pièces du puzzle. Autrement dit, je pense que nous avons découvert le cadre conceptuel dont parlait Crick, qui n'explique pas seulement les principes de base du fonctionnement du néocortex, mais qui ouvre toute une nouvelle façon de se représenter l'intelligence. Nous ne possédons pas encore de théorie complète du cerveau – et nous en sommes loin. En général, au moment d'aborder un nouveau champ de la science, on s'arme d'un cadre théorique pour ne se consacrer aux détails qu'ensuite. Le plus bel exemple de ceci nous est peut-être fourni par la théorie de l'évolution. Darwin a commencé par proposer une manière nouvelle et audacieuse de considérer l'origine des espèces, mais les détails, tels que le rôle des gènes et de l'ADN, ne seraient connus que bien plus tard.

Pour être intelligent, le cerveau doit apprendre beaucoup de choses au sujet du monde. Pas seulement celles que l'on apprend à l'école, mais des choses élémentaires, comme l'aspect, le son ou la sensation que nous procurent les choses du quotidien. Il faut apprendre comment se comportent les objets, de l'ouverture à la fermeture d'une porte à ce que font les applis de nos smartphones quand on touche l'écran. Il faut apprendre où se trouvent toutes les choses du monde, que ce soit l'endroit de la maison où nous conservons nos effets personnels ou celui où se situent le bureau de poste et la bibliothèque de notre ville. Et puis, bien entendu, nous acquérons aussi des concepts plus élevés, comme la « compassion » ou l'« État ». Par-dessus tout cela, chacun de nous apprend encore le sens de dizaines de milliers de mots. Nous possédons tous une quantité phénoménale de connaissances à propos du monde. Certaines de nos facultés élémentaires, comme manger ou éviter la douleur, sont déterminées par nos gènes. Mais l'essentiel de ce que nous savons du monde est acquis.

La recherche nous dit que le cerveau acquiert un modèle du monde. Le terme « modèle » suppose que notre savoir n'est pas entreposé sous forme d'un empilement de faits, mais qu'il est organisé d'une façon qui est le reflet de la structure du monde et de tout ce qu'il contient. Pour savoir ce qu'est un vélo, par exemple, on ne

retient pas une liste de faits concernant les vélos. Notre cerveau crée un modèle de vélo qui comprend ses différentes pièces, leur disposition les unes par rapport aux autres et la façon dont elles se meuvent et opèrent ensemble. Pour reconnaître une chose, il faut d'abord apprendre à quoi elle ressemble et la sensation qu'elle procure. Et pour accomplir quelque chose, il faut apprendre comment se comportent généralement les choses du monde lorsqu'on interagit avec elles. L'intelligence est intimement liée au modèle du monde que contient notre cerveau ; par conséquent, pour comprendre comment le cerveau crée de l'intelligence, il nous faut découvrir comment cet organe constitué de simples cellules acquiert un modèle du monde et de tout ce qui s'y trouve.

Ce que nous avons découvert en 2016 permet d'expliquer cette acquisition. Nous en sommes venus à comprendre que le néocortex entrepose tout ce que nous connaissons, notre savoir tout entier, à l'aide de ce qu'on appelle des référentiels. J'y reviendrai plus loin, mais établissons pour l'instant une analogie avec un plan sur papier. Un plan est un type de modèle : le plan d'une ville est un modèle de cette ville et les lignes du quadrillage, comme la latitude et la longitude, sont un certain référentiel. Ce quadrillage, ce référentiel, procure à la carte sa structure. Le référentiel situe les choses les unes par rapport aux autres, il vous dit par exemple comment remplir l'objectif consistant à se rendre d'un point à un autre. Nous avons alors compris que le cerveau se fabrique son modèle du monde en usant de référentiels similaires à ceux d'une carte. Il n'y a pas qu'un référentiel, mais des centaines de milliers. Pour tout dire, nous sommes en train de découvrir que la plupart des cellules du néocortex sont dédiées à la création et à la manipulation des référentiels qu'utilisera le cerveau pour planifier et réfléchir.

Sous ce nouveau jour, nous avons vu poindre les réponses à certaines des grandes questions que se posent les neurosciences. Comment nos diverses entrées sensorielles s'unissent-elles pour donner lieu à une expérience unique ? Que se passe-t-il quand on pense ? Comment deux personnes peuvent-elles aboutir à des croyances différentes à partir des mêmes observations ? Et pourquoi sommes-nous dotés d'une conscience de soi ?

