POUR LA SCIENCE
INTELLIGENCE ARTIFICIELLE Jusqu'où ira-t-elle?
Hors-série numéro 115

Pour la Science

HORS-SÉRIE

La science expliquée par ceux qui la font

n° 115 - 05.22/06.22

L 13264 - 115 H - F: 9,90 € - RD

« L'informatique nous permet d'avoir l'ambition la plus dingue, celle d'une société égalitaire où tout le monde vivrait bien »

Serge Abiteboul, directeur de recherche à Inria

Jusqu'où ira L'INTELLIGENCE ARTIFICIELLE?

IA et *fake news* : liaisons dangereuses

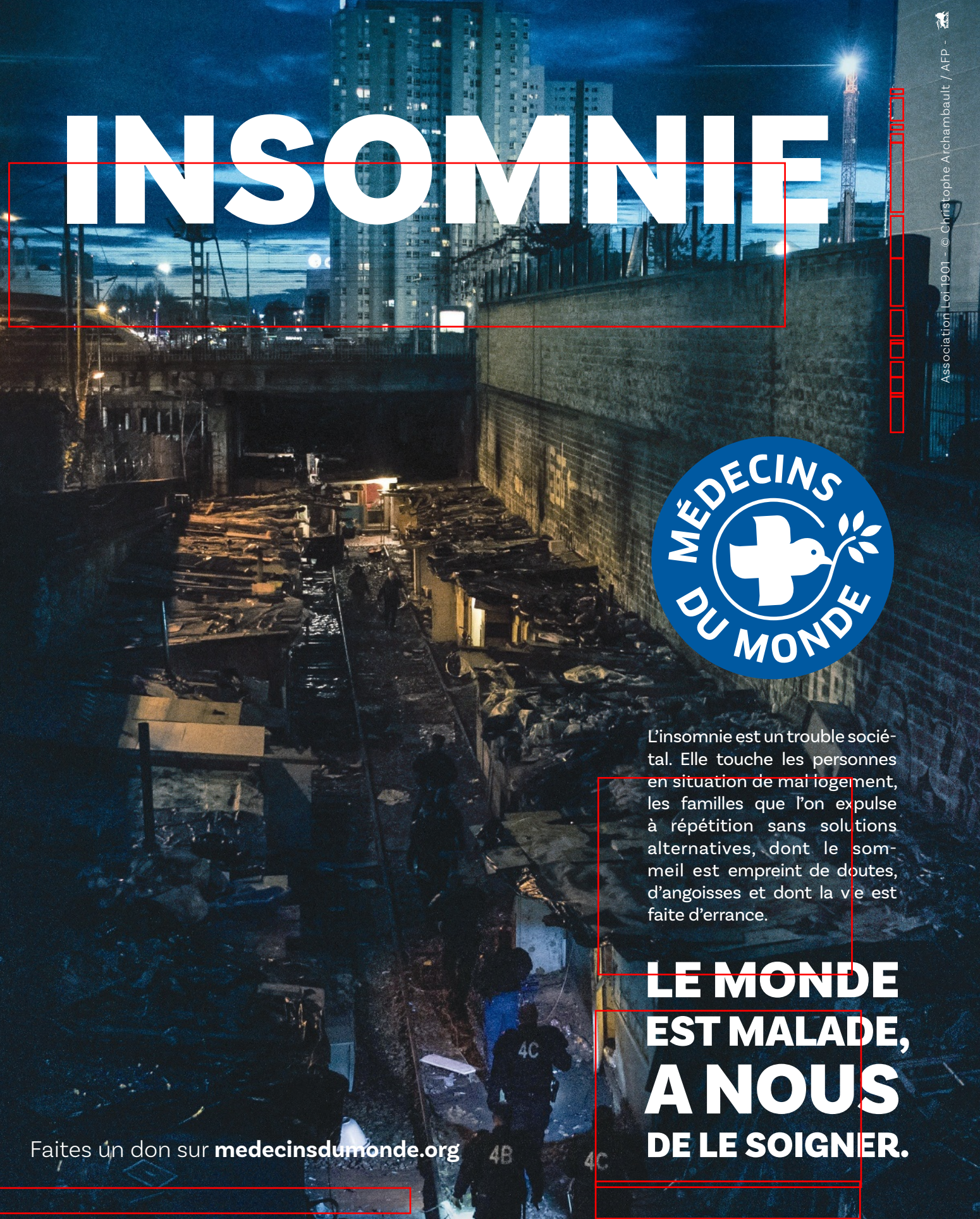
Une aide précieuse en médecine

Quand l'électronique mime le cerveau

Où en est le véhicule autonome?

Édition française de Scientific American - BEL./LUX.: 11,40 € - CH: 17,10 FS - CAN.: 16,99 $CA

INSOMNIE



Association Loi 1901 - © Christophe Archambault / AFP -

MÉDECINS DU MONDE

L'insomnie est un trouble sociétal. Elle touche les personnes en situation de mal logement, les familles que l'on expulse à répétition sans solutions alternatives, dont le sommeil est empreint de doutes, d'angoisses et dont la vie est faite d'errance.

LE MONDE EST MALADE, A NOUS DE LE SOIGNER.

Faites un don sur **medecinsdumonde.org**

05.22 / 06.22

Pour la Science HORS-SÉRIE

La bêtise artificielle

par **Loïc Mangin**
Rédacteur en chef adjoint
à *Pour la Science*

Déchiffrer le grec ancien, comprendre les vocalisations des cochons, contrôler les plasmas pour la fusion nucléaire, inspecter les ponts, proposer au parfumeur de nouvelles pistes olfactives, converser sur la santé sexuelle… Ce sont les nouvelles applications de l'intelligence artificielle (IA), et plus précisément de son avatar le plus récent, l'apprentissage profond, qui ont été annoncées rien que ces dernières semaines… Pas un recoin de nos vies n'échappe à la déferlante au point que l'on peut se demander où elle s'arrêtera. Égalera-t-elle un jour la nôtre ? Ce n'est sans doute pas pour tout de suite, car malgré d'indéniables succès, décrits dans ce numéro, et à supposer que l'on se mette d'accord sur une définition de l'intelligence, des difficultés, aussi décrites dans ces pages, se dressent encore. Par exemple, on comprend mal le fonctionnement intime des algorithmes.

Au micro de Xavier de La Porte, dans le podcast « Le Code a changé », Milad Doueihi, titulaire de la chaire d'humanisme numérique à Sorbonne Université, livrait quelques pistes pour améliorer aussi bien les machines que notre cohabitation avec elles. Premier conseil, introduire des mécanismes d'oubli pour que les programmes se concentrent sur l'essentiel. Ensuite, s'inspirer des enfants, de leur apprentissage sans fin et de leur acquisition de la sociabilité pour tenir compte du contexte. Enfin, il appelait à réintroduire un peu de « bêtise » dans ces machines pour qu'elles nous laissent une marge de liberté. Un prochain *Hors-Série* sera donc consacré à la bêtise artificielle !

Ont contribué à ce numéro

Nicholas Ayache
est directeur de recherche à Inria et directeur scientifique de l'Institut interdisciplinaire d'intelligence artificielle (3IA) Côte d'Azur, à Nice.

Laurence Devillers
est professeuse en intelligence artificielle à Sorbonne Université et dirige une équipe de recherche au CNRS-LISN (Saclay), où elle est responsable de la chaire IA Humaaine.

Julie Grollier
est chercheuse à l'unité mixte de physique CNRS/Thales, à Palaiseau. Elle a reçu le prix Irène Joliot-Curie 2021 de la Femme scientifique de l'année.

Christof Koch
est neurobiologiste, directeur scientifique et président de l'institut Allen pour les sciences du cerveau, à Seattle, aux États-Unis.

SOMMAIRE

Jusqu'où ira L'INTELLIGENCE ARTIFICIELLE?

p. 6 Repères
Des schémas, des chiffres, des définitions… L'essentiel pour apprécier ce numéro.

p. 10 Grand témoin
Serge Abiteboul

> J'ai un problème: je ne sais pas trop ce qu'est l'intelligence artificielle

01 Prouesses et promesses

p. 18 La fée IA au chevet des malades
Nicholas Ayache
Les algorithmes brillent à interpréter l'imagerie médicale.

ENTRETIEN
p. 24 La traduction automatique doit son succès aux statistiques
Thierry Poibeau

p. 28 Les protéines, nouveau terrain de jeu de l'IA
Michael Eisenstein
Leur forme en 3D se calcule de mieux en mieux.

p. 34 Les agents jouent collectif
Philippe Mathieu, Sébastien Picault et Jean-Christophe Routier
Il y a de l'intelligence collective dans les systèmes multi-agents.

p. 42 Communication, les deux visages de l'IA
Laurence Devillers
Elle favorise la désinformation, et offre un moyen de lutter contre.

02

Chantier en cours

p. 54 Interro écrite pour l'IA
Melanie Mitchell
Comment s'assurer que les machines comprennent ?

p. 60 Un seul pixel vous manque et tout est désappris
Jean-Paul Delahaye
L'apprentissage profond manque surtout de… profondeur.

p. 68 Théorie en voie de construction
Kevin Hartnett
Entre architecture et fonction, on ne sait pas bien faire le lien.

p. 76 L'ordinateur sera-t-il un jour le boss des maths ?
Jean-Paul Delahaye
Un programme fait des maths. De là à devenir mathématicien…

03

Se rapprocher de l'humain

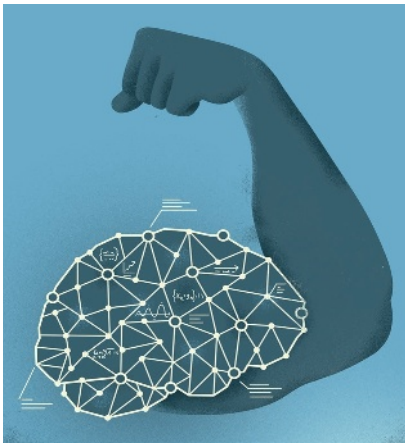
p. 86 Itinéraire bis pour la voiture autonome
Steven Shladover
Confier le volant à son auto ? Ce n'est pas encore pour demain.

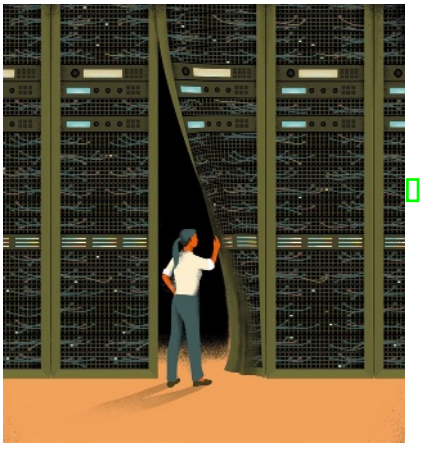
p. 94 Bien vivre avec les machines
Laurence Devillers
Entre les machines et nous, il y a de l'émotion. Et là, gare !

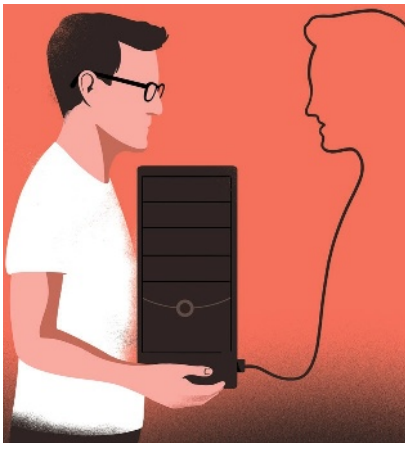
p. 102 S'inspirer du cerveau pour mieux l'imiter
Julie Grollier et Damien Querlioz
Côté économie d'énergie, notre matière grise est un modèle.

p. 112 L'improbable conscience artificielle
Christof Koch
Une machine consciente, est-ce possible ? On peut en douter.

p. 120 Les incontournables
Des livres, des expositions, des podcasts… Les coups de cœur de la rédaction.







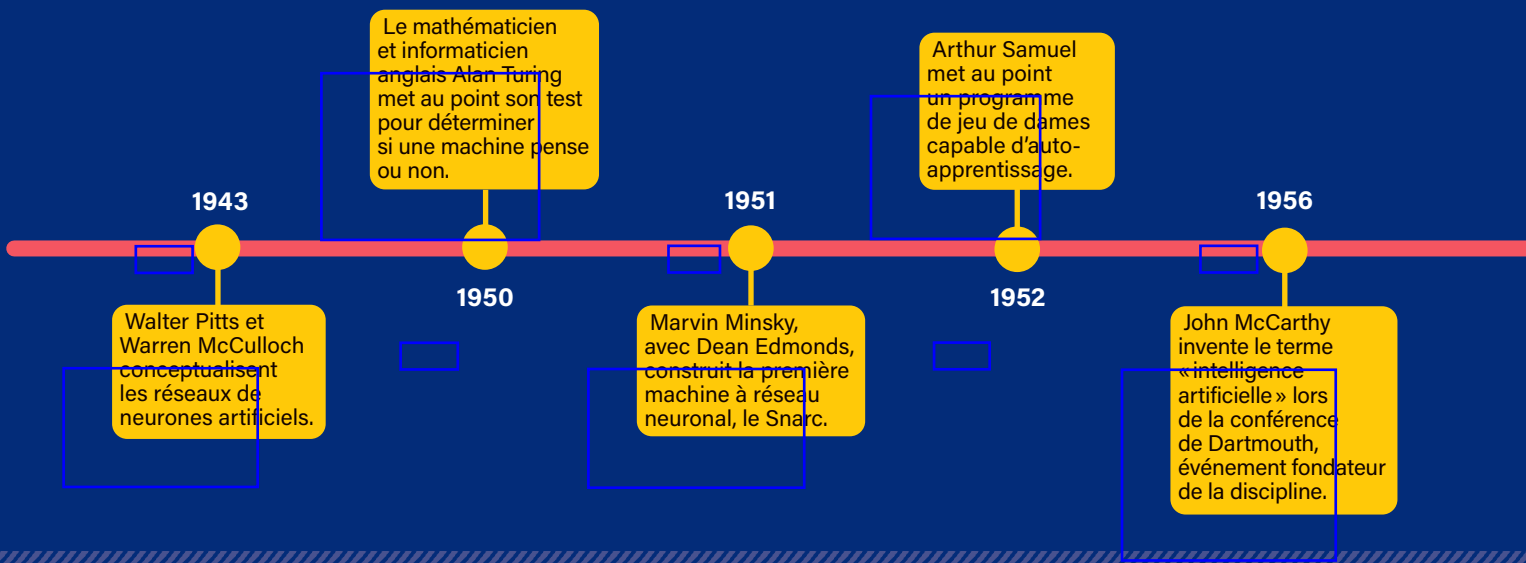
Intelligence artificielle

Du symbole aux statistiques

L'intelligence artificielle (ou IA) émerge dans le champ plus global des sciences informatiques au lendemain de la Seconde Guerre mondiale. Il n'en existe pas une définition universelle, les chercheurs ayant du mal à tracer une limite entre IA et informatique… et plus globalement à définir ce qu'on entend par « intelligence » ! L'approche la plus ancienne, dite **« symbolique »**, se fonde sur la logique et les mathématiques. Le programme fixe des règles (déduction, classification, hiérarchisation, etc.), que la machine exécute automatiquement. Avantage : le raisonnement est compréhensible par l'esprit humain. Jusqu'aux années 1990, cette approche est ultradominante et irrigue toutes les innovations. Les systèmes actuels d'aide à la décision, de gestion des connaissances ou d'e-santé s'inscrivent dans cette filiation.

Très différente est l'approche dite **« statistique »**. Ses algorithmes cherchent des régularités dans les données disponibles pour extraire des connaissances, sans modèle préétabli. Le programme n'est plus ce qui pilote le système, mais… ce qui en résulte. Une vraie révolution ! Née sous sa forme actuelle dans les années 1980, l'IA statistique est sous les projecteurs depuis une dizaine d'années, notamment parce qu'elle a permis de considérables progrès dans la reconnaissance d'images ou la traduction automatique. Inconvénient : la démarche est propre à la machine, et n'est pas compréhensible étape par étape.

Une histoire à rebondissements



GLOSSAIRE

> Algorithme
Suite précise d'instructions permettant d'obtenir un résultat en un nombre fini d'étapes. Un algorithme reçoit des données à traiter (nombres, textes, etc.) ; les données peuvent affecter le cours des instructions. Le mot vient du vieil espagnol *alguarismo*, lui-même issu d'un mot arabe signifiant « celui de Huwārizm », surnom du mathématicien persan Muhammad Ibn Mūsā al-Khuwārizmī, introducteur de l'algèbre dans l'Europe médiévale.

> Apprentissage automatique (ou statistique)
Approche de l'IA visant à tirer des généralités plausibles à partir d'un très grand nombre d'observations. L'algorithme apprend la tâche qui lui a été assignée par « essais et erreurs » avant de se débrouiller seul.

> Apprentissage profond
Sous-catégorie de l'apprentissage automatique, caractérisée par des réseaux de neurones artificiels organisés en différentes couches, échangeant les uns avec les autres *(voir pages suivantes)*.

> Connexionniste
Se dit de l'approche qui utilise les réseaux de neurones, offrant une voie différente de l'approche symbolique.

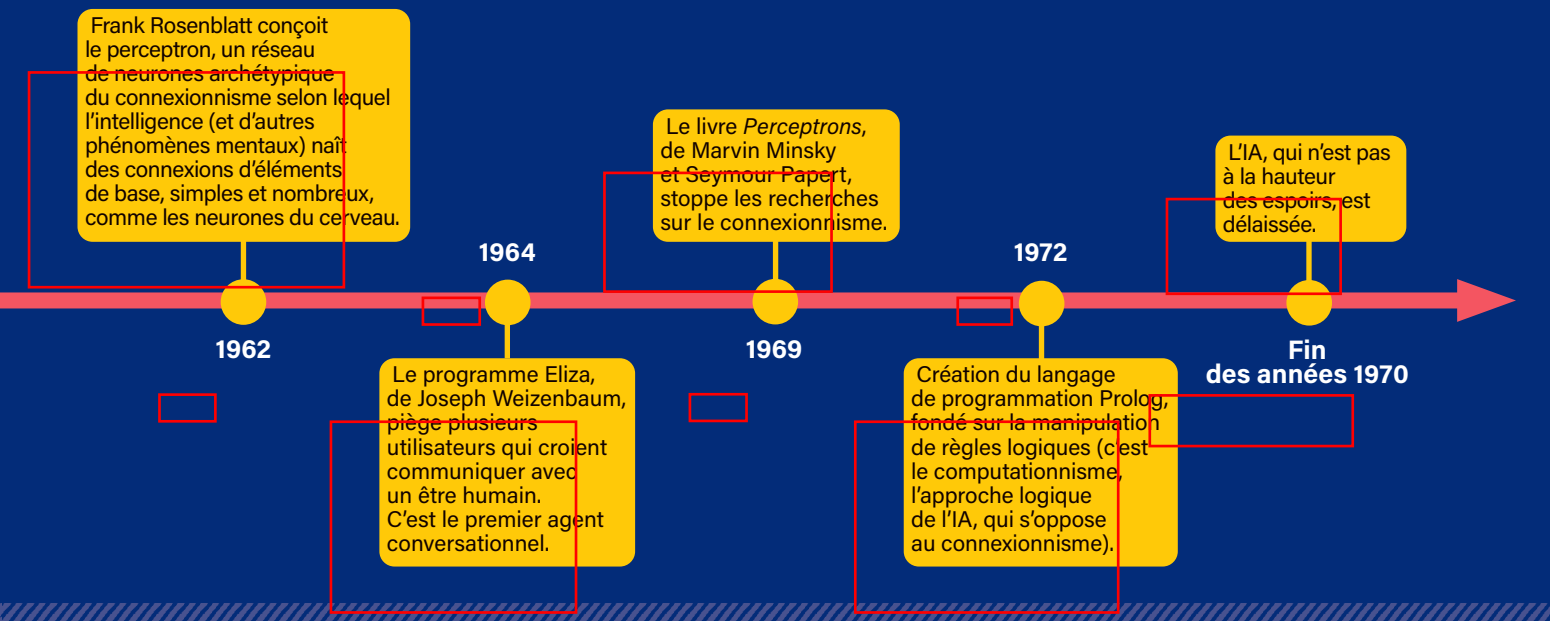
> Hivers de l'IA
Phases pendant lesquelles se manifestent un désintérêt et une baisse des financements pour l'intelligence artificielle. Le premier survient au début des années 1970 et le second à la fin des années 1980 : dans les deux cas, les résultats s'avéraient très inférieurs aux promesses.

> Jeu d'échecs
Pour les mécanismes intellectuels supposés formalisables qui régissent sa pratique, ce jeu a longtemps inspiré les chercheurs en informatique. Selon le mathématicien russe Alexandre Kronrod, il a été aux informaticiens ce que la drosophile fut aux généticiens !

© Frederic Lewis/Getty Images

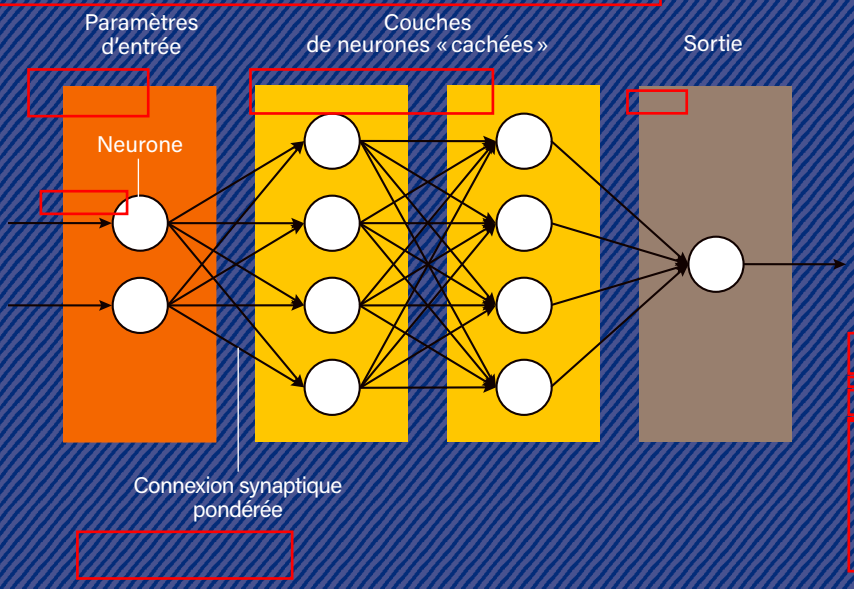
> Perceptron
Conçue pour la reconnaissance d'images et financée par la marine américaine, c'est la première véritable machine connexionniste *(photo)*.

> Réseau de neurones
Algorithme inspiré du fonctionnement collectif des cellules de notre système nerveux *(voir pages suivantes)*.



Comment fonctionne un réseau de neurones

1 S'ils s'inspirent des cellules de notre système nerveux, les réseaux de neurones artificiels sont des algorithmes et non des objets physiques. Ils représentent des nœuds, les « neurones », reliés entre eux par des « synapses ». Dans le cas de l'apprentissage profond, les neurones sont organisés en couches. Le nombre de couches et de neurones par couche dépend de la tâche à accomplir par le réseau : analyse d'images, de textes, de sons… Néanmoins, on ne sait pas encore très bien adapter le nombre et la largeur des couches en fonction de l'objectif recherché *(voir Théorie en voie de construction, de K. Hartnett, page 68).*



© Shutterstock/all_is_magic, Pour la Science

1982

John Hopfield imagine un nouveau type de réseau neuronal capable d'apprendre et de traiter de l'information. C'est le retour du connexionnisme.

