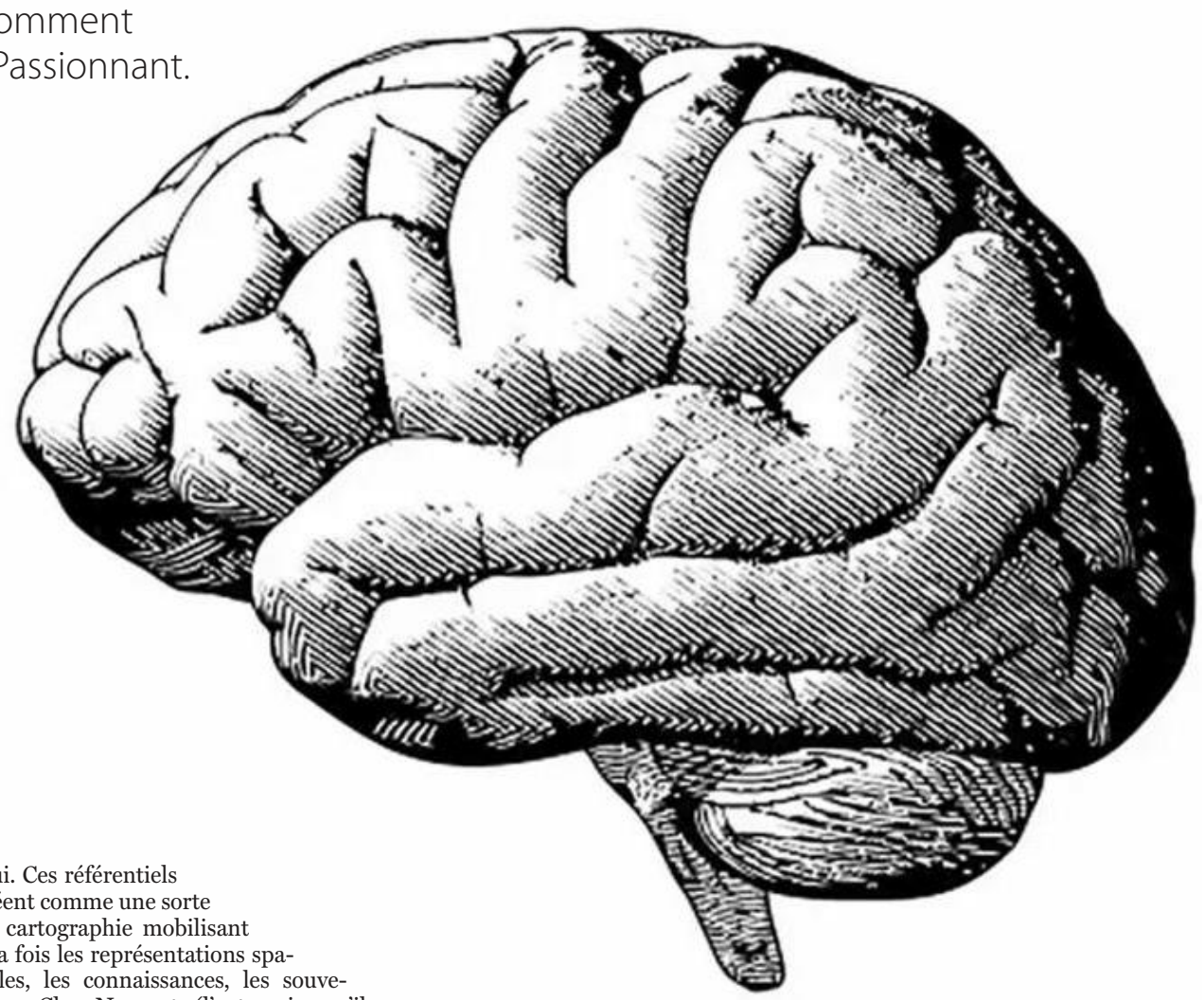
« Les **machines** n'ont pas d'intelligence. Mais elles en auront »

Dans son nouveau livre, Jeff Hawkins, un ingénieur en informatique devenu neuroscientifique, explore comment notre cerveau appréhende le monde et comment l'intelligence artificielle s'en inspire. Passionnant.



Le néocortex occupe environ 75% de l'espace du cerveau et est responsable de toutes les fonctions cognitives « supérieures », comme le raisonnement spatial ou le langage. © NUMENTA.

Tribune de Genève

ENTRETIEN
GREGORY WICKY

Parcours peu banal que celui de Jeff Hawkins. Ingénieur en informatique, il fonde dans les années 1990 l'entreprise Palm, qui proposa au monde ses premiers ordinateurs de poche. Désormais neuroscientifique, il dirige une équipe dans la Silicon Valley, en Californie, qui tente de modéliser le cerveau humain, notamment dans le but de développer des intelligences artificielles (IA) qui s'en inspirent.