Ce livre décrit ces découvertes et leurs implications pour l'avenir. L'essentiel de son contenu a paru dans des revues scientifiques dont vous trouverez les références en fin d'ouvrage. Mais les articles scientifiques n'ont pas pour vocation d'expliquer des théories de grande envergure, et certainement pas d'une manière qui soit compréhensible aux profanes.

J'ai organisé ce livre en trois parties. La première est une description de notre théorie des référentiels, que nous appelons la « théorie des mille cerveaux ». Étant donné qu'elle se fonde en partie sur la logique de déduction, je vous emmènerai pas à pas sur la voie qui nous a conduits à nos conclusions. Je vous fournirai aussi un minimum de contexte historique pour que vous ayez une idée de la façon dont notre théorie s'inscrit dans l'histoire de l'étude du cerveau. J'espère que vers la fin de la première partie, vous comprendrez un peu mieux ce qu'il se passe dans votre tête quand vous pensez et agissez dans le monde, et ce qu'être intelligent veut dire.

La deuxième partie traite de l'intelligence des machines. Le XXI^e siècle sera transformé par les machines intelligentes comme le XX^e l'a été par l'ordinateur. La théorie des mille cerveaux explique pourquoi l'IA d'aujourd'hui n'est pas encore de l'intelligence et comment rendre les machines vraiment intelligentes. Je décrirai à quoi ressembleront ces machines de demain et quels usages nous pourrions en faire. J'expliquerai pourquoi certaines d'entre elles seront conscientes et ce qui mérite d'être fait à ce sujet, si tant est qu'il faille absolument faire quelque chose. Enfin, beaucoup s'inquiètent aujourd'hui de voir les machines intelligentes constituer un risque existentiel, du fait que nous soyons sur le point de créer une technologie qui conduira l'humanité à sa perte. Je ne le crois pas. Ce que nous avons découvert explique pourquoi l'intelligence des machines est en soi inoffensive. Il n'en s'agit pas moins d'une technologie très puissante, dont le risque réside en vérité dans l'usage qu'en feront les humains.

Dans la troisième partie de ce livre, j'observerai la condition humaine sous la perspective du cerveau et de l'intelligence. Le modèle du monde que contient notre cerveau comporte un modèle de notre soi. Et cela conduit à cette étrange vérité que vos perceptions et les miennes, à chaque instant, sont une simulation du

monde, pas le monde réel. L'une des conséquences de la théorie des mille cerveaux est que nos croyances au sujet du monde peuvent être fausses. J'expliquerai comment c'est possible, pourquoi les fausses idées peuvent être tenaces et comment les fausses croyances associées aux plus primitives de nos émotions risquent de menacer notre survie à long terme.

Il sera question dans les derniers chapitres de ce qui m'apparaît comme le choix le plus fondamental qui s'offre à notre espèce. Nous avons deux façons de nous représenter nous-mêmes. Nous sommes des organismes biologiques issus de l'évolution et de la sélection naturelle. Sous cet angle, l'humain se définit par ses gènes, et la vie a pour objectif de les répliquer. Mais nous sommes en train d'émerger de notre passé purement biologique. Nous sommes devenus une espèce intelligente. La première espèce sur la Terre à connaître la taille et l'âge de l'univers. La première à savoir comment la Terre a évolué et comment nous sommes venus à l'existence. La première à développer des outils qui nous permettent d'explorer l'Univers et d'en mettre au jour les secrets. Sous cet angle-là, ce sont notre intelligence et notre savoir qui nous définissent, pas nos gènes. Face à l'avenir, le choix s'offre à nous de continuer à nous laisser mener par notre passé biologique ou d'embrasser notre intelligence récemment apparue.

Il n'est pas dit que nous puissions faire l'un et l'autre à la fois. Nous élaborons des technologies assez puissantes pour profondément altérer la planète, manipuler la biologie et, bientôt, créer des machines plus intelligentes que nous. Mais nous possédons encore les comportements qui nous ont conduits jusqu'ici. Cette combinaison constitue le véritable risque existentiel, celui qui mérite toute notre attention. En choisissant de nous définir par l'intelligence et le savoir plutôt que par nos gènes, peut-être pourrions-nous forger un avenir plus durable et aux finalités plus nobles.