1984

Douglas Lenat élabore Cyc, un projet d'IA reposant sur une base de connaissances générales et censé raisonner à la façon d'un être humain.

Fin des années 1980

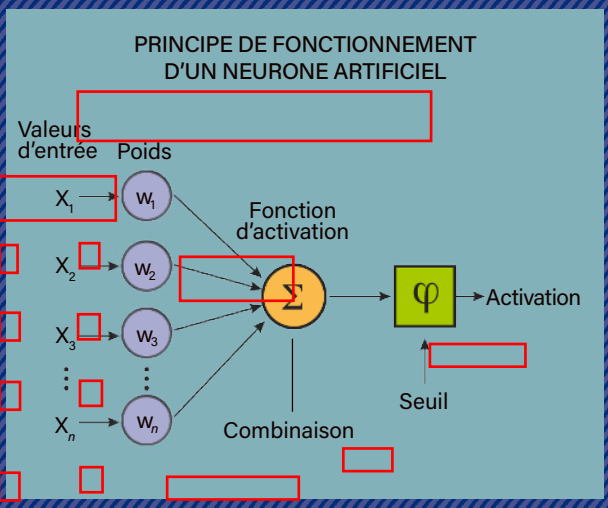
Second hiver de l'IA au moment de l'apparition des premiers ordinateurs de bureau. Les grosses machines dédiées aux recherches sur l'IA n'ont plus la cote.

1997

Le 11 mai, Deep Blue, d'IBM, bat aux échecs le champion du monde en titre, Garry Kasparov.

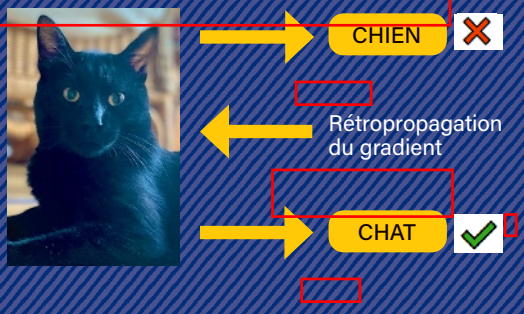
Début des années 2000

L'IA profite des progrès spectaculaires de l'informatique, notamment en ce qui concerne la puissance des ordinateurs.



2 Le signal d'entrée est envoyé aux neurones de la première couche, qui transmettent un signal à ceux de la suivante, etc. Au niveau de chaque neurone, les signaux issus de la couche précédente sont additionnés suivant une somme pondérée par les coefficients (ou poids) portés par les synapses, coefficients qui définissent la fonction d'activation. Laquelle est configurée grâce à une phase d'apprentissage du réseau.

3 On soumet des données en entrée du réseau, qui effectue un calcul et compare la solution obtenue à celle qu'on lui fournit. Exemple : s'il s'agit de reconnaître des images de chats, à chaque fois qu'on soumet au réseau une photo et qu'il se trompe, les pondérations synaptiques sont corrigées par une méthode mathématique nommée « rétropropagation du gradient ». Celle-ci s'appuie sur des outils statistiques pour calculer la variation d'une erreur au niveau d'un neurone, de la dernière couche vers la première. Après un certain nombre d'itérations, la fonction d'activation est alors correctement configurée pour que le réseau de neurones reconnaisse les images de chats avec un taux d'erreurs très faible.



© Loïc Mangin

2005 — Un robot de l'université Stanford conduit de manière autonome pendant plus de 200 kilomètres sur la piste d'un désert sans reconnaissance préalable.

2006 — Yoshua Bengio, Geoffrey Hinton et Yann LeCun réussissent à entraîner des réseaux neuronaux « profonds », c'est-à-dire constitués de plus d'une couche de neurones artificiels.

2011 — Le programme Watson, conçu par IBM, gagne au jeu télévisé Jeopardy contre les deux plus grands champions.

2016 — AlphaGo, un logiciel conçu par la société Google DeepMind, bat Lee Sedol, champion du monde du jeu de go.

2017 — AlphaGo se fait humilier au jeu de go (100 à 0) par... AlphaGo Zéro, le nouveau logiciel de Google DeepMind.

J’ai un problème : je ne sais pas trop ce qu’est l’intelligence artificielle

Serge Abiteboul
est directeur de recherche à Inria et à l’ENS Paris

La créativité ? Les algorithmes en sont encore incapables

Peut-on dire que, plus que le jeu d'échecs, les réseaux de neurones ont dopé la recherche sur l'intelligence artificielle ?

Ce serait un peu du bourrage de crâne. Un système comme Deep Blue, d'IBM, qui a battu le champion du monde d'échecs Gary Gasparov en 1997, embarquait des années de recherches, le plus souvent développées pour autre chose. Toutes les techniques informatiques ont des applications considérables qui sont à l'œuvre tous les jours, qui marchent très bien et ont pour nom « gestion de données », « système d'exploitation », « compilateur », « communication numérique », « interface humain-machine », « calcul parallèle », « raisonnement logique », etc. Ce qu'on observe depuis une dizaine d'années, c'est l'arrivée d'algorithmes d'apprentissage automatique, qui ont obtenu des résultats superbes dans des domaines qui nous bloquaient jusque-là. Mais internet et votre téléphone portable fonctionnent en grande partie sans eux, même si on utilise de plus en plus l'apprentissage automatique, par exemple dans les assistants vocaux.

Pouvez-vous préciser l'apport de l'apprentissage automatique ?

Sur certains problèmes, les approches dites « symboliques », fondées sur le calcul et le raisonnement, ne progressaient presque plus. Je pense notamment à des problèmes tout bêtes comme de distinguer l'image d'un chat de celle d'un chien, ou, plus intéressant, de reconnaître une tumeur cancéreuse. Et, assez soudainement, des techniques connues depuis longtemps et qui avaient souvent des résultats médiocres, les réseaux de neurones, se sont mises à fonctionner. La traduction automatique des langues, par exemple, s'est améliorée considérablement *(lire l'entretien avec Thierry Poibeau, page 24)*. Avec l'apprentissage automatique et les méthodes statistiques, ainsi nommées parce qu'elles s'appuient sur de gros volumes d'informations, un logiciel apprend de données fournies par des humains, en observant son environnement, en simulant des situations, etc. Pour faire une analogie, vous pouvez apprendre à jouer au tarot parce que des amis vous en expliquent les règles, mais vous pouvez aussi vous former en regardant des joueurs. Souvent, les deux modes coexistent. On connaissait depuis des années des algorithmes pour faire apprendre aux machines, notamment les réseaux de neurones. Pourquoi tout à coup sont-ils devenus plus performants ?

Il y a eu une sorte de conjonction de planètes avec l'arrivée, presque au même moment, de beaucoup plus de puissance de calcul, de plus en plus de corpus de données pour nourrir l'apprentissage, et du développement de nouveaux algorithmes dits « d'apprentissage profond » (*deep learning,* en anglais). D'un coup, des problèmes qui nous résistaient depuis des années se sont mis à tomber. Cela avait un côté génial. Mais, encore une fois, cette approche n'a pas remplacé ce qui existait avant. Même quand AlphaGo, de DeepMind, a battu le joueur de go Lee Sedol, il n'utilisait pas uniquement des algorithmes d'apprentissage profond.

Le revers de ces méthodes n'est-il pas leur opacité ?

En effet, quand on fait tourner un algorithme d'apprentissage profond, on ne sait pas expliquer pourquoi on arrive à un résultat particulier. Les longs calculs réalisés ne font pas à proprement parler un raisonnement, en tout cas un raisonnement qu'un humain serait capable de comprendre. On peut penser que, tant pis, seule l'efficacité prime, mais ce n'est pas si simple. Prenons deux exemples en médecine : un algorithme d'apprentissage qui aide à retrouver des tumeurs cancéreuses *(voir La fée IA au chevet des malades, par N. Ayache, page 18)* examine des milliers d'images annotées par des médecins, et à partir de toute cette connaissance se prononce sur les images qu'on lui soumettra. Impossible pour un humain de se former en étudiant toutes ces images : il y en a trop. Et de toute façon, il y aura un médecin, voire une équipe, pour discuter l'avis de la machine qui sera un avis comme un autre, pris comme tel. En revanche, en matière de diagnostic médical, si vous rentrez dans un programme un grand nombre d'informations sur un patient, et qu'à la fin ce logiciel décide « c'est une hépatite », ça ne peut pas suffire au médecin qui a besoin d'explications. Il ou elle a besoin d'entendre que, en fonction des observations du malade, statistiquement ce peut être telle maladie avec 95 % de chances, mais aussi telle autre avec 5 % de chances, et qu'il faudrait poser telle question au malade pour écarter telle possibilité, ou demander tel examen complémentaire, etc. Dans les deux cas, il y a un travail collaboratif entre machine et humains. Dans le premier, nul besoin d'explications (les techniques d'apprentissage automatique un peu brutales sont efficaces). Dans le second, des explications sont indispensables.

Où finit l'informatique « ordinaire », où commence l'intelligence artificielle ?

Cette distinction n'a pas vraiment de sens. On a, à l'intérieur de l'informatique, un vrai continuum.

Quelle est alors votre définition de l'intelligence artificielle ?

Pour Alan Turing, une activité d'une machine sera qualifiée d'« intelligence artificielle » si elle est considérée comme intelligente quand un humain s'y livre. À ses yeux, ce ne peut être qu'une imitation de l'intelligence humaine, une

> Je ne connais pas beaucoup d'informaticiens qui essaient d'écrire des programmes idiots…

simulation. J'utilise la définition de Turing mais, honnêtement, ça ne me dit pas grand-chose puisque, tout comme lui, je ne sais pas définir l'intelligence humaine. En fait, j'ai un problème : je vous en parle, mais je ne sais pas trop ce qu'est l'intelligence artificielle ! L'expression fait fantasmer. Mais qu'est-ce qu'elle signifie ? Depuis ma thèse, je travaille sur des systèmes de gestion de base de données, qui répondent aux questions des humains. C'est quand même intelligent de répondre à des questions ! J'ai travaillé sur des bases de connaissances qui font de la déduction. Là encore, c'est intelligent de raisonner. Plus récemment, l'apprentissage automatique m'a permis d'introduire de nouvelles fonctionnalités dans des systèmes sur lesquels nous travaillons avec des étudiants. Distinguer ce qui en informatique tient de l'intelligence artificielle ou pas, ça n'aide en rien. Pour moi, c'est avant tout un *buzzword*, surtout utile pour récupérer des financements ou impressionner des amis. Le truc cool, aujourd'hui, n'est pas l'intelligence artificielle, mais l'apprentissage automatique qui vient compléter d'autres techniques essentielles de l'informatique.

Donc vous ne cherchez jamais à développer des programmes « plus intelligents » ?

Je cherche à faire des programmes qui résolvent des problèmes, qui répondent aux besoins de leurs utilisateurs. Cela dit, je ne connais pas beaucoup d'informaticiens qui essaient d'écrire des programmes idiots… même si on ne sait pas définir l'intelligence.

Quel est l'objectif de la recherche en intelligence artificielle ? Dépasser l'humain ?

Vous l'avez compris, je ne sais pas distinguer recherche en intelligence artificielle et en informatique. Les chercheurs en informatique veulent repousser les limites de la science. Certains se posent des questions théoriques, par exemple sur la calculabilité ou la puissance du raisonnement, presque du ressort des mathématiques pures. À l'autre bout du spectre, d'autres développent des produits informatiques prêts à être utilisés le mois d'après comme les logiciels scikit-learn (une bibliothèque Python pour l'apprentissage automatique) et Caml (un langage de programmation et un environnement populaire). Parfois, une recherche très théorique comme celle des universitaires Ronald Rivest, Adi Shamir et Leonard Adleman débouche sur un algorithme de chiffrement très pratique, le RSA, qui est à la base de tous les échanges chiffrés sur internet. Pour moi, cette diversité est la grande richesse de la recherche dans ma discipline. Les humains sont de magnifiques machines à résoudre des problèmes. Pourquoi ne pas essayer de les imiter avec des ordinateurs ? C'est le genre de défi qui fait avancer les sciences. Quant à les dépasser… pourquoi pas ? À vrai dire, l'informatique accomplit déjà des tas de choses dont nous sommes incapables. Reproduire à la main des calculs que votre smartphone traite à toute vitesse prendrait un temps dingue à des centaines de milliers de personnes qui commettraient des

millions d'erreurs. Les ordinateurs calculent bien mieux que nous. Faire mieux que l'humain n'est pas si difficile.

Mais, d'une calculette, on ne dit pas qu'elle est intelligente…

Un des trucs qu'on apprend à l'école primaire, c'est calculer, non ? Moi, je trouve ça intelligent. Le chien du voisin, sait-il calculer ?

Peut-être ce déni s'explique-t-il parce que la calculette est devenue ordinaire ?

En effet, comme c'est un objet de notre quotidien, on lui interdit d'être vraiment intelligent. Peut-être que ce qu'on sait expliquer par une suite d'opérations est dénué de vraie intelligence. La preuve automatique d'un théorème mathématique, on peut la décortiquer pas à pas. Et une machine reproduira «bêtement» le calcul, donc ça ne doit pas être bien sorcier. Mais comme on ne comprend pas comment fonctionne l'apprentissage automatique, alors c'est forcément intelligent.

Est-ce qu'il y a, sur l'intelligence artificielle, une approche particulière à la France ?

Non. La recherche en informatique est devenue extrêmement mondiale, il ne peut pas y avoir d'approche hexagonale. Il y a une grande fluidité entre les pays. J'ai fait ma thèse aux États-Unis comme beaucoup de collègues, on interagit sans cesse avec des collègues américains, européens, africains, asiatiques, nos labos sont peuplés de doctorants, postdoctorants, visiteurs, etc., de multiples nationalités. Si on voulait vraiment chercher une coloration française, ce serait plutôt du côté de la formation scolaire et universitaire. Nos étudiants avaient jusqu'à récemment, en moyenne, une formation plus mathématique que ceux venus d'ailleurs. Cela leur donnait des bases théoriques vraiment solides. J'espère que cela ne va pas changer.

Dans ce *Hors-Série*, Jean-Paul Delahaye nous rappelle que les algorithmes ne sont pas encore capables de créativité mathématique.

Ils ne sont pas capables de créativité tout court : ni en maths, ni en biologie, ni en littérature. On y travaille, on fait des progrès, mais les poèmes que nos algorithmes créent sont encore médiocres. Ce qu'on sait faire, c'est donner plein d'exemples de beaux tableaux d'un peintre à une machine, et lui demander de produire une œuvre dans le même genre. Elle ne crée pas vraiment, elle singe. D'ailleurs, on retrouve la même difficulté de définition qu'avec l'intelligence, je ne sais pas définir formellement la beauté ou la créativité. À ce sujet, les travaux du jeune chercheur en IA Antoine Cully dans sa thèse m'ont passionné. Il montre, par exemple, comment un robot à six pattes a pu inventer une nouvelle façon de marcher avec une patte abîmée ou manquante. Mais ce robot a-t-il vraiment découvert une nouvelle façon de marcher ? Ou cette nouvelle démarche était-elle plus ou moins déjà inscrite dans tous les calculs qu'on lui avait demandés avant ?

Sauriez-vous développer un algorithme de bêtise artificielle ?

Vous ne trouvez pas qu'il y a assez de bêtise naturelle ?

Imaginez qu'il n'y en ait pas autour de nous et qu'on ait besoin d'une machine bête pour nous divertir.

S'il s'agit de programmer un générateur de formules fausses, je peux facilement le faire. Mais si vous voulez en plus qu'elles soient drôles, c'est de l'humour. Là, c'est encore plus dur que la créativité.

L'intelligence artificielle est sortie des laboratoires, elle est entrée dans la cité, et elle y produit des effets. Lesquels vous paraissent les plus importants ?

Il y a deux questions qui me semblent particulièrement critiques en ce moment, et elles sont liées : c'est la sobriété énergétique et le travail. Selon les sources, le numérique représenterait aujourd'hui de 3 à 4 % des émissions de gaz à effet de serre dans le monde, et cela croît. Je ne sais pas chiffrer la proportion de l'intelligence artificielle dedans. Ce n'est pas énorme, mais cela augmente aussi. Pour limiter notre impact sur l'environnement, il va nous falloir changer nos modes de vie,

Aujourd'hui partenaire de l'homme (ici, la programmation d'un bras robot), la machine intelligente pourrait, demain, capter une partie importante de l'emploi. Une catastrophe… ou bien une chance ?

par exemple arrêter de changer de téléphone tous les deux ans ou de passer du temps à visionner des films en haute résolution sur un téléphone cellulaire. Il y a beaucoup de gaspillage, comme avec les « chaînes de blocs » *(blockchains)*, ces procédés de stockage sécurisés et décentralisés, qui pourraient fonctionner en consommant plusieurs ordres de grandeur d'énergie en moins pour le même résultat. Dans le numérique comme pour le reste, il va nous falloir apprendre à être frugaux.

Et concernant le monde du travail ?

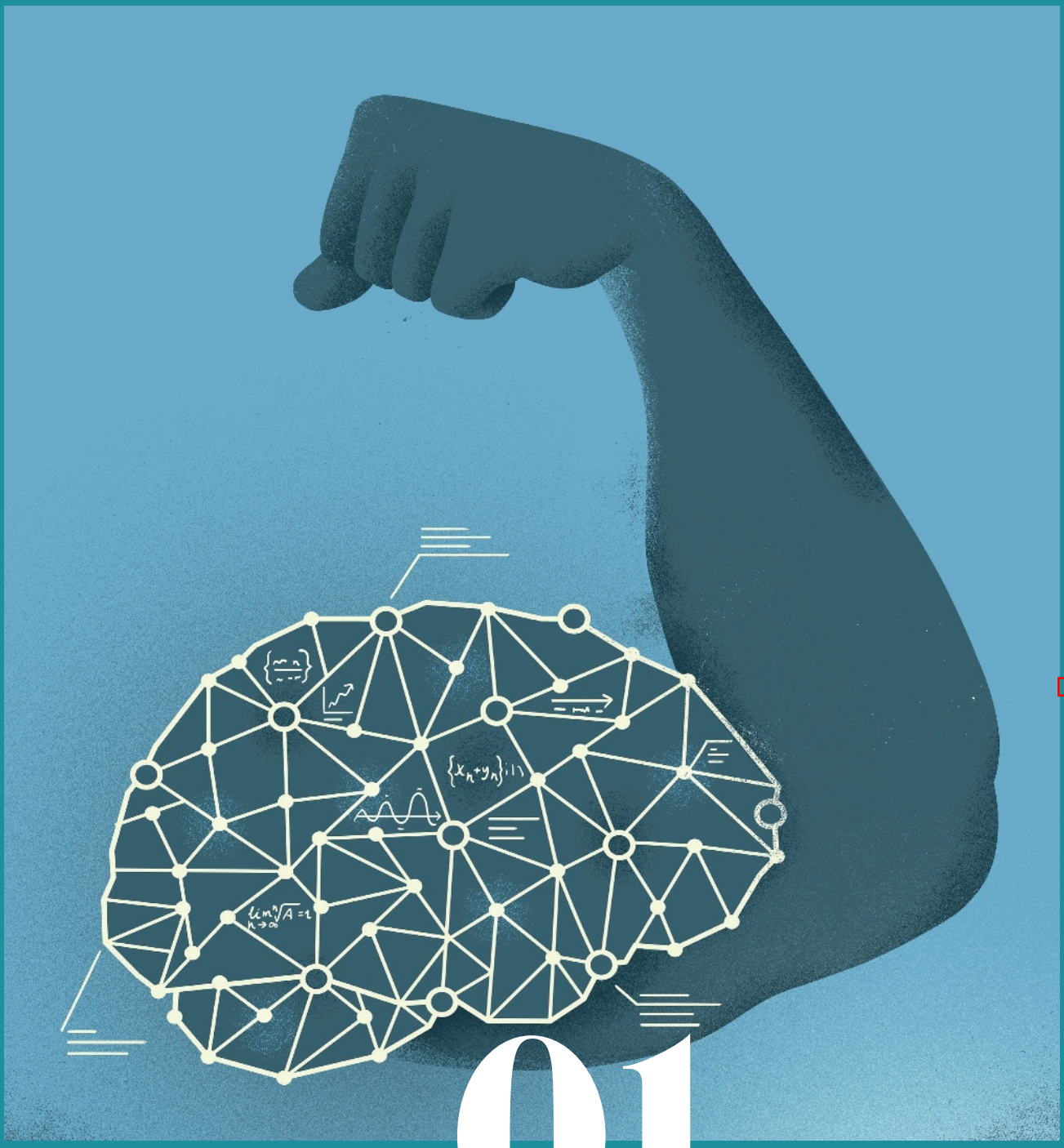
L'ensemble de la technologie numérique a une incidence sur l'emploi, pas seulement l'intelligence artificielle. Aujourd'hui, on peut faire fonctionner une usine avec très peu d'individus grâce à l'informatique en général. Il est vrai qu'avec l'intelligence artificielle on va aller encore plus loin dans le remplacement de l'humain. Après sa force physique, son travail intellectuel devient de plus en plus remplaçable. Le hiatus est qu'on veut une société plus sobre énergétiquement, qui produise et pollue moins, et qu'on veut aussi moins travailler, donc utiliser plus de machines. Or il faut de l'énergie pour fabriquer les machines et elles ont des rendements souvent moins bons que les nôtres. Pour y arriver, de sérieuses avancées scientifiques et d'importantes mesures d'économie seront nécessaires. Et puis, si les machines remplacent les humains, qui va être rétribué ? Uniquement ceux qui les possèdent ? Dans ce cas, la grande masse de la population sera non seulement privée d'activité, mais aussi de quoi se nourrir. Ce ne sera pas socialement tenable. Dans les vingt à cinquante ans à venir, une transformation complète de la société s'imposera, exigeant que l'économie soit beaucoup plus redistributive. Voilà pour le volet sociétal, qui se double d'un volet humain. Dans notre culture, on nous apprend dès l'enfance que le travail est la grande valeur. Comment fera-t-on dans un monde où une grande partie d'entre nous sera sans emploi, ou avec de l'emploi partiel, ou des travaux associatifs ou d'aide à la personne, non « productifs » dans le sens économique actuel ? Il nous faut inventer une nouvelle philosophie du travail, de son utilité sociale, une nouvelle philosophie des loisirs. Le côté génial, c'est que, si on ne se plante pas écologiquement ou socialement, l'informatique nous permet d'avoir l'ambition la plus dingue, celle d'une société égalitaire où tout le monde vivrait bien, en s'éduquant, avec autant de loisir que souhaité. Ce n'est pas de la science-fiction… Enfin, je l'espère.

Propos recueillis par Olivier Voizeux

© Didgeon Oudomsi/shutterstock.com

Prouesses et promesses

On a longtemps pensé qu'une machine serait intelligente le jour où elle battrait le champion du monde d'échecs. C'est arrivé en 1997, et depuis, « l'intelligence artificielle », qui se distingue mal de l'informatique plus « ordinaire », irrigue nos ordinateurs, nos téléphones et nos réseaux. Mieux : depuis les années 2010, une approche ancienne, mais un peu délaissée, celle du réseau de neurones numériques, est venue à bout de problèmes jusque-là sans solution. Ainsi les records se sont-ils enchaînés : traduction automatique, prédiction des formes tridimensionnelles des protéines, analyse d'images médicales…



© Sébastien Thibault/Pour la science

01

Aider au diagnostic, au pronostic et au traitement : les objectifs de l'intelligence artificielle appliquée à l'imagerie médicale ne sont pas minces. Après des premiers succès spectaculaires mais coûteux, la recherche sur les algorithmes vise désormais à établir un « jumeau numérique » du patient.