Dans son livre, *1000 cerveaux*, le chercheur explore en détail le fonctionnement de notre néocortex. Partie « récente » du cerveau, occupant environ 75 % de son espace, celui-ci est responsable de toutes les fonctions cognitives « supérieures », comme le raisonnement spatial ou le langage. On le distingue des régions « anciennes » de l'appareil cérébral, présentes également chez les reptiles ou les amphibiens, sur lesquelles il s'est greffé au fil de l'évolution. Vulgarisé juste ce qu'il faut pour des lecteurs peu versés dans les sciences, l'ouvrage développe la théorie – saluée en préface par le célèbre biologiste Richard Dawkins pour sa fulgurance digne de Darwin – selon laquelle notre néocortex fait sens du monde en en générant non-stop des milliers de modèles. Hawkins expose ensuite les pistes qui permettront à l'intelligence artificielle de l'imiter, et ainsi faire progresser l'humanité – nous avons affaire à un optimiste.

Pourriez-vous résumer simplement la théorie des « 1.000 cerveaux » ?
Pour appréhender le monde, notre néocortex construit un modèle permettant de recréer dans notre esprit tout ce que nous savons. Un peu comme un architecte ferait une maquette de son projet, mais en infiniment plus complexe : tous les gens et les endroits que nous connaissons, tout ce que nous avons appris.
Deux facteurs sont déterminants pour nous permettre de créer ce modèle. D'abord, nous le faisons par le mouvement : de nos mains sur des objets, de notre corps dans l'espace, de nos yeux… C'est comme ça que nous apprenons. Le deuxième facteur est que nous ne créons en réalité pas un seul modèle, mais des milliers, que nous mobilisons en permanence.

Comment cela fonctionne-t-il ?
Le cortex est structuré par quelque 150.000 colonnes, qui font environ la taille d'un grain de riz. Les neurones y sont particulièrement nombreux et densément connectés. Chacune de ces colonnes est une machine à créer des modèles, qu'on appelle également des référentiels. C'est une construction mentale multiple, distributive, et pas monolithique, comme on l'a longtemps pensé. Les colonnes, présentes à l'identique dans toutes les régions du cortex, créent leur propre modèle qu'elles mettent en commun, puis « votent » sur la perception à adopter et la réponse à apporter.

Vous avancez que nous fabriquerons un jour des intelligences artificielles fonctionnant selon ce principe ?
Oui. Ces référentiels créent comme une sorte de cartographie mobilisant à la fois les représentations spatiales, les connaissances, les souvenirs… Chez Numenta (l'entreprise qu'il dirige, NDLR), nous réalisons déjà des versions rudimentaires de tels référentiels. A terme, les IA seront capables d'apprendre par elles-mêmes, de s'ajuster toutes seules à leur environnement, sans qu'on doive investir des millions de dollars et des semaines entières à les abreuver de données.

Vous dites qu'il n'y a, pour l'heure, pas d'intelligence dans l'IA ?
Les nouveaux systèmes dont on parle beaucoup aujourd'hui, comme ChatGPT ou DALL_E, sont très bons, et continueront de se développer, car ils ont un gros potentiel commercial. Mais ils sont fragiles : ils font des erreurs, on les trompe facilement. Ils utilisent une quantité incroyable de données pour tenter d'imiter ce que ferait un être humain en répondant à des requêtes. Mais c'est un tour de passe-passe. Il n'y a pas d'ancrage dans le monde, pas de réflexion. La plupart des chercheurs dans le domaine s'accordent à le dire : à terme, l'IA sera tout autre chose.

Vous prenez l'exemple de robots chargés de créer sur Mars des environnements capables d'accueillir des êtres humains…
C'est un exemple un peu fantaisiste, mais pas non plus irréaliste. Pour collaborer, surmonter des obstacles, apprendre de leurs erreurs, de telles machines devraient bénéficier d'intelligence artificielle générale, c'est-à-dire pas destinées à une seule fonction, comme la plupart des IA actuelles. Pour l'heure, un enfant de 5 ans qui observe le monde et voit comment une porte s'ouvre ou un verre d'eau se renverse est bien mieux équipé que n'importe quelle IA.
Ces machines devraient également être équipées de senseurs, donc posséder quelque chose comme un corps, qui les ancre dans le monde. Car c'est le mouvement dans l'espace qui permet d'apprendre. Ce ne serait pas forcément un corps de robot, type cyborg de science-fiction, mais il lui faudrait une forme de matérialité.