Le parcours qui nous a menés à la théorie des mille cerveaux a été long et tortueux. Après des études supérieures d'ingénieur électricien, je venais de décrocher mon premier emploi chez Intel quand j'ai lu l'article de Francis Crick. L'effet qu'il a produit sur moi était tel que j'ai changé de voie et choisi de consacrer ma vie à l'étude

du cerveau. Après avoir vainement tenté d'obtenir un poste dans ce domaine chez Intel, j'ai postulé à une place d'étudiant de troisième cycle dans le laboratoire d'intelligence artificielle du MIT. (Il me semblait que pour construire des machines intelligentes, mieux valait commencer par étudier le cerveau.) Les enseignants du MIT qui m'ont fait passer les entretiens d'admission ont rejeté ma proposition de créer des machines intelligentes en se fondant sur la théorie du cerveau. On m'a expliqué que le cerveau n'était qu'un ordinateur cafouilleux et qu'il ne rimait à rien de l'étudier. Déçu, mais pas dissuadé, je me suis inscrit au programme de doctorat de l'Université de Californie, à Berkeley. J'y ai entamé mes études en janvier 1986.

Dès mon arrivée à Berkeley, je me suis adressé au D^r Frank Werblin, détenteur de la chaire du groupe de troisième cycle de neurobiologie. Il m'a demandé de décrire dans un article les recherches que je souhaitais conduire pour ma thèse. Dans ce papier, j'ai expliqué que je voulais travailler à une théorie du néocortex en abordant la question sous l'angle de la façon dont le néocortex fait des prédictions. Le professeur Werblin a fait lire mon article à plusieurs de ses confrères, qui l'ont favorablement accueilli. Reconnaissant que mes ambitions étaient admirables, que mon approche était bien fondée et que le problème auquel je m'attaquais avait un jour été parmi les plus importants de la science, il m'a dit – et ça, je ne l'avais pas vu venir – qu'il ne voyait pas de quelle façon je pourrais poursuivre mon objectif à cette époque-là. En tant qu'étudiant de troisième cycle en neurosciences, je serais appelé à m'occuper pour le compte d'un professeur des choses sur lesquelles lui-même travaillerait. Or, personne à Berkeley, ni où que ce soit à sa connaissance, ne s'intéressait de près ou de loin à ce qui m'habitait.

Développer une théorie générale de la fonction du cerveau paraissait trop ambitieux, et par conséquent trop risqué. Un étudiant qui y consacrerait cinq ans sans faire de réels progrès n'obtiendrait pas son diplôme. Le risque était réel pour l'enseignant également, qui pourrait y perdre sa titularisation. Et les agences qui distribuent le financement de la recherche auraient elles aussi trouvé cela trop risqué. Toute proposition d'études s'articulant sur une théorie était systématiquement rejetée.

J'aurais alors pu travailler dans un laboratoire expérimental, mais j'ai compris après quelques entretiens que ce n'était pas pour moi. J'y aurais consacré l'essentiel de mes heures à entraîner des animaux, à construire du matériel expérimental et à collecter des données. Toute théorie que je développerais serait limitée à la portion du cerveau qu'étudierait ce laboratoire.

J'ai passé les deux années suivantes à hanter les bibliothèques universitaires, dévorant les articles de neurosciences les uns après les autres. J'en ai ainsi lu des centaines, et notamment les plus importants des cinquante dernières années. J'ai également lu ce qu'avaient écrit à propos du cerveau et de l'intelligence les psychologues, les linguistes, les mathématiciens et les philosophes. La formation que j'ai reçue est certes peu conventionnelle, mais de tout premier ordre. Après deux ans d'études en autodidacte, un changement s'imposait. J'ai élaboré un plan. Je retournerais travailler dans l'industrie pendant quatre ans, après quoi je réévaluerais mes chances dans les études. Je suis donc revenu dans la Silicon Valley travailler sur les ordinateurs individuels.

C'est là que j'ai rencontré mes premiers succès en tant qu'entrepreneur. Entre 1988 et 1992, j'ai créé l'une des premières tablettes numériques, le GridPad. Puis, en 1992, j'ai fondé Palm Computing, inaugurant une décennie qui me verrait concevoir certains des premiers ordinateurs de poche et smartphones, comme le PalmPilot ou le Treo. Tous mes collaborateurs chez Palm savaient que mon cœur battait pour les neurosciences, que mon travail dans l'informatique portable était provisoire. Concevoir certains des premiers ordinateurs de poche et smartphones a été une activité passionnante, et je ne doutais pas que des millions de gens finiraient par adopter ces appareils, mais l'étude du cerveau était plus importante encore à mes yeux. Je sentais qu'une théorie du cerveau aurait plus de répercussions positives pour l'avenir de l'humanité que l'informatique. Il fallait donc absolument que je retourne à la recherche sur le cerveau.