La fée IA au chevet des malades

Nicholas Ayache

© Have a nice day photo/Shutterstock

Face Shi
Protective Isolation

Pour l'imagerie médicale, l'informatique n'est pas exactement une nouveauté. On peut même affirmer qu'il y a des algorithmes depuis qu'il y a de l'imagerie numérique car, sans eux, il serait impossible d'exploiter les signaux originaux fournis par les dispositifs, IRM et autres, qui échappent totalement à nos sens et avec lesquels sont construites les images du corps humain. Toutefois, depuis quatre à cinq ans, un nouveau pas a été franchi dans le traitement informatique des images médicales numériques avec l'avènement des algorithmes d'apprentissage profond, qui utilisent, avec succès, des masses de données considérables qu'aucun cerveau humain n'est capable d'appréhender en totalité.

Des exemples récents ont frappé les esprits. En dermatologie, un logiciel d'apprentissage profond, préalablement entraîné sur plus de 1 million d'images générales annotées, a été ajusté sur les images de 130 000 anomalies dermatologiques pour apprendre à distinguer automatiquement les mélanomes des lésions bénignes avec l'expertise d'un dermatologue. En radiologie, l'entreprise française Therapixel a soumis son logiciel aux 640 000 mammographies proposées lors d'une compétition mondiale et a remporté celle-ci en repérant mieux que tous ses concurrents les images suspectes. Ce programme se mesure aujourd'hui aux radiologues experts. En pneumologie, un logiciel développé par Google, entraîné sur des milliers de scanners et testé sur plus de 1 000 nouvelles images, est capable de détecter la présence de nodules suspects dans les scanners volumiques des poumons, et de les caractériser aussi bien qu'un professionnel.

En bref

> Algorithmes d'apprentissage profond et vastes bases de données d'images médicales offrent à l'IA une expertise comparable à celle de l'humain pour détecter certaines pathologies.

> Ce savoir-faire est néanmoins très spécialisé et coûteux à acquérir.

> La recherche en IA travaille aussi à établir des « jumeaux numériques » du patient, permettant un pronostic et une thérapie personnalisés.

> Un large partage des données et une formation de tous les professionnels de santé à l'IA sont nécessaires pour faire progresser la discipline.

Enfin, en ophtalmologie, d'autres logiciels d'apprentissage, entraînés sur plus de 130 000 images de la rétine, identifient les rétinopathies diabétiques avec la fiabilité d'un ophtalmologue. C'est le cas d'IDx-DR, approuvé en 2018 par la FDA, l'agence américaine du médicament, pour faire ce diagnostic automatiquement sans que l'image ne soit vue par un spécialiste. Depuis, plus d'une douzaine de systèmes d'imagerie assistés par l'IA ont reçu ce label.

UN DOUBLE NUMÉRIQUE POUR CHAQUE PATIENT

Bien que ces résultats soient spectaculaires, ils restent pour l'instant confinés à des tâches précises, relativement étroites, pour lesquelles d'immenses bases de données renseignées par des spécialistes sont disponibles. Or ces bases ne sont pas légion, leur acquisition et surtout leur analyse par des experts coûtent très cher *(voir l'encadré page ci-contre)*. De plus, en cas de doute, les algorithmes d'apprentissage profond sont caractérisés par des millions de paramètres opaques à l'interprétation. Ils sont aussi sensibles aux attaques contradictoires. En 2019, une étude a révélé qu'ajoutées dans les images, de petites perturbations totalement invisibles à l'œil nu mais calculées très précisément bouleversaient les résultats *(voir Un seul pixel vous manque, et tout est désappris, par J.-P. Delahaye, page 60)*. Dans ces conditions, une lésion de la peau initialement classée bénigne avec 100 % de confiance devient maligne avec le même taux de confiance, alors qu'à l'œil nu la différence est imperceptible. L'inverse est aussi envisageable,

et plus grave, puisqu'une base piratée pourrait conduire à l'absence de traitement de cancers.

Comment rendre les algorithmes d'IA plus robustes ? Une piste consiste à les combiner avec des modèles numériques du patient, ces derniers exploitant les connaissances générales que nous avons en anatomie et en physiologie, mais en se limitant à un nombre réduit de paramètres plus facilement interprétables. Plus précisément, des algorithmes et des modèles mathématiques, physiques, et biologiques du vivant permettent aujourd'hui de construire, à partir d'images médicales et d'autres données qui lui sont propres, une représentation numérique et personnalisée du patient, un « patient numérique personnalisé », « e-patient », ou « jumeau numérique ». Les paramètres et les modèles de cette réplique sont ensuite utilisables par des algorithmes de médecine numérique conçus pour l'aide au diagnostic, au pronostic et à la thérapie.

Les modèles biophysiques du patient numérique aident aussi à enrichir les bases de données en simulant, par exemple, des maladies rares ou des cas peu ou mal représentés parmi les malades répertoriés. Des simulations intensives réalisées *a priori* sur des masses d'informations variées sont utiles pour accélérer les calculs scientifiques sur de nouveaux cas. Les approches statistiques de l'apprentissage automatique sont donc complémentaires d'approches plus déterministes.

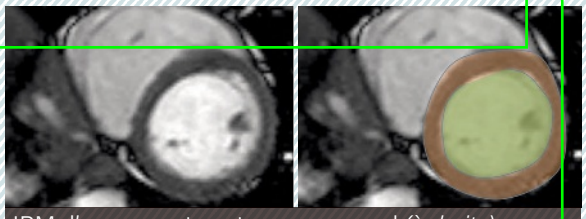
Enfin, le cadre théorique du patient numérique et les méthodes modernes de traitement des mégadonnées favorisent l'intégration de données qu'on peut qualifier d'holistiques sur le patient, incluant non seulement ses images anatomiques et fonctionnelles, mais aussi des informations biologiques très complètes (génétiques, métabolomiques…), voire comportementales et même environnementales selon le style de vie (on parle alors d'« exposome »).

Analyser l'ensemble de ces renseignements très hétérogènes et de grandes dimensions est un défi scientifique, mais la recherche est très active sur ces sujets, notamment dans le cadre des 3IA (instituts interdisciplinaires d'intelligence artificielle) déployés en France. L'objectif est de faire progresser la compréhension des maladies et de mieux guider leur prise en charge, notamment le dépistage et le traitement de cancers comme ceux du poumon et du sein, ainsi que de maladies neurodégénératives.

À terme, le champ d'application de l'intelligence artificielle en imagerie médicale est amené à s'étendre à toutes les disciplines médicales (radiologie, cardiologie, oncologie, neurologie, radiothérapie, chirurgie…). En conséquence, les médecins et, au-delà, tous les professionnels de santé, notamment les pharmaciens, le personnel médical… doivent être formés à l'IA, afin de

Quelques précautions s'imposent

Les réseaux de neurones utilisés en imagerie médicale ont besoin d'ingurgiter des dizaines, voire des centaines de milliers de scanners ou d'IRM préalablement préparés. Des dizaines de spécialistes doivent les annoter à la main, en précisant s'il y a ou non une pathologie (annotation faible), voire en contourant les éléments pathologiques (annotation forte, comme ici un ventricule gauche). Cette phase de préparation, puis d'apprentissage est très coûteuse, car elle nécessite une main-d'œuvre hautement qualifiée et une forte puissance de calcul. Une fois correctement entraînés, les algorithmes détectent automatiquement des pathologies dans de nouvelles images… à condition de veiller à certaines précautions : s'assurer de la bonne représentativité des cas rares (qui par définition ne sont pas nombreux !), de l'absence de biais (genre, âge, ethnicité…), ou de la présence de sujets sains en nombre suffisant (ce qui pose un problème éthique lorsque la modalité d'imagerie est irradiante par exemple).



IRM d'un cœur et contourage manuel *(à droite)* du ventricule gauche.

> Les algorithmes sont sensibles aux attaques contradictoires. Une perturbation peut faire passer une lésion classée 100 % bénigne en 100 % maligne

mieux comprendre son utilité et ses limites. Et déjà des 3IA développent des diplômes universitaires qui y sont consacrés (c'est le cas de celui de la Côte d'Azur, à Nice-Sophia Antipolis).

RASSEMBLER LES DONNÉES

Un partage plus large de l'information est aussi nécessaire. Au Royaume-Uni, la base de données UK-Biobank met à la disposition de la communauté scientifique l'ensemble des informations (images, analyses génétiques et paramètres environnementaux) acquises progressivement sur plus de 100 000 participants volontaires suivis dans le temps. Aux États-Unis, la base de données ADNI propose, pour une immense cohorte de patients Alzheimer, des images anonymisées de leur cerveau, ainsi que des données cliniques et génétiques.

La France se lance à son tour : Health Data Hub (HDH), Entrepôt des données de santé de l'AP-HP (EDS), France Life Imaging, Drim France IA, etc., autant d'initiatives pour rassembler des données de santé associant imagerie et informations complémentaires sur des patients, et pour les mettre à la disposition des chercheurs en préservant bien sûr la confidentialité et avec l'accord des participants. Des projets indispensables pour espérer des progrès, possiblement considérables, dans la recherche, et pour cela nous devons notamment veiller à ce que la réglementation, tout en protégeant suffisamment les patients, ne pénalise pas les chercheurs français et européens face au reste du monde.

L'un des moyens d'y parvenir consiste à développer l'apprentissage fédéré. Les données restent dans les centres qui les ont produites, et ce sont les algorithmes d'apprentissage qui sont exécutés sur les ordinateurs de chaque centre et qui partagent entre eux les apprentissages locaux pour fournir un résultat global. Un tel projet d'apprentissage fédéré pour l'imagerie de médecine nucléaire est mené par nos 3IA en partenariat avec plusieurs centres de lutte contre le cancer, parmi lesquels le centre Antoine-Lacassagne à Nice, l'institut Curie, à Paris, et le centre Henri-Becquerel, à Rouen.

En définitive, intelligence artificielle, imagerie médicale et patient numérique forment un ensemble d'outils informatiques au service d'une médecine aux « 4 P » : plus personnalisée, précise, prédictive et préventive. Ces outils numériques feraient gagner beaucoup de temps au thérapeute pour mieux prendre en charge son patient. Ils assisteraient également des généralistes en leur proposant d'orienter les patients, lorsque cela est nécessaire, vers un spécialiste. Ces nouveaux outils ont vocation à seconder le médecin, pas à le remplacer.

— L'auteur —

> Nicholas Ayache est directeur de recherche à l'Inria et directeur scientifique de l'Institut interdisciplinaire d'intelligence artificielle (3IA) Côte d'Azur.

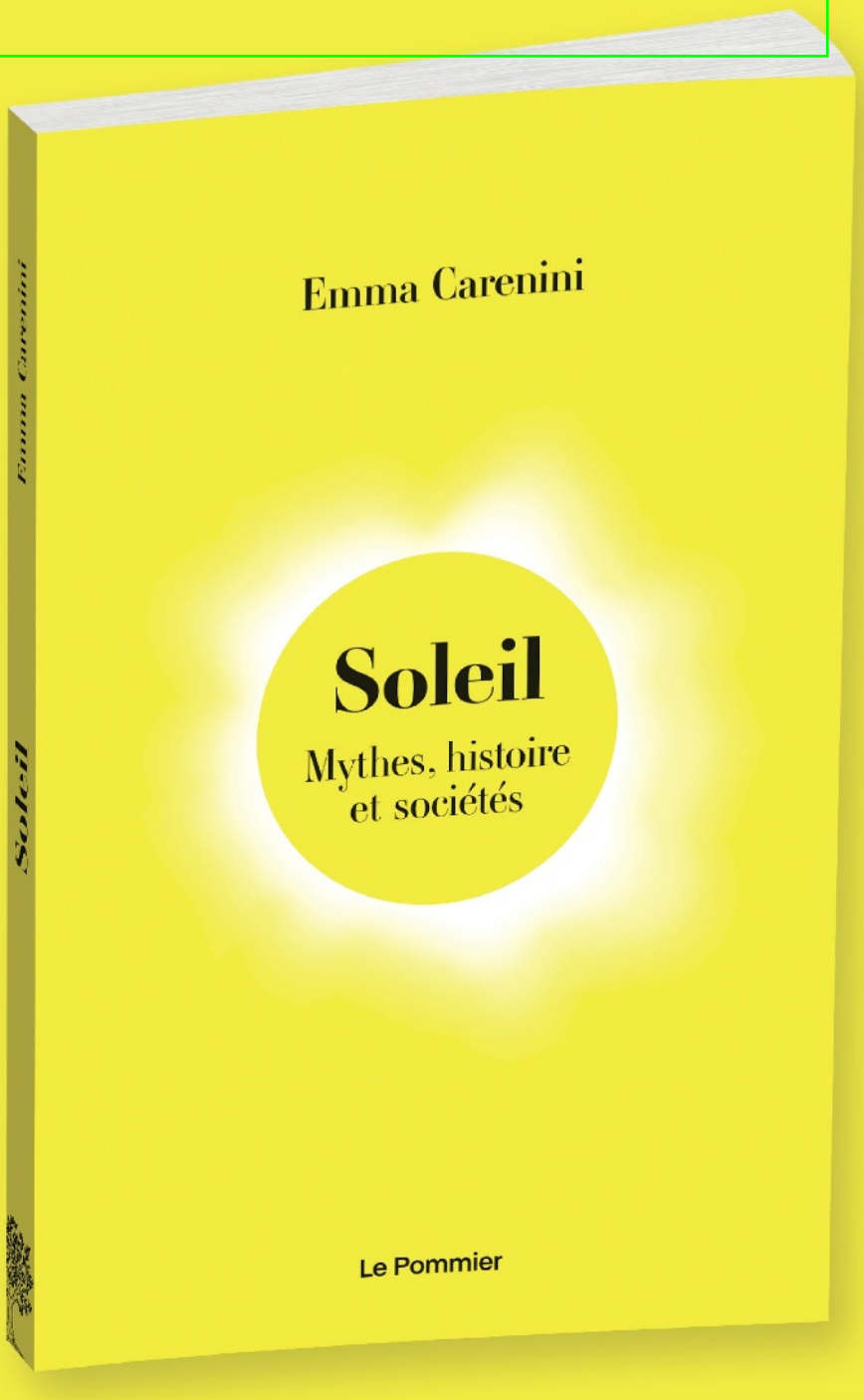
— À lire —

> N. Ayache *et al.,* *Nouvelle enquête sur l'intelligence artificielle*, Champs Actuel, Flammarion, 2020.

> S. G. Finlayson *et al.,* Adversarial attacks on medical machine learning, *Science*, 2019.

> N. Ayache, Des images médicales au patient numérique, *Leçon inaugurale du Collège de France, chaire annuelle 2014*. https://bit.ly/3pSrnin

Du cadran solaire à la fusion nucléaire



En librairie

Éditions Le Pommier

La traduction automatique doit son succès aux statistiques

Thierry Poibeau
est directeur de recherche au CNRS et directeur adjoint du laboratoire Lattice (Langues, textes, traitements informatiques et cognition).

Longtemps tournés en dérision, les systèmes de traduction automatique ont accompli d'incroyables progrès en quelques années seulement. Leur secret ? Puissance de calcul et statistiques.

La traduction automatique est-elle un domaine de recherche ancien ? Ou son essor est-il récent, lié à celui d'internet ?

Les deux sont vrais. La traduction automatique s'est énormément développée ces dernières années, et je vous expliquerai pourquoi. Mais l'idée d'automatiser le processus de traduction a toujours fasciné. Des philosophes comme Descartes et Leibniz s'interrogeaient déjà sur la possibilité de formaliser le langage pour le dénuer de toute ambiguïté, et donc le rendre facilement traduisible. Pendant la Seconde Guerre mondiale, grâce au déchiffrement automatisé des messages cryptés allemands, les Alliés ont développé des connaissances quantifiées sur les langues, l'organisation ou la fréquence d'emploi de telle ou telle lettre. Après 1945, les pionniers de l'informatique, qui avaient souvent eux-mêmes travaillé pour la défense, ont imaginé appliquer les mêmes techniques à la traduction, en faisant des comptages sur les mots, et non plus sur les lettres. D'autant qu'avec le début de la guerre froide, les Américains se sont penchés sur les grandes masses de textes en russe, concernant par exemple les avancées scientifiques et techniques des Soviétiques, avec l'idée de les indexer et d'y chercher des informations. Les besoins en traduction automatique sont devenus évidents.

Sur quelles difficultés a-t-on buté ?

Le problème, bien connu, est celui de l'ambiguïté du langage naturel : il y a plusieurs sens par mot. Et beaucoup de mots, de l'ordre de 50 000 dans un dictionnaire usuel du français, 400 000 quand on prend toutes les formes possibles des mots, et jusqu'à 1 million ou plus si on y ajoute les noms propres, les mots composés, les termes techniques… L'ambiguïté existe aussi au niveau des phrases. Si vous écrivez : « J'ai vu Clara avec les jumelles », voulez-vous dire que vous avez regardé Clara à l'aide de jumelles ? Ou bien que c'était elle qui portait des jumelles ? Ou bien encore qu'elle était accompagnée de… deux filles ? La somme des possibilités conduit à une explosion combinatoire qu'à l'époque il aurait fallu coder à la main, sans recours à l'électronique : c'était absolument impossible. En réalité, pour arriver à déchiffrer les mots, il faut percevoir l'intention d'emploi derrière, laquelle ne se trouve pas dans le texte, mais dans le contexte. Et le contexte, l'ordinateur ne sait pas ce que c'est. Pour le prédire, il lui serait nécessaire d'avoir, sur le monde, des connaissances de sens commun, ce qu'il n'a pas.

Qu'entendez-vous par « connaissance de sens commun » ?

Ce n'est pas simple de répondre, car les chercheurs en discutent sans toujours le définir. Je vous donne deux exemples : « On peut manger

des fraises » et « S'il y a un feu rouge, je dois arrêter ma voiture ». Beaucoup de connaissances de sens commun sont acquises par l'expérience, et varient selon la culture : ainsi se repérer dans l'espace peut se faire avec une boussole ou bien en regardant le Soleil. Même si votre culture ignore le sens du feu rouge, vous présumez que la présence de cet objet a une signification : voilà une autre connaissance de sens commun. Si vous croisez un fruit dont vous ignorez tout, la première question que vous allez vous poser est : est-ce comestible ? Vous allez avoir une interrogation par rapport à l'objet. Tout cela échappe à l'ordinateur. En revanche, on peut alimenter les bases de connaissances utilisables par les ordinateurs avec les milliards de textes disponibles sur le web. Une bonne partie des connaissances que nous acquérons au contact du monde, l'ordinateur va les trouver dans ce corpus, comme la proximité fréquente de « manger » et « fraises » (ce qu'on appelle des « cooccurrences ») qui lui permettra d'établir qu'une fraise est comestible. Le lien entre feu rouge, voiture et arrêt est plus difficile à déduire, mais pas impossible. L'ordinateur n'aura pas une connaissance formalisée du lien entre ces différents éléments, mais l'apparition simultanée et fréquente de ces mots dans des phrases permet malgré tout d'inférer qu'il y a un lien entre eux.

Comment l'ordinateur regarde-t-il une langue ? Comme un ensemble de symboles ?

On a longtemps pensé que les langues étaient des systèmes symboliques, et que traduire revenait à effectuer des calculs sur des symboles. Au nom de quoi on a tenté de nourrir la machine avec des règles qui correspondent en partie à ce qu'on apprend à l'école : une phrase se compose d'un groupe nominal et d'un groupe verbal, etc. (ce qui revient à modéliser la syntaxe). Mais ce qui convient pour la syntaxe est très difficile à transposer à la sémantique. Comment modéliser des règles permettant de déterminer *a priori* le sens d'un mot dans une phrase ? C'est là que les statistiques deviennent utiles. Le linguiste américain George Zipf avait remarqué, dès les années 1930, qu'en français, par exemple, il y a des mots comme les déterminants qu'on trouve dans toutes les phrases, et d'autres, très rares, susceptibles de n'apparaître qu'une seule fois même dans un texte long. C'est aussi vrai au niveau de l'usage des mots : la plupart ont un sens fréquent (la fraise comme fruit) et d'autres plus rares (pensez à la fraise de veau). Par défaut, il vaut mieux considérer le premier usage, beaucoup plus fréquent.

Que faire de ces informations statistiques sur les mots ?

À condition de lui fournir des millions, voire des milliards d'exemples d'usage, l'ordinateur va déterminer le sens des mots en contexte. Pour reprendre l'exemple précédent, « J'ai vu Clara avec les jumelles », du fait de l'apparition conjointe fréquente de « voir » et de « jumelles », il va inférer que « jumelles » est sans doute rattaché au verbe : ce serait l'interprétation la plus probable, même si l'autre reste possible. Au-delà, les traitements sont fondés sur une hypothèse aussi simple que puissante : les mots employés dans les mêmes contextes ont des sens proches. S'il y a un trou dans une phrase, l'ordinateur saura fournir l'ensemble des possibilités qui conviendraient pour la compléter, en proposant une liste de mots ordonnés, du plus ou moins probable. Il a ainsi une vision complète du lexique, des proximités sémantiques entre mots, et des liens possibles entre les mots (entre « voir » et « jumelles », par exemple).

La traduction automatique doit donc son succès aux statistiques ?

Oui, et plus largement à deux choses : à la masse de textes aujourd'hui disponibles, sur le web essentiellement, et à la puissance de calcul de l'ordinateur. Désormais, il n'y a plus vraiment de règles au sens où on l'entend dans la grammaire traditionnelle, mais des milliards de contraintes qui s'appliquent sur les mots, comme on l'a vu ou, quand on est dans un cadre de traduction automatique, sur les liens entre mots d'une langue à l'autre. Étant donné l'énorme masse de données disponibles pour l'entraînement, c'est-à-dire pour la mise au point du modèle de traduction, l'ordinateur traduit le plus souvent sans trop d'erreurs, même dans le cadre d'applications grand public où on a affaire à une grande variété de couples de langues, et à des textes de différente nature, sur n'importe quel sujet.

> On a longtemps pensé que les langues étaient des systèmes symboliques, et que traduire revenait à effectuer des calculs sur des symboles

Certaines langues sont-elles mieux traduites que d'autres ?

La plupart des traducteurs automatiques en ligne grand public proposent une centaine de langues, qui sont toutes des langues écrites (sur les quelque 4000 qui disposent d'une forme d'écriture plus ou moins développée). Celles pour lesquelles la traduction fonctionne le mieux, une vingtaine, sont celles pour lesquelles on a le plus de données – la taille du corpus est le premier critère. Et la traduction est meilleure si l'anglais est la langue source ou la langue cible. Cette langue présente deux avantages : la masse des données disponibles, mais aussi sa faible complexité morphologique, c'est-à-dire que la forme des mots change très peu : au présent simple le verbe reçoit un *s* à la troisième personne du singulier, le passé se forme souvent en *ed*, etc., mais c'est peu, comparé au français. En allemand, la forme du mot varie selon sa fonction dans la phrase, ce qui veut dire qu'il faut déterminer sa fonction syntaxique avant de traduire. D'autres langues agrègent encore plus d'informations sur une seule forme graphique, par exemple, en finnois, le nombre, le possessif, etc. Produire la bonne forme est dès lors très difficile, alors qu'en anglais beaucoup d'informations passent par des « petits » mots isolés (préposition, déterminant, possessif) : vous pouvez tout décomposer, et la traduction mot à mot ne marche pas si mal.