On suppose que vous n'êtes pas de ceux qui pensent, comme Elon Musk, qu'il faut un moratoire sur l'IA.
Non, en effet. Je pense qu'il faut qu'on fasse attention à cette technologie, car elle est puissante et se développe vite. C'est un peu comme les réseaux sociaux : ils ont eu des conséquences négatives et positives. Il faut surveiller, légiférer, poser des garde-fous. Mais là n'est pas mon champ de réflexion.
La plupart des gens qui ont signé l'appel dont vous parlez ne comprennent pas très bien comment cette technologie fonctionne. Ils ont peur que, soudain, nous fassions le progrès de trop, et hop, ce sera la fin du monde. Il n'y a aucune chance que cela se produise. Quant à Elon Musk, il a certes du talent, mais je crois qu'on lui accorde un peu trop de crédit…

Vous êtes plutôt optimiste quant aux menaces que peut représenter cette technologie ?
Les êtres humains sont complexes. Même si notre néocortex est très développé, nous restons des animaux, avec une partie de cerveau ancienne et des instincts de survie, de reproduction… Nos aspirations plus élevées, nos questionnements philosophiques sont souvent en conflit avec nos pulsions et nos désirs, qui peuvent nous pousser à mentir, à voler, à souhaiter du mal à autrui. La plupart des gens peinent à séparer ces deux choses. Ils pensent que l'une n'est pas possible sans l'autre. Mais le type de machine intelligente que nous voulons créer ne sera pas sujet à ces pulsions. Et elle ne pourra pas les développer. Les IA ne peuvent pas nous poser de risque existentiel, vouloir nous détruire ou prendre le contrôle du monde. Le vrai danger lié à cette technologie, c'est qu'elle soit utilisée à des mauvaises fins : désinformation, abus de pouvoir, etc. Ce risque-là, bien sûr, est réel.

Il faudra bien pourtant que ces machines soient dotées d'objectifs…
Oui, bien sûr, il faudra leur poser des buts. La voiture automatisée doit vous amener d'un point A à un point B sans percuter d'enfants en chemin ! Mais elle ne va jamais développer des objectifs du genre : « J'aimerais avoir une plus grosse maison que mon voisin. » Ça, ça nous est propre. L'IA ne fera rien à moins que vous ne lui donniez quelque chose à faire, c'est facile à implémenter. On peut imaginer quelque chose comme les trois lois de la robotique d'Isaac Asimov.

Un enfant de 5 ans qui observe le monde et voit comment une porte s'ouvre ou un verre d'eau se renverse est bien mieux équipé que n'importe quelle IA

Reste la question de la conscience. Vous pensez que les machines en seront équipées…
Je dis dans le livre qu'elles ne le seront pas forcément, mais que nous pourrons le faire si nous le souhaitons. Le concept de conscience, je pense qu'on a tendance à en faire une trop grande affaire. A mon sens, il se décompose en deux éléments : le sens de sa propre existence, et une forme de mémoire. Ce deuxième élément implique de se souvenir de ses états précédents, ainsi que des choses qu'on a faites jusqu'ici. C'est naturellement indispensable pour apprendre. Les IA pourront être dotées de ces fonctions sans problème.

Ce ne sera donc pas un meurtre que de les débrancher ?
Non. Déjà nous, humains, nous nous débranchons toutes les nuits au moment du coucher, et nous nous rallumons au réveil. Quant au problème que pourrait poser le fait de détruire une machine consciente, là aussi, je n'en vois pas. La peur de la mort nous vient de notre cerveau ancien. Sans lui, pas de crainte ou de chagrin. Les machines s'en ficheront bien qu'on les éteigne, qu'on les démonte, ou qu'on les envoie à la casse.

Les trois lois de la robotique

Les trois lois de la robotique de l'auteur de science-fiction Isaac Asimov furent proposées pour la première fois dans sa nouvelle *Cercle vicieux*, en 1942, avant d'être peaufinées par lui-même et d'autres au fil des décennies. Elles sont les suivantes :
– un robot ne peut porter atteinte à un être humain ni, restant passif, laisser cet être humain exposé au danger ;
– un robot doit obéir aux ordres donnés par les êtres humains, sauf si de tels ordres entrent en contradiction avec la première loi ;
– un robot doit protéger son existence dans la mesure où cette protection n'entre pas en contradiction avec la première ou la deuxième loi.
En 2020, un député français a déposé une proposition de loi pour promulguer une charte de l'intelligence artificielle et des algorithmes. Son article 2 est constitué de ces trois lois.
G.W.