Il n'y avait pas de bon moment pour s'en aller, alors j'ai choisi une date au hasard et je me suis retiré des entreprises que j'avais contribué à créer. En 2002, avec le concours et les encouragements de quelques amis travaillant dans les neurosciences (notamment

Bob Knight à UC Berkeley, Bruno Olshausen à UC Davis et Steve Zornetzer à NASA Ames Research), j'ai fondé le Redwood Neuroscience Institute. Tout entier voué à la théorie du néocortex, le RNI employait dix chercheurs à plein temps. Nous étions tous passionnés par les théories générales du cerveau, et le RNI a été l'un des seuls endroits du monde où cette optique n'était pas seulement permise, mais attendue. Pendant les trois ans que j'ai passés à la tête du RNI, nous avons reçu plus de cent collaborateurs visiteurs, pendant quelques jours ou quelques semaines. Nous avons organisé des conférences hebdomadaires, ouvertes au public, qui donnaient généralement lieu à des débats qui se prolongeaient pendant des heures.

Tous ceux qui ont travaillé au RNI, moi le premier, ont trouvé cela formidable. J'ai eu l'occasion de rencontrer et de passer du temps avec bon nombre des principaux spécialistes en neurosciences du monde. Cela m'a permis de me constituer un savoir dans diverses branches des neurosciences, ce qui n'est pas évident lorsqu'on occupe un poste universitaire ordinaire. Le problème, c'est que je cherchais la réponse à un ensemble précis de questions, et que je ne voyais pas l'équipe s'orienter vers un consensus à leur sujet. Chacun des chercheurs se bornait à faire sa cuisine de son côté. Alors, après trois années passées à la tête d'un institut, j'ai compris que le meilleur moyen d'atteindre mes objectifs serait de diriger ma propre équipe de chercheurs.

À tous autres égards, le RNI se portait à merveille, au point qu'on a décidé de le transférer à l'Université de Berkeley. Oui, l'institution même qui m'avait dit que je ne pouvais pas étudier la théorie du cerveau a décidé, dix-neuf ans plus tard, qu'un centre dédié à la théorie du cerveau était très exactement ce dont elle avait besoin. Le RNI existe encore aujourd'hui sous le nom de Redwood Center for Theoretical Neuroscience.

C'est lorsque le RNI s'est établi à Berkeley qu'avec plusieurs collègues j'ai lancé Numenta. C'est une société de recherche indépendante, dont le premier objectif est de mettre au point une théorie du fonctionnement du néocortex. Le second est d'appliquer ce que nous apprenons du cerveau à l'apprentissage machine et à l'intelligence des machines. Numenta ressemble à un laboratoire de recherche

universitaire ordinaire, mais en plus souple. Je peux y diriger une équipe, veiller à ce que nous ramions tous dans le même sens et tester de nouvelles idées aussi souvent qu'il le faut.

À l'heure où j'écris ces lignes, malgré plus de quinze ans d'âge, Numenta demeure à bien des égards une start-up. L'étude du fonctionnement du néocortex est particulièrement ardue. Elle exige un niveau de souplesse et de focalisation qu'on rencontre généralement dans une start-up. Il faut aussi beaucoup de patience, ce qui en revanche n'est pas le propre des start-up. Notre première découverte d'importance – la façon dont un neurone fait des prédictions – n'a eu lieu qu'en 2010, soit cinq ans après la fondation. Il faudrait encore en attendre six pour que survienne celle dans le néocortex de référentiels similaires à des cartes, en 2016.

En 2019, nous avons entamé notre deuxième mission, l'application des principes cérébraux à l'apprentissage machine. C'est aussi l'année où je me suis mis à écrire ce livre, pour communiquer ce que nous découvrons.

Je trouve fascinant que la seule chose dans l'Univers qui sache que l'Univers existe soit cette masse d'un peu plus de deux kilos de cellules qui flotte en suspension dans notre tête. Ça me rappelle cette ancienne énigme : lorsqu'un arbre tombe dans la forêt, s'il n'y a personne pour l'entendre, a-t-il vraiment fait du bruit ? On peut pareillement se demander : si l'Univers venait à entrer et sortir de l'existence sans qu'il y ait de cerveaux pour le savoir, aurait-il vraiment existé ? Comment en être sûr ? Les quelques milliards de cellules qui flottent sous votre crâne savent non seulement que l'Univers existe, mais qu'il est vaste et ancien. Ces cellules ont appris un modèle du monde, un savoir qui, pour autant qu'on sache, n'existe nulle part ailleurs. J'ai passé ma vie à me démener pour comprendre comment le cerveau s'y prend, et cela m'a fait découvrir des choses exaltantes. J'espère aujourd'hui vous transmettre cette exaltation. Alors, commençons sans attendre.