Quel est l'avenir de la traduction automatique ?

Pour la recherche, l'un des enjeux importants est de mieux traduire les langues pour lesquelles on a le moins de données, voire celles pour lesquelles on n'a même pas deux corpus qui peuvent être « alignés » au niveau des phrases, c'est-à-dire pour chaque phrase dans la langue source, une phrase en regard dans la langue cible, afin d'acquérir les connaissances permettant de passer d'une langue à l'autre. Pour les opérateurs de télécommunication, l'enjeu est plus économique : il s'agit de traduire de l'oral vers l'oral afin de pouvoir converser de manière transparente, même sans connaître la langue de l'autre. En pratique, la parole est d'abord transcrite en texte, lequel est traduit, puis retranscrit à l'oral. Les performances s'améliorent et on peut imaginer que, dans quelques années, on utilisera ce type de technologie dans nos systèmes de communication à distance, en visio.

Est-ce la fin du métier de traducteur ?

Il est sûr que le métier change. Par exemple, les sites d'information sur le web proposent désormais des vidéos dont les sous-titres sont souvent traduits automatiquement. Le résultat de la traduction par le programme est juste retouché au minimum lors d'une phase de post-édition : le traducteur relit le texte produit, regarde l'original et ne corrige que les contresens. Tant pis si, esthétiquement, le français n'est pas idéal. Même si la traduction automatique va ouvrir de nouveaux marchés (pour les revues de presse, le journalisme ou l'édition par exemple), il est clair qu'elle propose un service plus rapide et moins coûteux, qui conduit dans certains cas à des remplacements purs et simples.

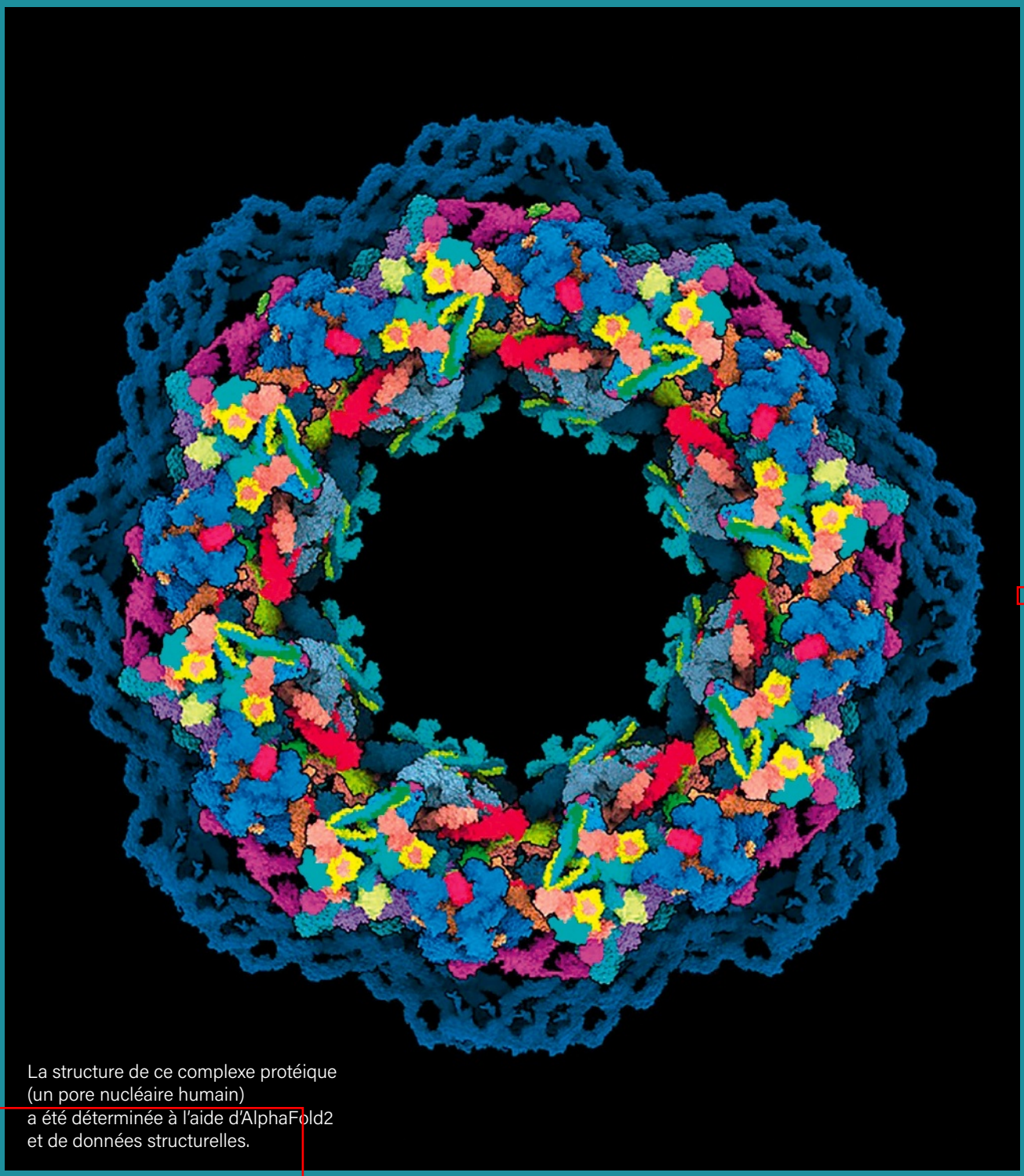
Propos recueillis par Olivier Voizeux

Les algorithmes d'apprentissage profond prédisent désormais la forme 3D d'une protéine à partir de sa séquence linéaire. Une révolution pour la biologie !

Les protéines, nouveau terrain de jeu de l'IA

Michael Eisenstein

© Agnieszka Obarska-Kosinska



La structure de ce complexe protéique (un pore nucléaire humain) a été déterminée à l'aide d'AlphaFold2 et de données structurelles.

«L'un des plus grands mystères de la biologie est en grande partie résolu par l'IA», pour la BBC. La «réalisation la plus importante de l'IA», selon Forbes. Rarement un logiciel scientifique aura fait l'objet de titres aussi sensationnels. De quoi parle-t-on ? D'un nouveau jeu vidéo ? Non, du programme AlphaFold2 mis au point par DeepMind, la branche de Google consacrée à l'intelligence artificielle, et dédié à l'un des problèmes les plus épineux de la biologie : prédire la structure tridimensionnelle, et donc fonctionnelle, d'une protéine à partir de sa séquence linéaire en acides aminés. L'association américaine pour l'avancement des sciences en fait une des «percées de l'année 2021».

Auparavant, les chercheurs déterminaient la structure 3D des protéines grâce à des techniques expérimentales lourdes, comme la cristallographie à rayons X et la cryomicroscopie électronique (cryo-EM). Les informaticiens ont fait des progrès constants en développant des logiciels qui complètent ces méthodes, mais malgré ces avancées, on ne disposait toujours pas d'informations structurelles pour environ 4800 protéines humaines : AlphaFold2 a réduit ce nombre à 29…

Le programme a été dévoilé en novembre 2020 lors de la 14e édition d'une compétition biennale, le CASP *(Critical assessment of structure prediction)*. L'objectif : mettre au défi biologistes et algorithmes avec des séquences de protéines dont les structures ont été déterminées expérimentalement sans être rendues publiques. Le logiciel de DeepMind a survolé la compétition.

«Au CASP14, il a livré la structure d'environ deux tiers des protéines proposées, dont certaines difficiles, avec une précision de type expérimentale», explique Yang Zhang, de l'université du Michigan, dont l'algorithme figurait parmi les finalistes. Depuis, plusieurs publications ont confirmé l'efficacité d'AlphaFold2. Le biologiste y voit une preuve éclatante de la puissance de l'apprentissage profond, mais une solution encore partielle au problème du repliement des protéines, car certains aspects restent inaccessibles. Comment fonctionnent ces algorithmes ?

En bref

> La fonction d'une protéine, un enchaînement d'acides aminés, dépend de sa structure tridimensionnelle.

> Pour la déterminer, on a longtemps eu recours à des études expérimentales complexes, comme la cristallographie à rayons X et la cryomicroscopie électronique.

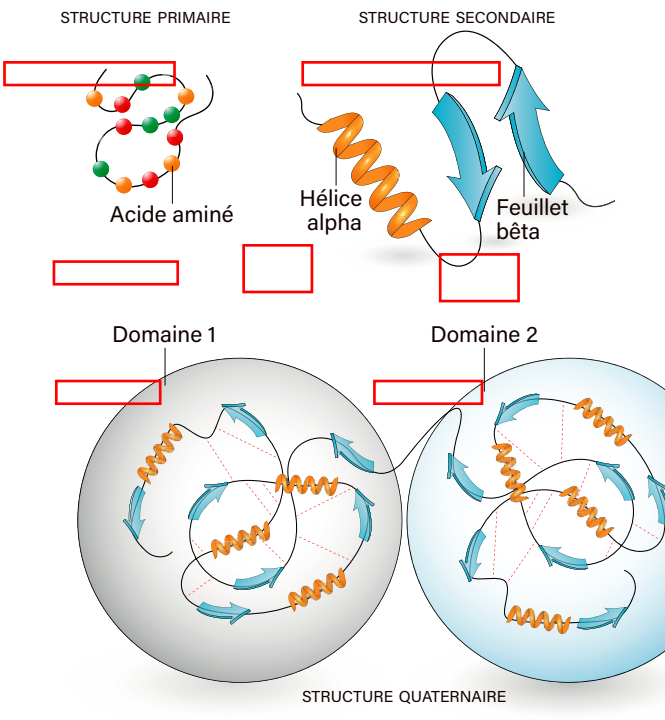
> Depuis peu, l'intelligence artificielle fondée sur l'apprentissage profond a révolutionné la discipline.

> Le logiciel AlphaFold2, conçu par Google, est à ce jour le programme le plus efficace, mais quelques obstacles, notamment la dynamique des protéines, restent à franchir.

«ENSEIGNEMENT SUPÉRIEUR»

Le processus intègre des phases d'apprentissage automatique durant lesquelles des réseaux neuronaux sont formés pour reconnaître et interpréter des modèles dans les données. «Ces modèles n'essaient pas de prédire la structure en une seule fois, explique David Baker, de l'université de Washington, à Seattle. Ils s'apparentent davantage à une simulation physique où les modèles apprennent à faire les bons mouvements pour améliorer la structure.» Entraînés avec de grandes quantités de données expérimentales annotées, ces algorithmes identifient progressivement des liens entre la séquence et la structure.

Ces cinq dernières années, les progrès ont été importants. La première itération d'AlphaFold a remporté le CASP13, en 2018, mais ses performances étaient loin de celles de son successeur en 2020. Plusieurs laboratoires universitaires ont par la suite développé des programmes basés sur l'apprentissage profond surpassant la première génération d'AlphaFold, notamment D-I-TASSER du laboratoire de Yang Zhang, trRosetta du groupe de David Baker, et RaptorX, développé par Jinbo Xu et son équipe de l'institut de technologie Toyota, à Chicago.



Une protéine est un enchaînement linéaire d'acides aminés qui se replie selon plusieurs degrés de complexité jusqu'à obtenir la configuration propre à sa fonction.

Mais ces algorithmes fonctionnaient le plus souvent au sein de logiciels plus larges, ce qui créait des risques d'erreurs. «Ils étaient en outre fragmentés, et les différents constituants communiquaient souvent mal», explique Mohammed AlQuraishi, de l'université Columbia, à New York. Ces difficultés sont levées avec des algorithmes dédiés qui gèrent l'ensemble du processus, de la séquence à la structure. C'est pourquoi John Jumper, le chercheur principal de DeepMind, à Londres, a abandonné AlphaFold après le CASP13 pour se consacrer à AlphaFold2.

L'ÈRE DES TRANSFORMEURS

Plusieurs aspects de ce programme reposent sur des techniques établies. Par exemple, l'algorithme commence par des alignements multiséquences (MSA), où la séquence d'une nouvelle protéine de structure inconnue est comparée à celles de protéines apparentées provenant d'autres espèces. En identifiant les acides aminés qui évoluent en parallèle, les algorithmes repèrent les plus susceptibles de s'associer dans la protéine repliée, en supposant qu'une modification de la séquence en un endroit a nécessité ailleurs une mutation compensatoire pour préserver la structure globale de la protéine.

Chris Sander, de l'institut de cancérologie Dana-Farber, à Boston, et Debora Marks, de l'université Harvard, à Cambridge, avaient mis au point cette technique basée sur la coévolution, en 2011. Aujourd'hui, l'apprentissage automatique la rend plus efficace. Pour construire leurs MSA, les développeurs d'AlphaFold2 ont utilisé les milliards de séquences protéiques compilées par Martin Steinegger, de l'université de Séoul, en Corée du Sud, et Johannes Söding, de l'institut Max-Planck de chimie biophysique, à Göttingen, en Allemagne.

L'équipe de DeepMind a également imaginé des solutions innovantes. L'une d'elles consiste en des transformeurs, des outils de reconnaissance de motifs couramment utilisés dans l'analyse d'images et le traitement du langage naturel. Les transformeurs sont conçus pour repérer des modèles locaux (des chaînes de mots fréquentes ou des éléments visuels souvent associés, par exemple) qui guident l'interprétation des données. DeepMind les a adaptés aux interactions à longue portée entre acides aminés éloignés de la séquence et susceptibles d'être importantes dans la forme finale.

AlphaFold2 traite le repliement des protéines sous plusieurs angles simultanément, et crée en parallèle plusieurs représentations de la structure possible. Celles-ci sont ensuite comparées de façon à affiner le processus de modélisation lors des itérations suivantes. Pour ce faire, John Jumper et ses collègues ont conçu une architecture de réseau neuronal qui favorise un échange d'informations fluide et efficace entre les composants du logiciel.

En raison du décalage entre les débuts publics d'AlphaFold2 et la parution des articles le décrivant, et de l'incertitude des universitaires quant à la diffusion de tous les détails, David Baker et Minkyung Baek, lui aussi de l'université de Washington, ont travaillé à partir d'informations éparses sur l'architecture du logiciel pour

© Art of Science/shutterstock.com

développer leur propre version, RoseTTAFold. Celle-ci utilise de nombreuses stratégies identiques à celles d'AlphaFold2, mais avec quelques particularités.

«Au moment où nous l'avons mis à disposition, c'était de loin le meilleur outil de prédiction de structure, mais il n'est pas aussi bon qu'AlphaFold2», déclare David Baker. Il souligne que DeepMind est une entité privée disposant, contrairement à la plupart des laboratoires universitaires, d'énormes ressources. Le succès d'AlphaFold2 «tient beaucoup à l'argent mis par Google, confirme Amelie Stein, de l'université de Copenhague, mais aussi aux compétences des nombreux experts réunis».

Depuis la sortie d'AlphaFold2 en juillet 2021, les laboratoires se sont rués dessus. Le logiciel est simple d'utilisation, bien qu'il nécessite plusieurs téraoctets d'espace disque pour les bases de données et plusieurs unités de traitement graphique (GPU) pour effectuer les analyses. «Les calculs sur une seule structure sont plutôt faciles», explique Arne Elofsson, de l'université de Stockholm. Mais les analyses de l'ensemble des protéines d'un organisme, ou protéome, sont encore hors de portée de la plupart des laboratoires universitaires.

Pour pallier cette difficulté, Martin Steinegger et ses collègues ont mis au point ColabFold, un système dématérialisé qui fait fonctionner

AlphaFold2 n'est pas encore adapté pour comprendre l'effet des mutations sur la structure des protéines

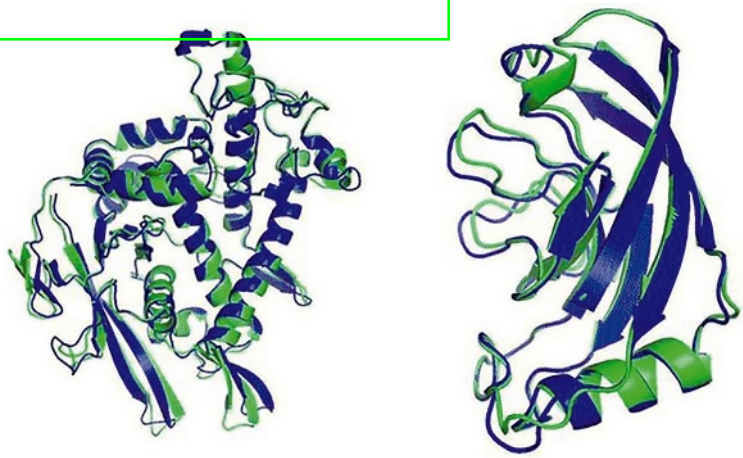
AlphaFold2 et RoseTTAFold en utilisant des bases de données distantes et la puissance de calcul fournie par Google. L'interface web est relativement simple et permet d'optimiser l'expérience.

Même l'équipe de DeepMind a été surprise par les performances d'AlphaFold2 lors du CASP14. Depuis, de nombreuses démonstrations des capacités du logiciel ont été apportées, mais également de ses limites. Dans une étude publiée parallèlement à l'article décrivant l'algorithme, l'équipe de Google a appliqué AlphaFold2 à un ensemble de données comprenant 98,5 % du protéome humain. L'algorithme utilise un test de différence de distance locale prédite (pLDDT, pour *predicted local distance difference test*) pour quantifier sa confiance dans la position et l'orientation qu'il propose pour un acide aminé particulier dans la protéine réelle. Finalement, 36 % de tous les résidus aminés du protéome ont pu être analysés avec un degré de confiance très élevé.

En août 2021, le groupe d'Alfonso Valencia, du Centre de supercalcul de Barcelone, en Espagne, a conclu de son côté qu'AlphaFold2 avait fait passer de 31 à 50 % la proportion d'acides aminés des protéines humaines cartographiables avec précision.

SANS FAUX PLI

Yang Zhang s'attend à ce que le logiciel fasse un travail rapide sur les «fruits les plus faciles à cueillir» du protéome. Mais l'analyse de nombreuses autres protéines relève du défi, notamment celles qui sont composées de plusieurs unités fonctionnelles indépendantes (des domaines) reliées de façon flexible. Ici, les domaines individuels sont



↑ Dans la détermination des structures de protéines, l'intelligence artificielle *(en bleu)* fait désormais presque jeu égal avec les études expérimentales *(en vert)*

© DeepMind

accessibles au logiciel, mais pas leur orientation les uns par rapport aux autres.

Les segments de protéines qui sont naturellement désordonnés dans leur état, qui correspondraient à plus d'un tiers de tous les acides aminés du protéome humain, constituent un défi encore plus grand. Aucun algorithme ne peut actuellement prédire comment ils se replient, mais John Jumper note que la méthode des pLDDT peut aider à délimiter ces segments dans une structure.

Une caractéristique inattendue d'AlphaFold2 et de RoseTTAFold est leur capacité à prédire les structures des homodimères, ces complexes formés de deux protéines identiques, et des hétérodimères, quand les protéines assemblées diffèrent. Arne Elofsson et son équipe ont réussi à modéliser jusqu'à 59 % des complexes à deux protéines qu'ils ont analysés avec AlphaFold2. Cependant, ce processus est plus difficile sur le plan informatique lorsque l'on tente d'identifier des complexes probables à partir de zéro, plutôt qu'en modélisant des paires d'interactions connues. Mais en appliquant plusieurs algorithmes d'apprentissage profond en parallèle, David Baker et son équipe ont identifié et modélisé des centaines de complexes multiprotéiques à partir de millions de paires d'interactions possibles dans le protéome de la levure *Saccharomyces cerevisiae*. Ici, pour filtrer les paires d'acides aminés qui interagissent le plus probablement, RoseTTAFold était environ cent fois plus rapide qu'AlphaFold2, celui-ci ayant ensuite été utilisé sur le sous-ensemble obtenu.

Sentant l'enthousiasme pour ce domaine, DeepMind a publié en octobre 2021 AlphaFold-Multimer, spécifiquement entraîné pour s'attaquer aux complexes protéiques. Le programme a prédit des interactions avec une haute précision pour 34 % des complexes homodimériques testés, et pour 23 % des complexes hétérodimériques.

FRONTIÈRES FONCTIONNELLES

De nombreuses questions restent encore hors de portée, note Debora Marks. « Si votre technologie copie la cristallographie, c'est formidable, mais de tels instantanés structurels statiques ne disent rien du comportement dynamique d'une protéine donnée. » Ainsi, AlphaFold2 produit généralement une seule réponse « correcte » pour chaque séquence. Toutefois, de nombreuses protéines ont plusieurs conformations, toutes pertinentes pour la fonction, par exemple l'état activé ou inhibé d'une enzyme. Pour résoudre ces problèmes, des techniques expérimentales qui montrent la structure de la protéine réelle dans de multiples états, comme la cryo-EM, sont encore nécessaires.

AlphaFold2 n'est pas non plus adapté pour prédire comment les changements d'acides aminés individuels modifient la structure des protéines, un facteur pourtant crucial pour comprendre l'effet des mutations. Une raison en est que l'algorithme converge vers une solution correcte à partir de nombreuses séquences légèrement différentes, explique Amelie Stein. Néanmoins, son équipe a obtenu des prédictions plus précises en couplant l'apprentissage profond à d'autres algorithmes d'analyse des mutations.

Les biologistes structurels ne sont pas près de perdre leur emploi. En fait, ils sont même désormais en mesure de consacrer plus de temps à d'autres questions fondamentales. Randy Read, de l'université de Cambridge, au Royaume-Uni, note, par exemple, que les prédictions d'AlphaFold2 aident déjà beaucoup les cristallographes dans l'interprétation de leurs données lorsqu'elles sont incomplètes. Les concepteurs de protéines *de novo* tireront également profit d'AlphaFold2 en simplifiant la phase de calcul des modèles qui précède les tests en laboratoire.

Pour Mohammed AlQuraishi, la biologie structurale est à l'aube d'une nouvelle ère qui privilégiera la fonction des protéines plutôt que leur forme. Cette dernière ne sera plus une fin en soi, et la discipline va mériter la composante « biologie » de son nom.

— L'auteur —

> **Michael Eisenstein** est journaliste scientifique, basé à Philadelphie.

Cet article a paru sous le titre « Artificial intelligence powers protein-folding predictions » sur le site Nature.com, le 23 novembre 2021.

— À lire —

> **J. Jumper *et al.*,** Highly accurate protein structure prediction with AlphaFold, *Nature*, 2021.

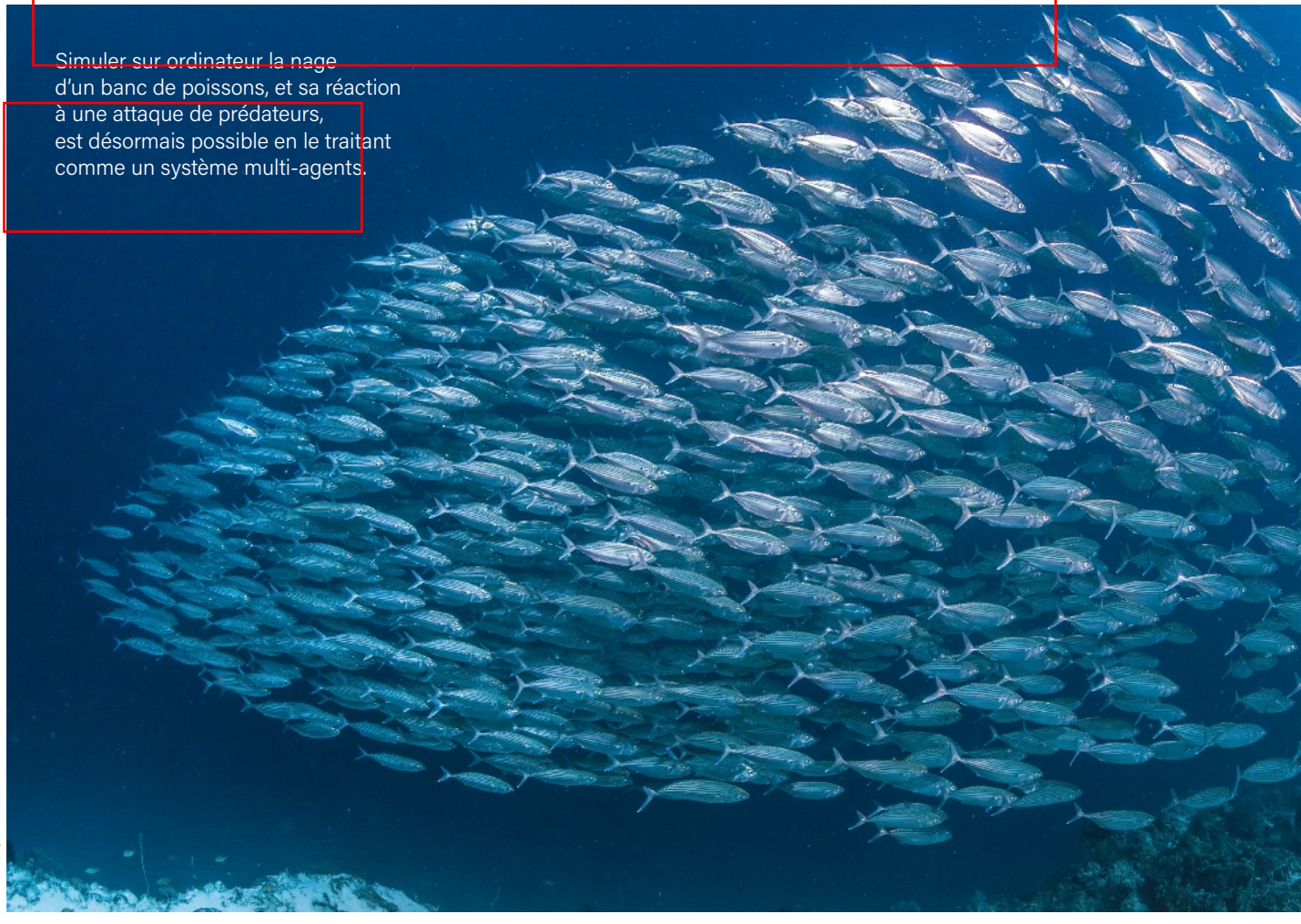
> **K. Tunyasuvunakool *et al.*,** Highly accurate protein structure prediction for the human proteome, *Nature*, 2021.

> **W. Zheng *et al.*,** Protein structure prediction using deep learning distance and hydrogen-bonding restraints in CASP14, *Proteins*, 2021.

> **I. Humphreys *et al.*,** Computed structures of core eukaryotic protein complexes, *Science*, 2021.

Les agents jouent collectif

Simuler sur ordinateur la nage d'un banc de poissons, et sa réaction à une attaque de prédateurs, est désormais possible en le traitant comme un système multi-agents.



© Paul Cowel/shutterstock.com

Programmer un système multi-agents est comme créer une fourmilière : chaque entité possède une petite intelligence, mais leur interaction et leur coordination en engendre une nouvelle forme. Elle est étonnante et parfois d'une efficacité déconcertante.

Philippe Mathieu, Sébastien Picault et Jean-Christophe Routier

Au terme des huit saisons et soixante-treize épisodes de *Game of Thrones*, le spectateur, éberlué, pourrait se livrer à un rapide décompte : combien de cavaliers, de fantassins, de spectres et autres Marcheurs blancs a-t-il vu défiler, combattre, périr ? Heureusement pour la production, depuis la trilogie du *Seigneur des anneaux*, récompensée aux Oscars, les scènes de groupe au cinéma ou dans les séries sont de plus en plus confiées à l'ordinateur, qui dirige des acteurs artificiels grâce à des algorithmes.

En fait, l'attribution de règles de comportements et d'objectifs à une multitude d'entités informatiques autonomes s'applique à de nombreuses situations. Avec les mêmes principes, on mime des phénomènes naturels, on anime des mondes artificiels à des fins ludiques, on teste l'aménagement du territoire, on étudie des comportements sociaux. Toutes ces applications sont fondées sur un schéma commun : une modélisation centrée sur les individus au moyen de la simulation par agents.

Rappelons que pour étudier un phénomène, les scientifiques aiment à en construire une représentation simplifiée, qu'ils appellent un « modèle ». Les orbites elliptiques des planètes de Kepler sont par exemple un modèle réaliste de leurs trajectoires autour du Soleil. Cependant, les modèles que l'on peut soumettre directement à l'expérience deviennent de plus en plus rares. On a alors recours à des simulations qui testent le modèle sur un substitut du réel. Aujourd'hui, la plupart des simulations consistent à résoudre de façon exacte ou approchée des équations qui sous-tendent les modèles physiques. Cette approche mathématique permet, par exemple, de modéliser l'évolution probable du climat à l'échelle planétaire ou d'une épidémie à l'échelle d'une population suffisamment homogène, de même que la régulation génétique au sein d'une cellule.

Mais une nouvelle façon d'envisager la simulation a émergé, au carrefour des recherches sur la vie artificielle (quand on laisse des ensembles d'entités virtuelles se développer et s'organiser) et des travaux sur la simulation des interactions d'individus. Le concept d'agents a repoussé les frontières de la modélisation au-delà des limites inhérentes à la seule mise en équations, notamment en fournissant des outils pour étudier les comportements en tant que tels. C'est ce qu'on appelle l'« intelligence collective ».

SE RECENTRER SUR L'INDIVIDU

En effet, l'une des faiblesses de l'approche mathématique de la simulation est son absence de portée explicative. On constate un comportement, telle une convergence vers un état, mais on ignore le détail des mécanismes en jeu. Les équations qui régissent le modèle portent sur le comportement moyen des individus et négligent les variations de comportement d'une entité à une autre. Le devenir d'un individu particulier et sa contribution à l'ensemble du phénomène restent inconnus.

À l'inverse, les modélisations dites « centrées sur des individus » sont focalisées sur les moyens plutôt que sur les fins. Elles posent le problème à l'échelle de l'individu : étant donné les entités qui interviennent dans un phénomène, quel doit être leur comportement pour que ce phénomène se manifeste ? Ce type de modélisation, qui permet de passer d'une simulation macroscopique à une

simulation microscopique, se décline principalement à travers deux paradigmes. Le premier est celui des automates cellulaires (le plus connu est le Jeu de la vie), où l'on divise l'espace en cellules dotées d'un état, et qui en changent en fonction de l'état des cellules voisines, selon des règles identiques pour toutes. Le second rassemble les simulations multi-agents qui offrent d'emblée aux entités qui les composent des comportements diversifiés : c'est ce que nous abordons ici.

L'industrie culturelle, en particulier les jeux vidéo et le cinéma, offre un débouché important pour ce type de simulations. Par exemple, dans les programmes comme le néo-zélandais Massive ou son homologue français Golaem, chaque « figurant » est un agent autonome à qui est attribuée une palette restreinte de règles de comportements, par exemple « Si je vois un ennemi, j'avance vers lui » ou « Si je suis attaqué sur le côté gauche, je pare avec mon bouclier ». Chaque agent prend ses décisions en fonction du contexte et selon des paramètres qui lui sont propres. De la même façon, le déplacement de vastes groupes d'animaux tels que les nuées d'oiseaux ou les bancs de poissons se simule typiquement dans le cadre d'un système multi-agents *(voir l'encadré de la page ci-contre)*.

Dans le champ scientifique, les agents connaissent aussi un grand succès. En physique ou en chimie, les corps semblables sont interchangeables : c'est bien le cas des atomes et des molécules de même nature. En revanche, en éthologie, en sociologie ou en économie, les acteurs sont dotés d'une histoire propre, et leurs comportements individuels varient. Les agents sont particulièrement bien adaptés à la modélisation de phénomènes collectifs émergeant de l'action d'individus « simples ».

En bref

> L'approche centrée sur des individus aide à construire des modèles explicatifs de phénomènes complexes.

> Un système multi-agents permet de construire des mondes artificiels pour tester des hypothèses.

> Tout peut être agent : un animal, un véhicule, un bâtiment, le feu…

> La conception des comportements des agents repose sur les résultats de la recherche en intelligence artificielle.

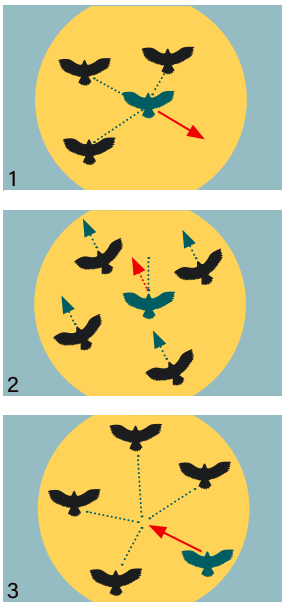
> Chaque agent agit selon le contexte et peut apprendre de ses expériences afin de mieux s'adapter.

> Les interactions des agents entraînent l'émergence de phénomènes collectifs non définis *a priori*.

Enfin, la simulation à visée ludique et celle qui fait office de « laboratoire virtuel » se rejoignent de plus en plus fréquemment, notamment à travers la montée en puissance des *serious games* à objectif pédagogique. Plus le jeu fait appel à des situations jugées vraisemblables par le joueur, plus ce dernier s'y implique et en assimile les connaissances. L'utilisation d'agents autonomes dans ce contexte permet une construction modulaire, facilement extensible et capable de répondre à des objectifs pédagogiques variés. C'est par exemple le cas dans Format Store, un logiciel destiné à former des vendeurs à la gestion de la relation clientèle, où l'apprenant est immergé, à travers son avatar, dans un supermarché virtuel, entièrement peuplé d'agents : les clients, mais aussi les articles, qui peuvent changer d'état, les panneaux d'informations, les caisses…

RÉACTIF OU COGNITIF ?

Le caractère ludique apparent de ces illustrations ne doit pas masquer les difficultés que pose la plupart du temps la définition de règles de comportements adéquates en fonction du but fixé aux agents. On doit d'abord déterminer et reproduire les caractéristiques physiques de l'environnement qui sont pertinentes par rapport au phénomène étudié : structure spatiale, lois physiques, événements qui y surviennent. Ensuite, chaque agent doit répondre aux informations perçues dans son environnement par des actions judicieuses. Le lien entre la perception et l'action peut être rudimentaire ou bien fondé sur des systèmes de planification ou de déduction. Enfin, l'environnement où les agents évoluent



Dans une nuée d'oiseaux, on a longtemps supposé qu'un meneur coordonnait l'ensemble du groupe. Craig Reynolds, de la société Sony, a montré que pour reproduire des groupes réalistes, trois règles simples appliquées à chaque individu suffisent : la **séparation** (1), qui tend à éloigner l'individu de ses congénères les plus proches ; l'**alignement** (2), qui force l'animal à s'orienter dans la direction moyenne de ses voisins ; enfin, la **cohésion** (3), qui le pousse vers leur position moyenne. Ce modèle, dit de « flocking », permet de simuler le comportement en vol, mais doit être enrichi pour les phases de décollage et de retour au sol.

est dynamique, car modifié sans cesse par leurs actions et leurs déplacements.

Placés dans un environnement changeant, complexe et imparfaitement connu, les agents doivent ainsi gérer des connaissances incomplètes et potentiellement erronées lorsqu'ils décident de leurs actions. Lorsque ces dernières conduisent à des impasses, elles peuvent être surmontées grâce au caractère adaptatif des comportements dont les agents sont dotés.

Pour ce faire, deux grandes approches du comportement individuel sont mises en œuvre. Les agents dits « réactifs » sont construits à l'aide d'instructions simples qui dictent la réponse à apporter en présence d'événements prédéfinis. Ils sont très dépendants de leur contexte et sont incapables d'envisager un objectif à long terme, par exemple « gagner la bataille ». Leurs capacités individuelles d'adaptation et d'autonomie sont donc limitées.

De plus, les règles comportementales peuvent être interdépendantes, si bien que l'ajout de nouvelles compétences demande une révision de l'ensemble des règles. Grâce à des mécanismes d'apprentissage, l'agent peut disposer de comportements qui vont s'adapter aux situations rencontrées de façon à y répondre plus efficacement, mais cela ne résout pas la gestion des situations imprévues. Néanmoins, le modèle réactif est aujourd'hui le plus utilisé pour la modélisation de comportements dans les jeux vidéo, en raison d'abord de sa simplicité de conception et de programmation et de son faible coût de calcul. En outre, il est parfaitement approprié à la modélisation d'entités physico-chimiques et fait fréquemment apparaître des comportements collectifs efficaces.

Cependant, simuler des comportements intelligents proches de ceux des êtres humains, prenant en compte la réflexion, l'opportunisme ou l'esprit d'équipe est de plus en plus souhaité. Ces compétences sont réservées à des agents plus évolués, dits « cognitifs ». Alors que le comportement

© Graham et Pour la Science (Shutterstock)

des agents réactifs dépend des événements qui surviennent dans leur contexte immédiat, la prise de décision des agents cognitifs s'appuie sur un raisonnement qui prend en compte les diverses possibilités d'action. Ils construisent ainsi des plans d'action afin de satisfaire des objectifs parfois lointains et abstraits, et sont capables d'adapter leurs plans à l'évolution du contexte.

Pour atteindre les buts qui leur sont fixés, les agents cognitifs élaborent et évaluent plusieurs plans d'action. Le principe le plus répandu est le chaînage arrière. L'agent cherche parmi les actions qu'il sait exécuter celle qui l'aiderait à atteindre le but fixé. Puis il étudie ses conditions d'application. Celles qui ne sont pas vérifiées deviennent de nouveaux buts, ou sous-buts, à satisfaire. La résolution suit ce principe de fonctionnement en cascade jusqu'à ce que l'agent arrive à une action dont toutes les conditions sont satisfaites. La plupart des agents cognitifs raisonnent ainsi en remontant de l'aboutissement des actions vers les conditions autorisant leur réalisation.

La tâche reste complexe, mais on peut recourir aux algorithmes conçus dans d'autres domaines de l'intelligence artificielle. Les résultats sur la planification aident chaque agent à établir l'ensemble des possibilités qui s'offrent à lui à chaque instant. Elles se traduisent par autant d'actions exécutables par l'agent dans le contexte actuel. Des mécanismes de sélection d'action déterminent ensuite l'action qui sera retenue. Les mêmes algorithmes génériques peuvent ainsi être utilisés par tous les agents au sein d'une simulation. Cette capacité des agents à intégrer les algorithmes les plus divers pour définir leurs comportements leur permet désormais de bénéficier des méthodes d'apprentissage automatique qui ont beaucoup progressé ces dernières années.

Les agents peuvent néanmoins être pourvus de capacités différentes. Ils sont ainsi construits à partir de briques élémentaires qui décrivent ce qu'ils peuvent faire au sein de la simulation et comment ils le font. Les variations dans les comportements obtenus résultent donc, d'une part, des contextes locaux distincts pour chaque agent et, d'autre part, des différences de compétences entre agents. Doté d'une compétence «soigner» avec zéro compétence «combat», un agent agira dans une bataille comme un infirmier, sans combattre. On peut aussi influer sur le comportement de l'agent en lui faisant préférer l'emploi d'une compétence à une autre. De la sorte, on obtient une diversité de comportements, voire de «personnalités», ce qui facilite la mise en place des simulations.

L'AGENT PEUT ÊTRE UN PIÉTON, MAIS AUSSI UNE VOITURE DE POMPIER

La modélisation par des agents permet certes de simuler l'action d'êtres vivants. Mais les moteurs de simulation actuels, pourvus de mécanismes d'optimisation, prennent en charge toutes sortes d'entités. Prenons un exemple, celui de l'élaboration de *smart cities*, un enjeu important consistant à doter «d'intelligence» des bâtiments, des routes, des dispositifs urbains, des véhicules, afin d'optimiser les futurs réseaux de ressources, d'énergie, de transport… La simulation peut, en amont de la réalisation technique *in situ,* apporter une aide dans la prise de décision en offrant des outils pour tester diverses hypothèses et comparer des scénarios. Pour ce faire, on doit modéliser

→ Dans la série ***Game of Thrones***, l'armée des morts semble sans fin… Seuls les figurants au premier plan sont réels, le reste est composé de milliers d'entités informatiques autonomes animées par le logiciel Golaem. Quelques règles simples de réaction à des événements qui surviennent dans leur entourage, combinées à l'utilisation de briques paramétrables (évitement de collision, suivi de chemin, etc.), suffisent pour régir leurs comportements.

© HBO/Rodeo FX

> Traiter un grand nombre d'entités soulève aussi une question méthodologique : à quel niveau de détail doit-on placer les agents ?

comme agents les divers artefacts et acteurs du système. Les réseaux de transport (électricité, gaz, eau, routes…) peuvent être coupés en tronçons qui sont autant d'agents.

Sur ces réseaux peuvent circuler les habitants (piétons, véhicules, transports collectifs), mais aussi les ressources, telle l'eau, dont la qualité peut être modifiée en retour par les autres activités urbaines. Les bâtiments deviennent aussi des agents qui prélèvent une part des ressources, des individus, des véhicules selon leur nature et leurs objectifs.

C'est ce genre de modélisation qui a été retenu par les concepteurs du jeu SimCity 2013 pour leur moteur GlassBox. Sa particularité est d'être complètement fondé sur les principes de la simulation multi-agents, la logique y étant poussée très loin puisque tout élément qui a une influence sur le jeu est un agent doté d'un comportement.

Cette nouvelle illustration du lien entre jeux vidéo et science montre en outre que l'approche « tout agent » peut supporter une montée d'échelle, à la fois en nombre total d'agents mais aussi en nombre de « familles » différentes. Elle offre également une grande souplesse et de nombreuses possibilités dans la réalisation de simulations. En unifiant les principes qui dirigent l'ensemble des entités impliquées dans la simulation, il devient plus facile d'y ajouter des nouveaux éléments et de les faire interagir afin d'enrichir la simulation et de multiplier les phénomènes qu'elle met en évidence. Néanmoins, l'apparition dans la dernière décennie de systèmes à très large échelle soulève de nouvelles problématiques sur lesquelles se focalise la recherche actuelle dans notre domaine.

En pratique, des simulations mettant en jeu des millions d'agents extrêmement diversifiés sont nécessaires pour l'étude de certains phénomènes, tels la propagation réaliste d'une maladie, un comportement économique global, la diffusion dans les réseaux sociaux… Le premier problème est alors celui du temps de calcul. La progression de la puissance des machines promet des simulations de plus en plus rapides, mais il s'agit surtout de programmer des modèles de comportements et des architectures d'agents avec des algorithmes concis et efficaces.

Traiter un grand nombre d'entités soulève aussi une question méthodologique : à quel niveau de détail doit-on placer les agents ? Prenons un exemple en biologie moléculaire, un des domaines les mieux adaptés à l'élaboration d'architectures et de modèles d'agents pour des applications à vaste échelle. Une bactérie renferme environ quatre mille types de molécules. Ne sachant pas toujours lesquelles sont pertinentes dans un mécanisme donné, on doit en simuler un très grand nombre. Si l'on se place à un niveau d'analyse plus fin, on atteint des populations de 10^{10} molécules dans la cellule ! Il est impossible de modéliser chacune d'elles. Doit-on alors créer un agent par espèce moléculaire ou pour un nombre fixé de molécules ? Comment faire le lien entre ces agents arbitraires et les concentrations ? Comment représenter les situations où quelques centaines d'agents molécules seulement sont impliqués ?

Un changement d'échelle de représentation devient intéressant. On regroupe des paquets d'agents en agents de plus haut niveau pour économiser du temps de calcul ; c'est aussi indispensable pour se placer à un niveau donné

Les aménageurs peuvent tester leurs idées sur des simulateurs de trafic routier comme PTV Vissim. Ce système multi-agents reproduit avec réalisme les réseaux routiers. On peut y introduire des flux de véhicules et des perturbations du trafic afin d'optimiser l'aménagement urbain.

et faciliter l'identification du niveau adéquat (molécules, groupes de molécules, cellules…).

La simulation par agents est ainsi un outil commode pour formuler des hypothèses en biologie moléculaire et les tester à faible coût. Dans cette situation, on est conduit à développer des simulations multi-agents « multiniveaux ». L'enjeu est d'automatiser le passage d'un niveau à l'autre, c'est-à-dire de détecter des groupes d'agents qui peuvent être remplacés par un agent macroscopique, et inversement.

LE MULTI-AGENTS FAIT LE BONHEUR

Enfin, la révolution du *big data* n'est pas sans conséquence sur la conception de simulations multi-agents : la disponibilité croissante de grandes masses de données concernant des systèmes complexes ne les rend pas plus faciles à analyser. Toutefois, elles sont utiles pour paramétrer avec plus de réalisme les propriétés et les comportements des agents. De plus, les techniques de recherche d'information qui sont employées pour traiter les données réelles peuvent être utilisées de la même façon sur les données produites par les simulations, ce qui offre de nouveaux outils de comparaison et de validation.

Les modèles multi-agents, où des entités autonomes communiquent, s'adaptent, voire se coordonnent et planifient leurs actions, autorisent la mise à l'épreuve des hypothèses de fonctionnement d'un système. On peut aussi comprendre l'apparition de caractères globaux et de structures macroscopiques à partir d'entités individuelles. Les modèles multi-agents sont ainsi un cadre naturel pour étudier des phénomènes complexes ou les comportements collectifs d'envergure.

Les outils nécessaires sont semblables quelle que soit la finalité de la simulation. L'expansion du domaine des jeux vidéo et la mise en place des métavers fournissent un banc d'essai remarquable, et laissent espérer le développement de techniques informatiques puissantes, applicables à des disciplines telles que la sociologie, la biologie ou l'industrie du commerce. De plus, la capacité de ces techniques à déborder du champ de la simulation pour fournir des outils de contrôle en temps réel d'objets physiques communicants (véhicules intelligents, réseaux de distribution, smartphones ou tablettes tactiles) permet d'envisager dans un avenir proche leur généralisation dans de nombreux secteurs d'activité. Jusqu'aux robots, bien entendu.

— Les auteurs —

> Philippe Mathieu dirige l'équipe Systèmes multi-agents et comportements (SMAC) du Centre de recherche en informatique, signal et automatique de Lille (Cristal).

> Jean-Christophe Routier est chercheur dans la même équipe.

> Sébastien Picault est chercheur Inrae dans l'unité Bioepar, à Nantes.

— À lire —

> D. Adam, Special report: The simulations driving the world's response to COVID-19, *Nature*, avril 2020. https://www.nature.com/articles/d41586-020-01003-6

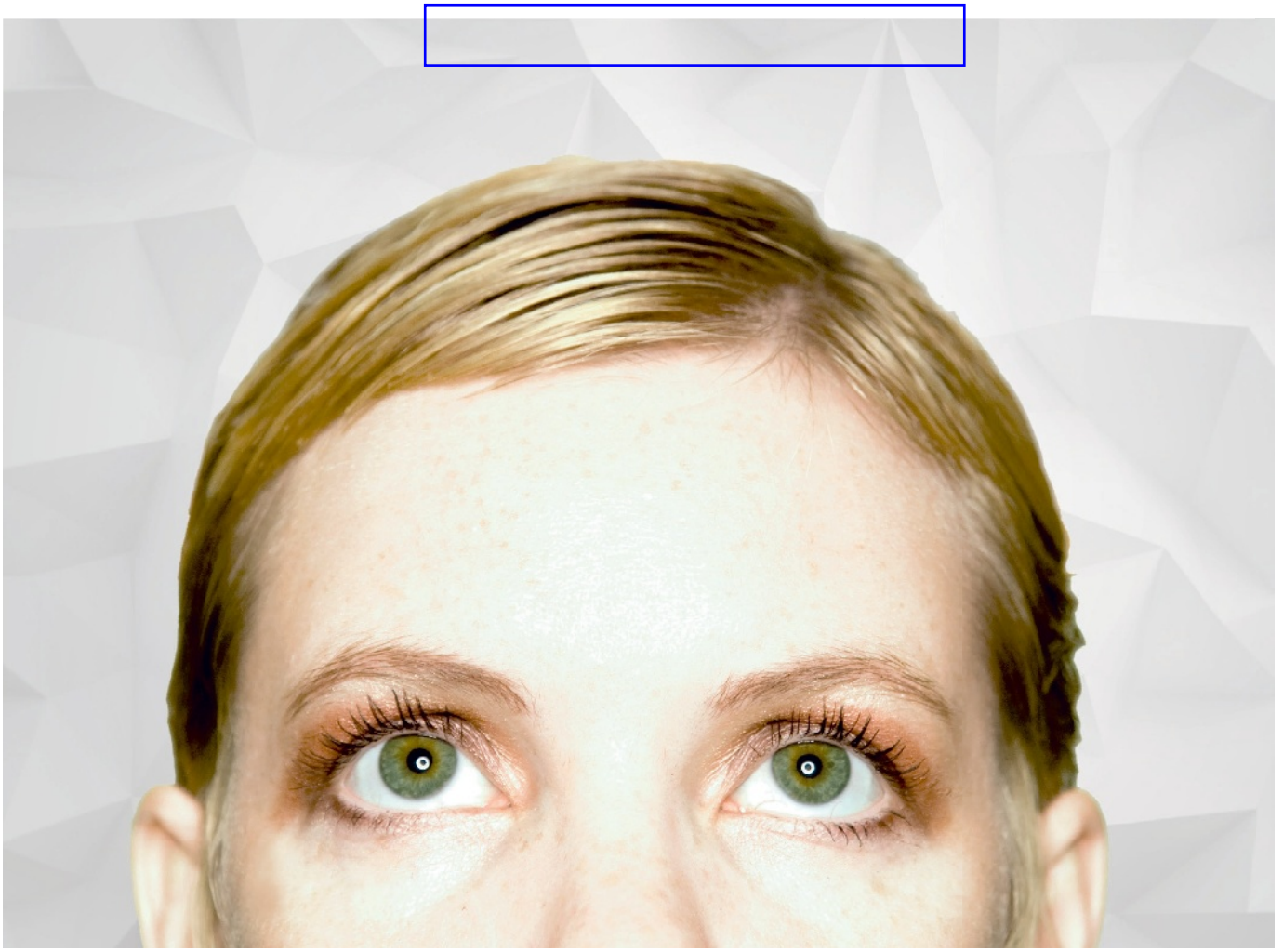
> J. D. Farmer et D. Foley, The economy needs agent-based modelling, *Nature*, vol. 460, pp. 685-686, 2009.

> P. Mathieu et S. Picault, An interaction-oriented model for multi-scale simulation, *in Ijcai 2011*, p 332-337, 2011.

© PTV Vissim

AcademiaNet

La base de données qui rassemble toutes les femmes scientifiques de renommée internationale



AcademiaNet offre un service unique aux instituts de recherche, aux journalistes et aux organisateurs de conférences qui recherchent des femmes d'exception dont l'expérience et les capacités de management complètent les compétences et la culture scientifique.

AcademiaNet, base de données regroupant toutes les femmes scientifiques d'exception, offre:

- Le profil de plus des 2.300 femmes scientifiques les plus qualifiées dans chaque discipline – et distinguées par des organisations de scientifiques ou des associations d'industriels renommées
- Des moteurs de recherche adaptés à des requêtes par discipline ou par domaine d'expertise
- Des reportages réguliers sur le thème »Women in Science«

Partenaires

Robert Bosch Stiftung

Spektrum der Wissenschaft

nature

Pour la Science

www.academia-net.org

Propos haineux, théories du complot et fausses nouvelles déferlent sur la Toile grâce, notamment, au renfort de robots logiciels. Contre ces fléaux issus de l'IA, il est heureusement un recours à fort potentiel : l'IA elle-même.

Communication

Les deux visages de l'intelligence artificielle

Laurence Devillers

© Alexander Pavlov/Shutterstock

En bref

> Sur internet, des outils d'intelligence artificielle sont utilisés de façon malveillante pour propager des informations fausses ou pour influencer les internautes.

> Ces outils exploitent nos biais cognitifs et nos réactions émotionnelles. Les *nudges*, qui consistent à inciter de façon imperceptible, renforcent leurs capacités de nuisance.

> Mais les *nudges* pourraient aussi aider les internautes à mieux résister à la désinformation.

> Des outils logiciels sont aussi développés pour traquer les contenus toxiques et repérer les robots malveillants œuvrant sur la Toile.

Avec la pandémie de Covid-19, nous avons vécu une période d'incertitude et d'anxiété qui a amplifié, en particulier sur internet, la propagation d'informations incomplètes ou fausses. Il y a eu d'abord la désinformation, c'est-à-dire la diffusion massive et délibérée d'informations visant à tromper ou à manipuler *via* les réseaux sociaux, les sites web, les forums, les messageries ou les *trolls* (internautes qui créent volontairement des polémiques). À laquelle s'est ajoutée la mésinformation, définie comme la propagation d'informations non vérifiées, incomplètes ou incertaines par des internautes qui n'ont bien souvent pas conscience des conséquences délétères de leurs actions.

Ainsi, compte tenu de l'arrivée massive d'informations provenant de sources multiples, contrôler la crédibilité et la fiabilité des données circulant en ligne devient une tâche urgente. D'autant que les techniques d'intelligence artificielle se développent et que les manipulateurs en tous genres n'hésitent pas à s'en servir.

Citons par exemple les techniques du *deepfake*, qui permettent de truquer des vidéos en leur superposant d'autres bandes-son ou images. Par ailleurs, sur les réseaux sociaux, l'utilisation des *bots*, ou robots logiciels, devient de plus en plus fréquente. Il existe ainsi des comptes automatisés, gérés par un algorithme plutôt que par une personne réelle, qui peuvent influencer le comportement des internautes, émettre des avis et renforcer des opinions. Ces faux profils alimentent des réseaux complexes en administrant des groupes, en augmentant le nombre de leurs membres et en « likant » les messages sur les pages. On estime que jusqu'à 15 % des comptes Twitter actifs sont de faux comptes pilotés par des *bots*.

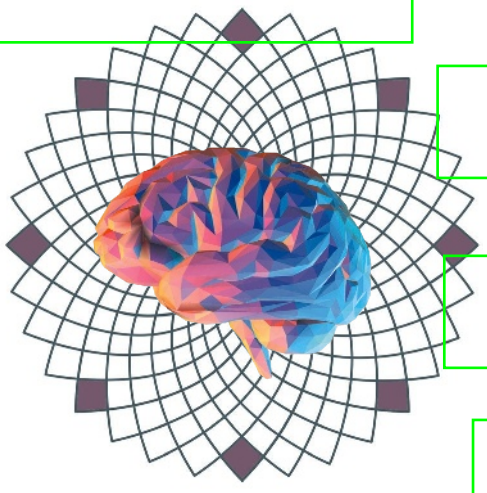
La menace va grandissant. Demain, ces *bots* seront capables d'interagir personnellement avec les internautes en détectant leurs émotions et leurs intentions. En exploitant nos biais cognitifs *(voir le schéma page ci-contre)*, ils pourront nous manipuler de façon très efficace et nous inciter par exemple à adhérer à un parti politique ou à acheter des produits censés nous apporter du bien-être. Les défis portent donc, d'une part, sur la vérification de la crédibilité des informations émises ou relayées par les internautes et par les *bots*, et, d'autre part, sur la vérification des « intentions » de ces machines.

Les développements dans le domaine de l'intelligence artificielle devraient conduire à des outils de plus en plus efficaces pour traquer sur internet les propos haineux, offensants ou menaçants *(voir l'encadré pages 48 et 49)*. Comme le préconisent la Commission européenne et le rapport 2019 de la mission française « Régulation des réseaux sociaux – Expérimentation Facebook », il est nécessaire de mener une réflexion d'ampleur en vue de constituer des bases de données communes pour améliorer les outils numériques de lutte contre la désinformation.

Des outils sont également disponibles pour détecter les faux profils et cartographier la diffusion de fausses nouvelles. Les ingénieurs de Facebook ont par exemple mis au point un outil d'apprentissage automatique qui a déjà permis de supprimer des milliards de faux comptes en 2020. Mais dès qu'un faux profil disparaît, un autre surgit. Pour mieux enrayer les dérives

Biais de confirmation
Tendance à accorder davantage de crédit et d'attention aux informations et événements qui confortent nos croyances, opinions ou préjugés.

Biais de négativité
Tendance à être davantage marqué par les expériences négatives que par les expériences positives.



Biais de l'angle mort
Incapacité de l'individu à reconnaître ses biais cognitifs, alors qu'il les remarque chez autrui.

Biais d'ancrage
Difficulté à se départir d'une première impression, qui pousse à se fier à l'information reçue en premier dans une prise de décision.

Illusion des séries
Tendance à trouver des régularités dans une série aléatoire d'événements.

Effet Google
Tendance à oublier les informations qui sont facilement accessibles par une recherche sur internet plutôt que de les mémoriser.

Effet de l'influence continue
Tendance à croire une information reçue antérieurement même si elle s'est révélée fausse et qu'elle a été corrigée.

Biais d'autocomplaisance
Tendance de l'individu à attribuer ses succès à ses qualités propres et ses échecs à des causes externes, indépendantes de lui.

informationnelles sur les réseaux sociaux, il est donc nécessaire de comprendre les ressorts de la désinformation sur ces plateformes, mais aussi les manipulations possibles à partir de la détection de nos comportements.

De nouveaux terrains de recherche en intelligence artificielle s'ouvrent donc : identifier les fausses informations, les rumeurs et les préjugés, résoudre les conflits entre des sources de données hétérogènes, mais aussi détecter les *trolls* ou les *bots* qui cherchent à nuire en propageant des rumeurs ou en manipulant les opinions.

LES BIAIS COGNITIFS, LES ÉMOTIONS ET LE « NUDGE »

Sur les réseaux sociaux, la diffusion « virale », massive et rapide, de certaines informations tient en grande partie à la capacité d'attention limitée des internautes, à leur appétence pour les réactions émotionnelles et aux biais cognitifs qui piègent nos raisonnements. Facebook, Twitter ou Instagram ont ainsi pour effet d'augmenter considérablement la visibilité des usagers malintentionnés ou de ceux qui sont prompts à exprimer sans frein leurs opinions. Certes, le principe de liberté d'expression est revendiqué par nos sociétés occidentales, mais les opinions exprimées sur les réseaux sociaux sont souvent des préjugés ou lieux communs dont la qualité ne devrait pas être estimée au nombre de clics qu'ils suscitent !

↑ Notre raisonnement est souvent affecté par des biais cognitifs que les manipulateurs exploitent ou qui, par eux-mêmes, jouent un rôle essentiel dans la propagation virale des informations sur les réseaux sociaux. Les psychologues ont recensé plusieurs dizaines de biais cognitifs ; quelques-uns, parmi les plus répandus, sont indiqués ici.

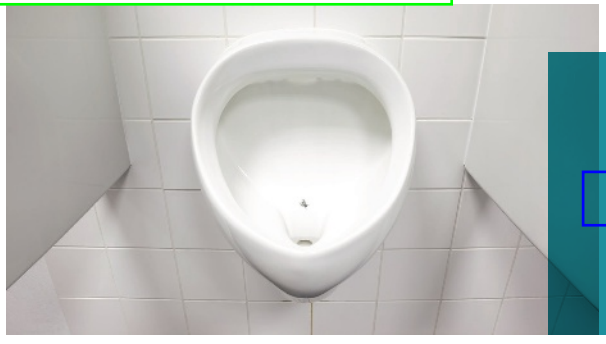
© Jolygon/Shutterstock

Les biais cognitifs jouent un rôle central dans la propagation de tels contenus. Et les futurs *bots* malveillants risquent de les exploiter de façon encore plus puissante et plus explicite. Dans les années 1970, les biais cognitifs ont été étudiés dans le cadre de l'économie comportementale, afin de comprendre nos décisions et agissements irrationnels. Les psychologues en ont recensé plusieurs dizaines, parfois classés en quatre catégories : les biais qui découlent d'un trop-plein d'informations, ceux dus à un manque de sens que l'individu cherche à combler, ceux liés au besoin d'agir vite et ceux induits par les limitations de la mémoire.

Le psychologue et économiste américano-israélien Daniel Kahneman, l'un des pères fondateurs de ce courant, Prix d'économie en mémoire d'Alfred Nobel en 2002, a montré dans ses travaux avec Amos Tversky le rôle important des émotions et des biais cognitifs dans nos processus de décision.

Une étape supplémentaire est venue avec un concept mis en exergue en 2008 par les Américains Richard Thaler et Cass Sunstein, et dont la mise en œuvre exploite les biais cognitifs et les réactions émotionnelles des humains. Ces deux chercheurs ont montré qu'à l'aide d'une conception appropriée des choix proposés aux individus, on peut améliorer de façon spectaculaire les décisions que ces individus prennent, sans forcer explicitement quiconque à faire quoi que ce soit. Richard Thaler et Cass Sunstein ont nommé *nudge* cette tactique de modification subtile du comportement.

Les *nudges* sont des sortes de coups de pouce imperceptibles qui prennent des formes très variées et qui sont utilisés dans divers contextes, notamment dans le marketing et la politique. L'exemple de *nudge* sans doute le plus connu est celui de la mouche peinte au fond de certains urinoirs : elle constitue une cible naturellement visée par les usagers, et aide ainsi à maintenir propres les toilettes.



← Le *nudge* le plus célèbre est celui des urinoirs dont le fond est orné d'une image de mouche qui incite les usagers à viser cette cible et, ainsi, à maintenir la propreté des toilettes. Ce concept d'incitation/manipulation douce se répand aussi dans le monde numérique.

© Shutterstock/Marcel Derweduwen

> Le « nudge » ? Un coup de pouce qui permet d'améliorer de façon spectaculaire les décisions d'un individu, sans le forcer explicitement à faire quoi que ce soit

En d'autres termes, les *nudges* sont des incitations douces, sans coercition, qui guident les individus vers un choix par défaut mais non obligatoire, la liberté de décision devant toujours être ménagée. Ils influencent une personne ou un groupe, dans l'intention de tenter de contrôler ou de modifier les actions d'une autre personne ou d'un autre groupe. Le *nudging* opère principalement à travers les éléments affectifs d'un système rationnel humain, l'attention, la mémoire, le jugement moral, la prise de décision étant influencés par les émotions.

Les *nudges* peuvent encourager des comportements ayant des effets à long terme pour l'individu ou la société. Par exemple, avant 1976, les Français devaient déclarer explicitement leur consentement à faire don, après leur mort, de leurs organes : cette règle a été modifiée et le consentement est désormais le choix par défaut. Or une étude de 2003 portant sur plusieurs pays et réalisée par Eric Johnson et Daniel Goldstein, de l'université Columbia, a montré que lorsque le consentement est le choix par défaut, le pourcentage effectif de donneurs d'organes est notablement supérieur : il est presque le double.

QUAND LE « NUDGE » DEVIENT NUMÉRIQUE

Les *nudges* peuvent être utilisés, dans le monde numérique, à bon comme à mauvais escient. Ils pourraient notamment jouer un rôle positif dans la lutte contre la désinformation. Par exemple, la sensibilisation des internautes à l'existence des *nudges* numériques permettrait de leur faire prendre conscience des biais cognitifs auxquels ils sont soumis lorsqu'ils réagissent de façon quotidienne sur les réseaux sociaux en propageant des informations non vérifiées. On pourrait ainsi se servir de *nudges* pour contrebalancer certaines opinions radicales, dans un objectif pédagogique : faire comprendre les manipulations et renforcer l'esprit critique, lequel se distingue de l'opinion par le questionnement, l'argumentation, l'approche contradictoire et le souci d'approcher une certaine vérité.

Le *nudge* devient de plus en plus pertinent pour les développeurs de technologies ciblant les soins de santé. Ainsi, l'université de Pennsylvanie a créé en 2016 au sein de son système hospitalier une unité dévolue aux *nudges*, la Penn Medicine Nudge Unit. Parmi ses nombreux projets réalisés, ce groupe a par exemple conçu des options par défaut pour les intégrer au système informatique utilisé par les médecins, options qui ont eu pour effet d'augmenter la prescription de médicaments génériques et réduire la prescription d'opioïdes.

Le concept de *nudge* s'intègre naturellement à l'informatique affective. Ce domaine regroupe des technologies d'intelligence artificielle comme la détection par les robots, qu'ils soient matériels ou purement logiciels, des émotions de l'interlocuteur humain, mais aussi la génération d'émotions chez celui-ci, la modulation des décisions du robot à partir des informations d'ordre affectif détectées dans le comportement humain, etc. On peut les déployer dans des *bots* pour orienter subrepticement les choix des individus en exploitant nos biais cognitifs tels que le biais de confirmation (tendance à privilégier les informations qui confirment nos idées préconçues), le biais de conformisme (tendance à penser et agir comme

IDENTIFIER LES CONTENUS TOXIQUES

Sur les plateformes sociales, la prévalence du langage haineux ou offensant en ligne a augmenté rapidement ces dernières années, et désormais le problème est grave. Dans certains cas, des commentaires toxiques en ligne ont entraîné des violences réelles, du nationalisme religieux au Myanmar à la propagande néonazie aux États-Unis.

Pourtant, des milliers de vérificateurs humains luttent pour modérer le volume toujours croissant de contenus préjudiciables. Mais, en 2019, il a été rapporté que les modérateurs de Facebook risquaient de souffrir de troubles de stress post-traumatique du fait d'une exposition répétée à des contenus aussi dérangeants. Confier au moins en partie ces activités de modération à des machines aiderait à gérer les volumes croissants de contenus préjudiciables, tout en limitant l'exposition humaine à ces derniers. Et en effet, de nombreux géants de la technologie intègrent des algorithmes dans leur modération de contenus depuis des années.

C'est le cas de Google Jigsaw. En 2017, cette entreprise a contribué à la création de *Conversation AI*, un projet de recherche collaboratif visant à détecter les commentaires toxiques en ligne. Cependant, un outil produit par ce projet, appelé Perspective, a fait l'objet de nombreuses critiques. Un reproche fréquent était qu'il créait un « score de toxicité » général qui n'était pas assez flexible pour répondre aux besoins variés des différentes plateformes. Par exemple, certains sites web souhaitent détecter les menaces mais pas les blasphèmes, tandis que d'autres ont des exigences opposées.

Un autre problème est que l'algorithme, après apprentissage, ne distinguait pas les commentaires toxiques des commentaires non toxiques contenant des mots liés au sexe, à l'orientation sexuelle, à la religion ou au handicap. Par exemple, un utilisateur a signalé que des phrases neutres et simples telles que « Je suis une femme noire lesbienne » ou « Je suis une femme sourde » donnaient des scores de toxicité élevés, tandis que « Je suis un homme » donnait un score faible.

À la suite de ces préoccupations, l'équipe de *Conversation AI* a invité les développeurs à entraîner leurs propres algorithmes de détection de toxicité et à les inscrire à trois concours (un par an) organisés sur Kaggle, une filiale de Google connue pour sa communauté de praticiens de l'apprentissage automatique, ses ensembles de données publiques et ses défis. Pour aider à entraîner les algorithmes, *Conversation AI* a publié deux ensembles de données publiques contenant plus d'un million de commentaires toxiques et non toxiques provenant de Wikipedia et d'un service appelé Civil Comments. Les commentaires ont été classés selon leur toxicité par des annotateurs, avec une étiquette « Très toxique » indiquant « un commentaire très haineux, agressif ou irrespectueux qui vous fera très vraisemblablement quitter une discussion ou renoncer à partager votre point de vue », et une étiquette « Toxique » signifiant « un commentaire grossier, irrespectueux ou déraisonnable qui est assez susceptible de vous faire quitter une discussion ou de renoncer à partager votre point de vue ». Certains commentaires ont été vus par beaucoup plus de dix annotateurs (jusqu'à des milliers), en raison de l'échantillonnage et des stratégies utilisées pour renforcer la précision des évaluateurs.

Detoxify pour purger le web

L'objectif du premier défi de Jigsaw était de construire un modèle de classification des commentaires toxiques avec des étiquettes telles que « toxique », « gravement toxique », « menace », « insulte », « obscène » et « haine identitaire ». Les deuxième et troisième défis se sont concentrés sur des aspects plus spécifiques : minimiser les biais involontaires concernant les mentions identitaires et ceux dus au fait que les modèles multilingues sont entraînés avec des données en anglais uniquement. Bien que les défis aient conduit à des moyens astucieux d'améliorer les modèles de langage toxique, notre équipe chez Unitary, une société d'intelligence artificielle modératrice de contenus, a constaté qu'aucun des modèles entraînés n'avait été rendu public.

C'est pourquoi nous avons décidé de nous inspirer des meilleures solutions de Kaggle et d'entraîner nos propres algorithmes dans l'intention de les rendre publics. Pour ce faire, nous nous sommes appuyés sur des modèles existants de traitement du langage naturel, comme Bert, de Google. Beaucoup de ces modèles ont leur code source en libre accès. C'est ainsi que notre équipe a créé Detoxify, une bibliothèque d'algorithmes de détection des

© Peter Dazeley/Getty Images

commentaires, ouverte et conviviale, pour repérer les textes inappropriés ou préjudiciables. Son utilisation est destinée à aider les praticiens à identifier les commentaires potentiellement toxiques.
Dans le cadre de cette bibliothèque, nous avons publié trois modèles différents correspondant à chacun des trois défis de Jigsaw. Alors que les meilleures solutions de Kaggle pour chaque défi utilisent des ensembles de modèles, qui font la moyenne des scores de plusieurs modèles entraînés nous avons obtenu une performance similaire avec un seul modèle par défi. Chaque modèle est facilement accessible en une seule ligne de programme et tous les modèles ainsi que le programme d'apprentissage sont disponibles publiquement sur GitHub.

« Stupide », un mot à toxicité variable

Bien que ces modèles soient performants dans de nombreux cas, il est important de noter également leurs limites. Tout d'abord, ils fonctionneront bien sur des exemples similaires aux données sur lesquelles ils ont été formés, mais ils risquent d'échouer s'ils sont confrontés à des exemples moins familiers. Nous encourageons les développeurs à affiner ces modèles sur des ensembles de données toxiques représentatifs de leur cas d'utilisation.
De plus, nous avons remarqué que l'inclusion d'insultes ou de blasphèmes dans un commentaire textuel entraînera presque toujours un score de toxicité élevé, indépendamment de l'intention ou du ton de l'auteur. Par exemple, la phrase « Je suis fatigué d'écrire cet essai stupide » donnera un score de toxicité de 99,7 %, tandis que la suppression du mot « stupide » fera passer le score à 0,05 %.
Enfin, malgré le fait qu'un des modèles publiés ait été spécifiquement formé pour limiter les biais involontaires, les trois modèles sont toujours susceptibles de présenter certains biais, ce qui peut poser des problèmes éthiques lorsqu'on les utilise pour modérer des contenus.

Bien que des progrès considérables aient été accomplis en matière de détection automatique des propos toxiques, beaucoup reste à faire avant que les modèles puissent saisir le sens réel, nuancé, de notre langue – au-delà de la simple mémorisation de mots particuliers ou de phrases particulières. Bien sûr, investir dans des ensembles de données plus représentatifs et de meilleure qualité, pour l'apprentissage des systèmes d'intelligence artificielle, permettrait des améliorations progressives. Mais nous devons aller plus loin et commencer à interpréter les données dans leur contexte, un élément crucial pour comprendre le comportement en ligne. Un texte apparemment anodin publié sur les réseaux sociaux, mais accompagné d'un symbolisme raciste dans une image ou une vidéo, passerait facilement inaperçu si nous ne regardions que les mots écrits. Il est bien connu qu'en ignorant ou en négligeant le contexte, nous commettons facilement des erreurs de jugement. Pour que l'intelligence artificielle remplace à grande échelle les humains dans la modération des contenus échangés sur les réseaux sociaux, il est impératif que nous donnions à nos modèles une image complète de ces contenus.

Laura Hanu
est ingénieure en vision par ordinateur dans la société britannique Unitary.

James Thewlis
est docteur en vision par ordinateur, cofondateur et directeur technique de Unitary.

Sasha Haco
est docteure en physique théorique, cofondatrice et PDG de Unitary.

la majorité), le biais d'ancrage (qui pousse à se fier à l'information reçue en premier dans une prise de décision)…

L'emploi de *nudges* à travers des *bots* pour influencer les internautes s'est développé ces dernières années. Il peut s'agir de *nudges* pour inciter ou pour freiner des comportements, que Richard Thaler appelle *sludges* (littéralement « boues », en anglais). L'un des défis consiste à évaluer l'apport ou la menace qu'ils représentent. Ainsi, le projet *Bad Nudge-Bad Robot* que je dirige au CNRS à l'université Paris-Saclay vise à mettre en évidence le danger des techniques de *nudging* pour les personnes vulnérables telles que les enfants ou les personnes âgées. Ce projet entre dans le cadre de ma chaire d'intelligence artificielle Humaaine (Human-Machine affective interaction & ethics) au CNRS, qui repose sur une forte collaboration interdisciplinaire entre l'informatique affective et la robotique émotionnelle, l'économie comportementale, la linguistique et le traitement du langage naturel.

DANS LA JUNGLE DU WEB

Par exemple, en 2019, dans une expérience de psychologie sociale conduite dans une école primaire, nous avons montré que les *nudges* des machines (robots, *chatbots*) étaient plus efficaces que les *nudges* d'humains. Dans cette expérience, la machine ou l'humain donne le même nombre de billes à chaque enfant. La première question posée à l'enfant est : combien il en garde pour lui et combien il en donne aux autres enfants. La deuxième question est un *nudge* pour le faire changer d'avis, en ajoutant par exemple que les autres en donnent plus (ce qui active le biais de conformisme). Plus les enfants étaient jeunes, plus ils étaient facilement manipulés par les machines. Cet exemple met en lumière le fort potentiel d'incitation associé à des agents capables de déployer un *nudge*. À n'en pas douter, de tels *bots* accroîtront les moyens de manipulation sur le web.

L'intelligence artificielle fournit ainsi des armes diverses, les unes permettant de traquer les propos haineux et repérer les *trolls* et les *bots*, les autres de propager des fausses informations et de manipuler les internautes, notamment grâce aux *nudges*. Ceux-ci sont aussi susceptibles d'être utilisés de façon positive. Imaginons par exemple qu'à la fin du mois on vous attribue des points en fonction de la qualité de ce que vous avez tweeté, et que vous puissiez comparer avec vos amis ou vos collègues. Ce *nudge* vous inciterait à moins relayer des informations non vérifiées et peut-être à adopter des pratiques plus respectueuses des autres.

Lutter contre la désinformation s'annonce donc comme un défi permanent. Le recours à l'intelligence artificielle dans ce combat n'en est certainement qu'à ses débuts. Si se doter de tels outils sera utile pour faire face aux propos toxiques ou aux *bots* malveillants, cela restera néanmoins insuffisant.

Pour protéger citoyens et démocraties, une régulation de cette jungle qu'est le web, par des lois et des garde-fous éthiques, est nécessaire. C'est assurément une question aussi difficile que cruciale pour l'avenir de nos sociétés. Elle suscite aujourd'hui de nombreuses réflexions. Notamment, le Comité national pilote d'éthique du numérique a publié en juillet 2020 un document portant sur les enjeux dans la lutte contre la désinformation et la mésinformation, lequel identifie les tensions éthiques relatives à la mise en œuvre, par les plateformes, d'outils de modération de contenus et aux mécanismes de lutte contre la désinformation. Ce texte interroge aussi l'autorité acquise par ces plateformes et les rapports qu'elles entretiennent avec l'État et la justice. En résumé, il convient de redéfinir la responsabilité des plateformes aux niveaux national et européen, tout en protégeant la liberté d'expression. De ces sujets de société, nous allons devoir débattre.

— L'autrice —

> **Laurence Devillers** est professeuse en intelligence artificielle à Sorbonne-Université et dirige une équipe de recherche au CNRS-LISN (Saclay), où elle est responsable de la chaire IA Humaaine (https://humaaine-chaireia.fr).

— À lire —

> **L. Devillers,** *Vague IA à l'Élysée*, L'Observatoire, 2022.

> **H. Ali Mehenni, L. Devillers *et al.*,** Nudges with conversational agents and social robots: a first experiment with children at a primary school, Conversational Dialogue Systems for the Next Decade, *Springer*, pp. 257-270, 2021.

> **C. Schneider *et al.*,** Digital nudging: guiding online user choices through interface design, *Communications of the ACM*, 2018.

> **D. Kahneman,** *Thinking, Fast and Slow*, Farrar, Straus and Giroux, 2011.

AVEZ-VOUS DES RESSOURCES QUE L'ON ÉCONOMISE QUAND ON RECYCLE ?

EN 2018, 1,3 MILLION DE TONNES
DE PAPIERS ONT ÉTÉ RECYCLÉES.
CE SONT 25 MILLIARDS DE LITRES D'EAU
ÉCONOMISÉS, L'ÉQUIVALENT
DE 8000 PISCINES OLYMPIQUES. ET ÇA,
C'EST GRÂCE À VOTRE GESTE DE TRI.

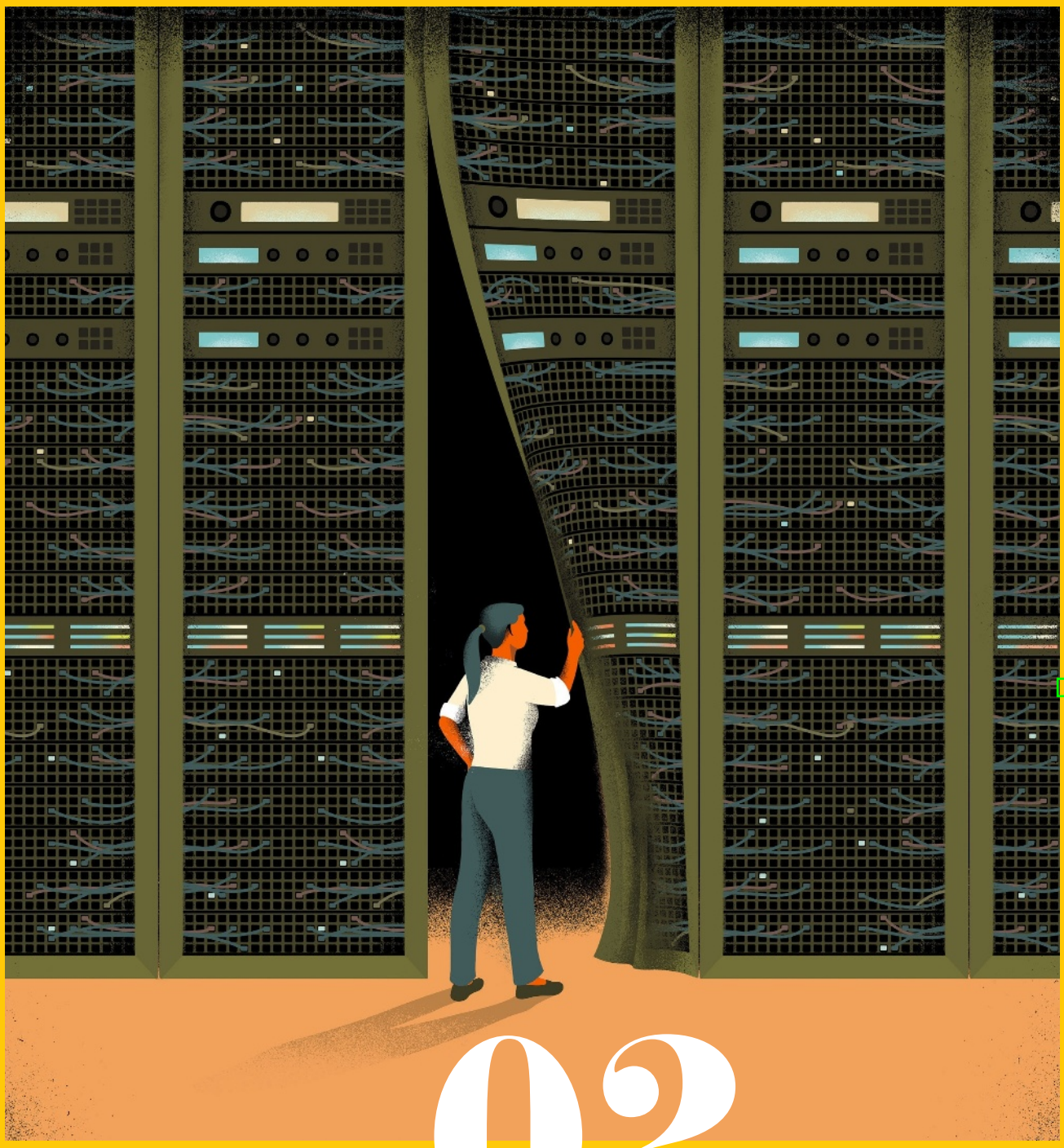
PLUS D'INFORMATIONS SUR LE RECYCLAGE
SUR TRIERCESTDONNER.FR

CITEO

Donnons ensemble une nouvelle vie à nos produits

Chantier en cours

C'est entendu, sous réserve d'être bien entraînés et gavés de données, les réseaux de neurones font des étincelles. Comment s'y prennent-ils ? C'est un problème : si l'intelligence artificielle symbolique était lisible, l'intelligence artificielle connexionniste, celle de l'apprentissage profond, est opaque. Parfois, la variation d'un seul pixel la fait dérailler. Beaucoup d'essais et d'erreurs sont nécessaires, car la théorie des réseaux neuronaux est très incomplète. Quant à savoir ce qu'ils comprennent… Pour la recherche, l'un des enjeux est donc de faire sortir cette branche de l'intelligence artificielle de son état de « boîte noire ».



© Sébastien Thibault/Pour la science

02

Interro écrite pour l'intelligence artificielle

© Maggie Chiang

L'IA comprend-elle ce qu'elle fait ? Si oui, encore faudrait-il le prouver par des tests appropriés. Pas si simple, d'autant que les réseaux de neurones actuels emploient des raccourcis statistiques qui brouillent les pistes.

Melanie Mitchell

Vous souvenez-vous du jeu Jeopardy diffusé en France de 1989 à 1992 ? Une sorte d'anti-quiz : l'animateur fournissait une réponse, et les candidats imaginaient la question. Aux États-Unis, patrie originelle du programme (la trente-huitième saison est en cours aujourd'hui !), une intelligence artificielle (IA) conçue par IBM et baptisée Watson est devenue un crack à ce jeu en battant, en février 2011, deux anciens champions humains. Une publicité de l'époque prétendait même que « Watson comprend le langage naturel dans toute sa complexité et son ambiguïté ». Cependant, les promesses n'ont pas été tenues, et Watson a échoué de manière spectaculaire dans sa quête pour « révolutionner la médecine avec l'intelligence artificielle ». Preuve qu'une facilité linguistique de surface est bien différente d'une réelle compréhension du langage humain.

Depuis longtemps, le traitement du langage naturel est un des objectifs majeurs de la recherche en IA. Les chercheurs ont d'abord essayé de programmer à la main tout ce dont une machine peut avoir besoin pour comprendre un magazine, un roman ou toute autre production écrite. Cette approche, Watson l'a montré, était vaine – il est impossible de coucher noir sur blanc tous les faits, règles et suppositions qui n'ont pas encore été écrits et sont nécessaires pour comprendre un texte.

Plus récemment, un nouveau paradigme a été établi : au lieu de les abreuver de connaissances explicites et les dresser à prédire des mots, nous laissons les machines comprendre le langage par elles-mêmes en leur faisant ingurgiter d'énormes quantités de textes. Le résultat est ce que les chercheurs appellent un « modèle de langage ». Le GPT-3 de la société OpenAI, inauguré en mai 2020, par exemple, peut produire de la prose (et de la poésie !) humaine à s'y méprendre, et conduire des raisonnements linguistiques apparemment raffinés.

Mais GPT-3, entraîné avec des textes issus de milliers de sites web, de livres et d'encyclopédies, va-t-il au-delà du simple vernis de Watson ? Comprend-il les mots qu'il produit et sur lesquels il raisonne ostensiblement ? C'est là un sujet de vif désaccord au sein de la communauté des chercheurs en IA. De telles discussions étaient l'apanage des philosophes, mais tout au long de la décennie passée, l'IA a jailli hors de sa bulle académique vers le monde réel. La question n'est pas que rhétorique, car le manque de compréhension de ce monde peut avoir des conséquences concrètes et parfois dévastatrices. Dans une étude portant sur les recommandations de Watson pour lutter contre le cancer, des « exemples multiples de traitements incorrects et peu sûrs » ont été pointés. D'autres travaux ont montré que le système de traduction de Google commettait d'importantes erreurs lorsqu'il s'appliquait à des instructions médicales destinées à des patients non anglophones.

Comment juger, en pratique, si une machine a la capacité de comprendre ? En 1950, le pionnier de l'informatique Alan Turing a essayé de répondre à cette question avec son fameux « jeu de l'imitation », aujourd'hui connu sous le nom

En bref

> Donner aux machines la maîtrise du langage naturel est, de longue date, un objectif de l'intelligence artificielle.

> Entraînés de façon adéquate, les réseaux de neurones comme GPT-3 semblent capables de performances spectaculaires. Mais comprennent-ils vraiment ?

> Du test de Turing aux schémas Winograd, il n'est pas aisé d'établir un test prouvant une vraie compréhension, notamment parce que les programmes recourent à des raccourcis statistiques.

> Les machines ne sauront sans doute comprendre le langage que lorsqu'elles sauront aussi comprendre le monde qui les entoure.

de « test de Turing ». Une machine et un humain, tous deux cachés, sont en compétition pour convaincre un juge humain de leur humanité en n'usant que de la conversation écrite. Si le juge est incapable de trancher, pensait Turing, nous devrions alors considérer que la machine pense – et, en effet, comprend.

Malheureusement, Turing a sous-estimé la propension des humains à se laisser duper. Même de simples robots de conversation, tels qu'Eliza – l'ersatz de psychothérapeute conçu par Joseph Weizenbaum en 1960 –, ont réussi à faire croire à des testeurs qu'ils discutaient avec un être doué de compréhension, alors même qu'ils savaient parler à une machine.

MAÎTRE DU MONDE POUR UN PRONOM

Dans un article de 2012, les informaticiens Hector Levesque, Ernest Davis et Leora Morgenstern ont proposé un test plus objectif, qu'ils ont baptisé « challenge du schéma Winograd ». Ce test a, depuis, été adopté par la communauté du langage de l'IA comme moyen, peut-être le meilleur, d'évaluer le degré de compréhension d'une machine – bien que, nous le verrons, il ne soit pas parfait. Un schéma Winograd, nommé d'après le chercheur en linguistique Terry Winograd, de l'université Stanford, consiste en une paire de phrases différant d'un mot exactement, chacune suivie d'une question. Voici deux exemples :

Paire 1, phrase 1 : J'ai versé de l'eau de la bouteille dans la tasse jusqu'à ce qu'elle soit *pleine*.

Question : Qu'est-ce qui était rempli, la bouteille ou la tasse ?

Paire 1, phrase 2 : J'ai versé de l'eau de la bouteille dans la tasse jusqu'à ce qu'elle soit *vide*.

Question : Qu'est-ce qui était vide, la bouteille ou la tasse ?

Paire 2, phrase 1 : L'oncle de Joe peut encore le battre au tennis, bien qu'il soit trente ans plus *vieux*.

Question : Qui est plus vieux, Joe ou l'oncle de Joe ?

Paire 2, phrase 2 : L'oncle de Joe peut encore le battre au tennis, bien qu'il soit trente ans plus *jeune*.

Question : Qui est plus jeune, Joe ou l'oncle de Joe ?

Dans chaque paire, la différence d'un mot *(en italique)* peut changer la personne ou la chose à laquelle le pronom fait référence. Répondre correctement à ces questions semble nécessiter une compréhension de type sens commun. Les schémas Winograd sont conçus précisément pour évaluer cela, en réduisant la vulnérabilité du test de Turing aux juges humains non fiables et aux astuces des robots conversationnels. En particulier, les auteurs ont conçu quelques centaines de schémas qu'ils pensent être imperméables à Google : une machine ne devrait pas être capable de lancer une recherche Google (ou tout autre moteur de recherche) pour répondre à ces questions correctement.

Ces schémas ont été l'objet d'une compétition en 2016 durant laquelle le programme vainqueur a répondu juste à seulement 58 % des phrases – un résultat à peine meilleur que celui qu'elle aurait obtenu en devinant les réponses au hasard. Ce qui fait dire malicieusement à Oren Etzioni, éminent chercheur en IA : « Quand

Quanta magazine

*Cet article a d'abord été publié en anglais par **Quanta Magazine,** une publication en ligne indépendante, soutenue par la Simons Foundation afin de favoriser la diffusion des sciences : bit.ly/3MFPKtr*

une intelligence artificielle ne peut déterminer à qui un pronom fait référence dans une phrase, il est difficile d'imaginer qu'elle puisse prendre un jour le contrôle du monde.»

Cependant, la capacité des programmes d'IA à résoudre les schémas Winograd a grimpé en flèche avec l'avènement des modèles de langage en réseau neuronal. En 2020, une étude d'OpenAI relatait que GPT-3, qui relève de cette catégorie d'IA, obtenait de bonnes réponses sur près de 90 % des phrases dans un lot de référence de phrases de ce type. D'autres modèles de langage sont encore plus performants après s'être entraînés spécifiquement à ces tâches. Fin 2021, les modèles de langage en réseau neuronal atteignaient 97 % de précision sur un lot particulier de schémas Winograd qui font partie d'une compétition de compréhension des langues par les IA, nommée SuperGlue. C'est à peu près équivalent aux performances humaines. Cela signifie-t-il que les modèles de langage en réseau neuronal ont atteint un niveau de compréhension comparable au nôtre ?

ÊTRE COLLÉ À LA COMPÉTITION SUPERGLUE

Pas forcément. Malgré tous les efforts de leurs créateurs, ces schémas Winograd n'étaient en fait pas imperméables à Google. Ces défis, comme beaucoup d'autres tests actuels de compréhension du langage par les IA, recourent parfois à des raccourcis statistiques grâce auxquels les réseaux neuronaux sont performants sans comprendre. Prenez, par exemple, les phrases «la voiture de sport a doublé la camionnette de la poste, car elle était plus rapide» et «la voiture de sport a doublé la camionnette de la poste, car elle était plus lente». Un modèle de langage entraîné avec un large corpus de phrases en français aura assimilé la corrélation entre «voiture de sport» et «rapide», et entre «camionnette de la poste» et «lent»; il peut donc répondre correctement uniquement grâce à ces corrélations. De fait, beaucoup de schémas Winograd de la compétition SuperGlue autorisent ce genre de lien statistique.

> Turing a sous-estimé la propension des humains à se laisser duper. Même de simples robots de conversation peuvent faire illusion

Plutôt qu'abandonner ces schémas en tant que test de compréhension, un groupe de chercheurs de l'institut Allen (du nom d'un cofondateur de Microsoft) pour l'IA, à Seattle, a essayé de régler certains de leurs problèmes.

En 2011, aux États-Unis, l'ordinateur Watson, d'IBM, triomphe au jeu Jeopardy. Nourri avec l'équivalent de 1 million de livres, il capte les questions posées en langage naturel par le présentateur, le sens des mots, mais aussi s'il faut ou non répondre à une question. Une performance pas si élémentaire…

En 2019, ils ont créé WinoGrande, un lot bien plus étoffé de schémas Winograd : il contient 44 000 phrases, contre quelques centaines jusque-là. Pour obtenir autant d'exemples, les chercheurs ont fait appel à Amazon Mechanical Turk, une plateforme pour proposer à des humains, contre rémunération, des tâches plus ou moins complexes. Il a été demandé à chaque travailleur d'écrire quelques paires de phrases, avec des contraintes pour s'assurer que l'ensemble aborderait divers sujets, mais avec la possibilité que ces paires puissent à présent différer de plus d'un mot.

Les chercheurs ont ensuite tenté d'éliminer les phrases propices à des raccourcis statistiques, en soumettant chacune d'entre elles à une IA assez peu sophistiquée et en écartant toutes celles qui étaient trop simples à résoudre. Comme ils s'y attendaient, les phrases qui restaient offraient un défi bien plus difficile pour les machines que la collection originale. Tandis que les humains continuaient d'obtenir des scores très élevés, les modèles de langage en réseau neuronal qui égalaient la performance humaine sur le lot d'origine ont vu leur score s'effondrer avec le lot WinoGrande. Ce nouveau défi restaurait le statut de « test de compréhension de sens commun » pour les schémas Winograd – du moment que les phrases étaient scrupuleusement filtrées pour assurer leur imperméabilité à des recherches Google.

> Il est souvent difficile de savoir si des systèmes d'IA comprennent véritablement le langage, ou d'autres données, qu'ils traitent

Cependant, une autre surprise allait surgir. Dans les deux années qui suivirent la publication de la collection WinoGrande, les modèles de langage en réseau neuronal ont continué de croître, et plus ils sont larges, meilleur semble être leur score à ce nouveau défi. Fin 2021, les meilleurs programmes actuels – d'abord entraînés avec des térabits de textes, puis avec des milliers d'exemples WinoGrande – obtenaient

© IBM

près de 90 % de bonnes réponses (contre 94 % pour les humains). Cette hausse de performance est presque entièrement imputable à l'augmentation de la taille des modèles de langage en réseau neuronal et de la quantité de leurs données d'entraînement.

Ces réseaux encore plus larges ont-ils enfin atteint un niveau de compréhension similaire au nôtre ? Encore une fois, c'est peu probable. Les résultats de WinoGrande s'accompagnent d'importantes mises en garde. Par exemple, parce que les phrases dépendent des travailleurs d'Amazon Mechanical Turk, la qualité et la cohérence de l'écriture sont assez inégales. L'IA utilisée pour filtrer les phrases « non imperméables à Google » peut avoir été trop peu sophistiquée pour repérer tous les potentiels raccourcis statistiques qu'un énorme réseau neuronal pourrait emprunter, et elle ne s'appliquait qu'à des phrases individuelles, si bien que certaines des phrases qui restaient ont fini par perdre leur « jumelle ». Une étude postérieure à ces travaux a montré que les modèles de langage en réseau neuronal testés uniquement avec des phrases jumelles, et devant répondre correctement aux deux, sont bien moins précis que les humains, ce qui indique que le résultat de 90 % vu plus tôt est moins significatif qu'il ne pouvait le paraître.

Au final, que retenir de cette saga Winograd ? La leçon principale est qu'il est souvent difficile de déterminer, à partir de leur performance lors d'un défi donné, si des systèmes d'IA comprennent véritablement le langage (ou d'autres données) qu'ils traitent. Nous savons que les réseaux neuronaux utilisent souvent des raccourcis statistiques – au lieu de vraiment faire preuve d'une compréhension semblable à celle des humains – pour obtenir de bonnes performances sur les schémas Winograd et sur d'autres bancs d'essais orientés vers une « compréhension générale du langage ».

Le nœud du problème, à mon avis, est que comprendre le langage nécessite de comprendre le monde, et notamment ce que signifie que « la voiture de sport a doublé la camionnette de la poste parce qu'elle était plus lente ». Cela suppose de savoir ce que sont des voitures de sport et des camionnettes de la poste, que des voitures peuvent se « doubler », et, à un niveau encore plus fondamental, que les véhicules sont des objets qui existent et interagissent dans le monde, conduits par des humains avec leurs propres objectifs.

Ces connaissances, nous, humains, les tenons pour acquises, mais ce n'est pas le cas des machines. Et il est peu probable que ce soit écrit explicitement dans le texte d'entraînement de n'importe quel modèle de langage. Certains chercheurs en cognition estiment que les humains, pour apprendre le langage, s'appuient sur un noyau de connaissances, prélinguistiques et innées, de l'espace, du temps et de nombreuses autres propriétés essentielles du monde. Si nous voulons que les machines maîtrisent le langage comme nous, nous devrons d'abord les doter des principes primordiaux avec lesquels nous naissons. Et pour évaluer leur niveau de compréhension, nous devrions commencer par évaluer leur capacité à saisir ces principes, ce qu'on pourrait appeler une « métaphysique infantile ».

Entraîner et évaluer des machines au niveau d'intelligence d'un nourrisson peut apparaître comme un pas de géant en arrière par rapport aux prouesses de Watson et de GPT-3. Mais si l'objectif est une compréhension authentique et digne de confiance, il se peut que ce soit le seul chemin vers des machines vraiment capables de comprendre à quoi « il » ou « elle » fait référence dans une phrase.

— L'autrice —

> Melanie Mitchell est professeuse de complexité à l'institut de Sante Fe, au Nouveau-Mexique, aux États-Unis.

— À lire —

> Y. Elazar *et al.*, Back to square one: Artifact detection, training and common sense disentanglement in the Winograd schema, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

> B. Taira *et al.*, A pragmatic assessment of Google Translate for emergency department instructions, *Journal of General Internal Medicine*, 2021.

> P. Amsili et O. Seminck, Schémas Winograd en français : une étude statistique et comportementale, *Actes de TALN 2017*, 2017.

> T. Brown *et al.*, Language models are few-shot learners, https://arxiv.org/abs/2005.14165.

La reconnaissance automatique fondée sur l'apprentissage profond a d'étonnantes faiblesses. Exploiter ces failles permet de s'amuser, mais aussi d'améliorer les procédures d'apprentissage… ou de concevoir de nouvelles attaques malveillantes.

Un seul pixel vous manque et tout est désappris

Jean-Paul Delahaye

© Shutterstock

L'apprentissage automatique à l'aide de réseaux neuronaux « profonds » est à la mode. Comme souvent quand il s'agit d'intelligence artificielle, un enthousiasme excessif s'est manifesté. Ce fut déjà le cas avec les premiers succès des méthodes d'apprentissage automatique dans la décennie 1950 ; les systèmes experts dans les années 1980 ; puis quelque temps après avec les réseaux de neurones ; et maintenant, c'est une forme de ces réseaux qui fait croire que nous sommes sur le point de mettre dans nos ordinateurs une intelligence générale susceptible de nous surpasser.

Sans dénigrer une remarquable technique, nous aimerions ici remettre sur terre les pieds de ceux qui l'imaginent comme une panacée informatique. Les neurones formels proposés en 1959 sont des modèles informatiques et simplifiés des neurones de notre cerveau. On conçoit des programmes qui en simulent un grand nombre en les regroupant en couches successives, comme les neurones du cerveau humain *(voir les Repères p. 6)*. Les neurones formels de la couche d'entrée du réseau reçoivent des informations, par exemple sous la forme d'images décomposées en pixels : chaque pixel est une entrée. En fonction de leurs paramètres internes, les neurones de cette première couche envoient des signaux aux neurones de la deuxième couche, qui eux-mêmes, en fonction de leurs paramètres internes, envoient des signaux à la troisième couche, etc.

Les influx de sortie, c'est-à-dire de la dernière couche, désignent les réponses possibles. Ce sont par exemple des lettres A, B, C …: si les images données en entrée sont des lettres manuscrites et imprécises, le but recherché est que le réseau, une fois instruit, sache correctement reconnaître les lettres qui lui sont proposées. La phase d'apprentissage consiste à ajuster les paramètres internes de chaque neurone pour que le réseau réponde correctement. Souvent, plusieurs milliers de paramètres sont à calculer. Tout l'art de la programmation des réseaux de neurones consiste, à l'aide d'un grand nombre d'images avec leurs classifications (par exemple des formes manuscrites dont on indique au réseau à quelles lettres elles correspondent), à faire converger les jeux de paramètres des neurones vers un état à peu près stabilisé et qui sera tel que lorsqu'on proposera des images, par exemple d'autres lettres manuscrites non utilisées, elles soient correctement identifiées.

On parle d'apprentissage supervisé : lors d'une phase initiale, on indique au réseau les réponses voulues ; il apprend et, si tout s'est bien passé, devient alors capable de trouver, seul, les bonnes réponses à des entrées nouvelles. Un autre type de méthode pour instruire un réseau de neurones formels se fonde sur l'idée de le récompenser quand il propose une bonne réponse et de le pénaliser dans le cas contraire ; ce sont les méthodes d'« apprentissage par renforcement ».

La puissance de calcul disponible aujourd'hui, les bases de données volumineuses collectées pour les exemples utilisés en phase d'apprentissage, ainsi que la multiplication des couches internes d'un réseau (on en utilise parfois plusieurs dizaines) ont conduit récemment à de spectaculaires réussites, comme la victoire des machines sur les humains au jeu de go.

En bref

> Les succès de l'apprentissage profond laissaient penser qu'un réseau de neurones bien entraîné à reconnaître une catégorie d'images ferait preuve de robustesse.

> Or de légères modifications introduites dans les images parviennent à le tromper, là où l'œil humain n'est pas dupe. La modification d'un seul pixel peut suffire.

> Ces images pièges peuvent exposer des systèmes informatiques à des attaques. Mais elles permettent aussi de forger des parades.

Performances : l'exemple du système CaptionBot

L'identification automatique d'objets sur une photo a fait récemment de grands progrès. En 2018, nous avions testé l'outil CaptionBot, conçu par Microsoft. Le système combine deux réseaux neuronaux. Le premier s'occupe de reconnaissance d'images, le second du traitement du langage naturel pour composer les réponses. En prenant en compte un grand nombre d'images accompagnées de leur description, le logiciel apprend à associer les caractéristiques de l'image à des descriptions écrites de ce qu'elles montrent. Il tente ensuite de faire de même avec les nouvelles images qui lui sont soumises. Voyez ci-dessous les résultats, traduits en français.

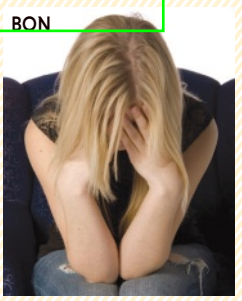
BON

Je pense que c'est un arbre avec en fond un coucher de soleil.

BON

Je pense que c'est une personne tenant une guitare.

BON



Je pense que c'est une femme assise sur un canapé.

PAS SI MAL

Je pense que c'est un homme sur une plage.

PAS SI MAL

Je n'en suis pas très sûr, mais je pense que c'est un ours brun assis sur une table.

RATÉ

Je pense que c'est un grand étalage de maïs.

RATÉ

Je pense que c'est une paire de chaussures bleues.

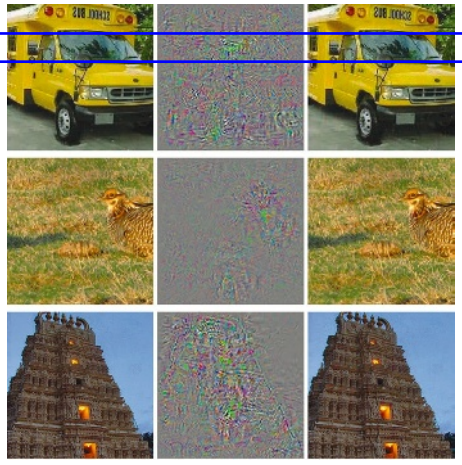
RATÉ

Je n'en suis pas très sûr, mais je pense que c'est un avion dans le ciel.

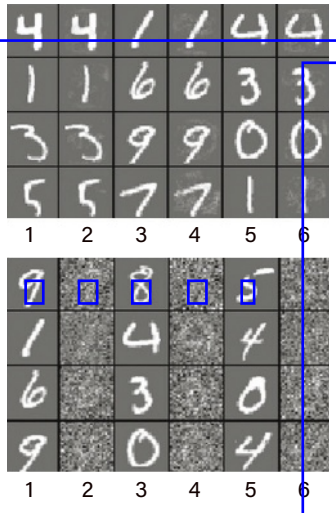
© Nasa, Jens Hackmann; © Shutterstock/@erics; © Shutterstock/Rawpixel; © Shutterstock/iofoto; © Shutterstock/David Tadevosian; © Pixabay; © Shutterstock/Billion Photos; © Superstudio

DES IMAGES PIÈGES

Un procédé de modification d'images permet de piéger les algorithmes neuronaux de reconnaissance d'images et les conduit à produire des réponses absurdes.



Les images proposées dans la colonne de gauche de la figure *ci-contre*, après la phase d'apprentissage du réseau de neurones, sont toutes correctement reconnues. Christian Szegedy et son équipe calculent alors des perturbations *(colonne centrale)* qui, appliquées aux images de la colonne de gauche, donnent les images de la colonne de droite. Ces dernières sont alors, dans chaque cas, reconnues par le système comme figurant… des autruches !



Dans le premier tableau de la figure *ci-contre*, les chiffres manuscrits des colonnes 1, 3 et 5 sont identifiés correctement par le réseau de neurones (après apprentissage). En revanche, le réseau identifie mal les mêmes images qui ont subi de légères modifications (colonnes 2, 4 et 6), alors que l'œil humain les identifie aussi bien et ne voit presque pas de différences avec les premières images. Dans le second tableau, on a placé en colonnes 2, 4 et 6 les images des colonnes 1, 3 et 5 soumises à un assez fort bruit aléatoire. Le réseau continue pourtant de fournir les bonnes identifications dans plus de la moitié des cas, alors que les humains en sont incapables.

Ces deux tableaux montrent donc que ce qu'un réseau de neurones apprend dans sa phase d'apprentissage est d'une nature assez différente de ce que nous-mêmes apprenons.

Les chercheurs du domaine étaient persuadés qu'un réseau ayant bien appris était nécessairement robuste et doté d'une propriété de continuité : des traits dessinant de manière vague et approchée une lettre seront reconnus correctement, et une image proche d'une image reconnue sera elle aussi reconnue.

Or ce n'est pas le cas : les chercheurs ont mis au point des images pièges, conçues pour qu'un réseau ayant appris à reconnaître une certaine catégorie d'images propose une réponse incongrue et fausse quand on lui soumet une image proche, qu'un humain classe aisément *(voir l'encadré ci-contre)*.

Insistons sur la prudence nécessaire. Yann LeCun, pionnier de l'« apprentissage profond », qui dirige la recherche en intelligence artificielle chez Facebook, est clair : « C'est un abus de langage que de parler de neurones ! De la même façon qu'on parle d'aile pour un avion comme pour un oiseau, le neurone artificiel est un modèle extrêmement simplifié de la réalité biologique. » Il ajoute : « Les intelligences artificielles les plus abouties ont aujourd'hui moins de sens commun qu'un rat ! »

En somme, les réussites des réseaux de neurones, des méthodes d'apprentissage automatique et des méthodes de fouille de données massives (les *big data*) sont de nouveaux exploits des machines, mais ces succès ne signifient pas que les machines sont devenues véritablement intelligentes.

Revenons à nos exemples pièges. Une équipe de sept chercheurs affiliés aux universités de New York et Montréal, à Google et à Facebook, réunie autour de Christian Szegedy, a publié en 2014 une méthode permettant de tromper les systèmes de reconnaissance visuelle.

© Images extraites de C. Szegedy et al., Intriguing properties of neural networks, International Conference on Learning Representations, 2014, prépublication arXiv:1312.6199, 2014.

Les intelligences artificielles les plus abouties ont aujourd'hui moins de sens commun qu'un rat

Yann LeCun, patron de la recherche en IA de Facebook

Les chercheurs en réseaux neuronaux affirment volontiers que les réseaux « généralisent ». Si un réseau apprend à reconnaître un cheval à partir d'un ensemble de photos de chevaux, il aura la capacité de reconnaître un cheval sur des photos qu'il n'a jamais vues. Il semblait alors aller de soi qu'un réseau qui classe correctement une photo de cheval donnée classera correctement la même photo où quelques pixels seulement ont été modifiés.

L'ÉTRANGE EXPÉRIENCE DE 2014

Ce n'est pas le cas : le réseau indiquera parfois qu'il voit une vache à la place du cheval, que tous les humains identifient pourtant. Si l'hypothèse de continuité « Ce qui est proche d'une image correctement traitée l'est aussi » jusque-là admise était correcte, alors aucune méthode piège de ce type ne pourrait fonctionner. Il est troublant que la technique du groupe de Christian Szegedy fonctionne presque toujours et puisse fabriquer des exemples trompeurs pour une vaste catégorie de réseaux de neurones et de données d'apprentissage.

Pour construire les images pièges, on recourt à une fonction qui estime le risque d'erreur pour chaque image possible, puis on utilise un algorithme d'optimisation qui, de petite modification en petite modification, déforme l'image en augmentant le plus possible ce risque d'erreur, tout en changeant le moins possible l'image manipulée. De proche en proche, sans que l'œil humain ne perçoive de changements notables, l'algorithme construit une image qui fera trébucher le réseau.

Comme souvent en recherche, quand une bonne idée est découverte, une multitude de travaux la précisent et la prolongent. Depuis l'article de 2014, des dizaines d'équipes ont trouvé et perfectionné toutes sortes d'astuces produisant des images pièges, et cela quelle que soit la méthode d'apprentissage mise en œuvre pour instruire le réseau. Dans certains cas, un seul pixel modifié induit en erreur le réseau (*voir l'encadré page 66*). Assez souvent, une image qui piège un réseau trompera aussi d'autres réseaux ayant appris selon une méthode différente. Christian Szegedy explique : « Nos exemples pièges sont relativement robustes, et trompent aussi des réseaux de neurones ayant un nombre différent de couches que celui qui les a engendrés ou provenant de phases d'apprentissage n'utilisant que des sous-ensembles des données. »

En apprentissage automatique, l'erreur que commettent les systèmes est parfois due à ce qu'on nomme le « surajustement » (*overfitting*). La méthode apprend au système à réagir sur un nombre limité d'exemples, mais dès qu'on en sort, le système se trompe, un peu comme une courbe que l'on force à passer par tous les points de mesure disponibles et qui est incapable de prédire correctement une nouvelle valeur.

Cependant, on a du mal à y voir la bonne explication des exemples pièges, puisque la classification automatique par le réseau fonctionne en général. Ce qui se produit est plus subtil qu'un surajustement : le réseau a généralisé les exemples utilisés dans la phase d'instruction, mais cette généralisation a laissé des failles que détectent les méthodes d'optimisation

employées pour le piéger, en faisant apparaître des cas d'erreur inattendus.

De surcroît, on a su fabriquer des objets 3D qui, quand on les photographie, produisent des images pièges : ce n'est pas une image très spécifique qui trompe le réseau, mais toutes celles qu'engendre un objet réel. C'est inquiétant : on pourrait par exemple fabriquer un faux panneau « Stop » qu'un véhicule autonome interpréterait comme un panneau lui donnant la priorité ! Ces images pièges rendent possibles de nouveaux types d'attaques contre des systèmes informatiques intégrant des réseaux de neurones. Parmi les travaux menés, mentionnons ceux ayant établi qu'on peut piéger plusieurs images à la fois : on cherche à optimiser le pixel qui augmente le plus la mesure du risque d'erreur non pas sur une seule image, mais sur un ensemble d'images. On construit ainsi de proche en proche une perturbation unique, une matrice de valeurs, qui, appliquée à chaque image d'un vaste ensemble d'images, produira une image piège.

Ces découvertes donnent des pistes pour créer de nouvelles méthodes d'apprentissage plus résistantes. L'idée la plus simple consiste, après la première période d'apprentissage, à construire de nombreuses images pièges et à les faire apprendre en complément au réseau en lui indiquant les bonnes réponses, tout comme un professeur de langue indiquerait les faux amis à l'étudiant.

Une autre idée simple fournit facilement une première protection : la compression de données.

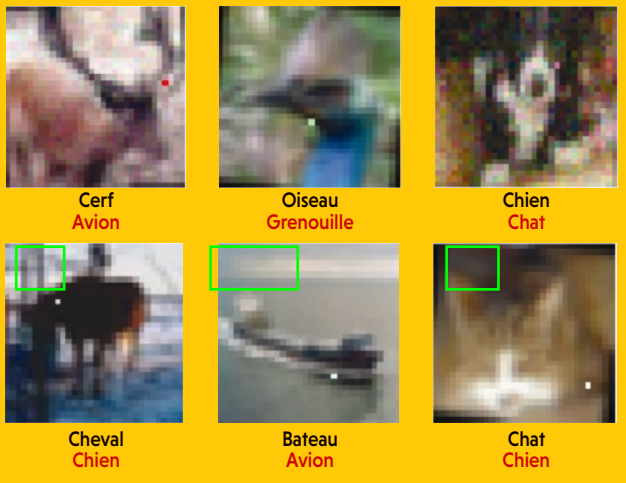
> Les images pièges sont conçues pour qu'un réseau ayant appris à les reconnaître propose une réponse incongrue

Gintare Dziugaite, de l'université de Cambridge, et ses collègues ont noté dès 2016 l'intérêt et les effets des méthodes de compression informatique pour se protéger des images pièges. Avant de faire apprendre une image au réseau, on lui appliquera un algorithme de compression d'images d'usage courant. Cela modifie un peu certains pixels, mais la compression, en simplifiant l'image, la débarrasse des caractéristiques inutiles que le réseau pourrait prendre en compte alors qu'elles ne sont qu'accidentelles.

On se retrouve une fois encore dans une situation où le perfectionnement du glaive conduit aux perfectionnements du bouclier. Quelle sera l'issue finale de cette coévolution ? Nul ne le sait,

TROMPER AVEC UN SEUL PIXEL

Les images composées de peu de pixels sont plus délicates à déchiffrer. Pourtant, nous savons le faire assez bien, comme l'illustrent ces six images en basse résolution, assez facilement identifiables (ici un cerf, un oiseau, un chien, un cheval, un bateau, un chat). Dans un de leurs articles, les chercheurs Jiawei Su, Danilo Vasconcellos Vargas et Sakurai Kouichi expliquent comment, après une phase d'apprentissage conduisant à de bonnes identifications, ils ont réussi, en ne modifiant qu'un seul pixel de chaque image (qu'on voit assez nettement), à berner le réseau, qui a alors reconnu autre chose, respectivement un avion, une grenouille, un chat, un chien, un avion et un chien.



mais cela fera globalement progresser la science de l'apprentissage automatique.

Les réseaux de neurones artificiels sont formidables pour une catégorie de tâches particulières qu'Andrew Ng, chercheur principal chez Baidu, entreprise chinoise, décrit comme suit : « Si un individu moyen est capable de réaliser une certaine tâche en moins de une seconde, alors il est probable que vous pourrez l'automatiser. » En revanche, leur mode de fonctionnement et leur programmation par apprentissage à base d'exemples entraînent des caractéristiques dont plusieurs sont des limitations.

ARTIFICIELLE ET… SUPERFICIELLE !

Gary Marcus, professeur émérite de psychologie à l'université de New York, a énuméré dix problèmes liés aux réseaux de neurones profonds, qui nous aident à percevoir ce qu'il faudrait faire pour aller au-delà.

1. L'apprentissage profond exige énormément de données : lors de la phase d'apprentissage, il faut donner au système plus d'exemples que n'en aurait besoin un humain.

2. L'apprentissage à l'aide de réseaux de neurones profonds est en réalité superficiel ; l'utilisation de l'adjectif « profond » concerne la structure des réseaux utilisés, mais pas la nature de ce qui est appris et maîtrisé.

3. L'apprentissage profond ne sait pas, jusqu'à présent, traiter de manière directe les structures hiérarchiques ou impliquant une combinatoire complexe entre éléments (par exemple, le fonctionnement d'un moteur de voiture).

4. L'apprentissage profond est inefficace pour mener des raisonnements logiques enchaînés, comme déduire de « Lucie est la mère de Jacques et la fille de Jean » que « Jacques est le petit-fils de Jean ». On sait construire des méthodes symboliques (non neuronales) qui réussissent ces tâches de raisonnement qu'il faudra probablement associer aux méthodes neuromimétiques.

5. L'apprentissage profond est opaque. Lorsque le réseau a appris à mener une certaine tâche, des milliers de paramètres ont été fixés et, sauf dans les cas les plus simples, il est impossible de les interpréter et de comprendre pourquoi ils ont les valeurs retenues.

6. L'apprentissage profond ne permet pas, jusqu'à présent, d'intégrer des connaissances *a priori* ou formulées abstraitement, par exemple du type : « Si sur le dos de l'animal il y a une selle, il est assez probable que ce soit un cheval. » Les humains ne comprennent pas tout à l'aide d'exemples, ils intègrent aussi des connaissances élaborées par d'autres auxquels ils font confiance, comme leurs enseignants.

7. L'apprentissage profond ne distingue pas une corrélation d'une cause. Il repère des liens, mais n'est pas en mesure de les analyser.

8. L'apprentissage profond suppose un monde stable, ce qui est ennuyeux dans un univers en mouvement ou en évolution rapide.

9. L'apprentissage profond ne garantit pas les résultats. La découverte de méthodes insoupçonnées qui engendrent des exemples pièges en a été une preuve spectaculaire.

10. L'apprentissage profond n'est pas facile à mettre en œuvre, car on ne sait pas à l'avance ce qui marchera, et à chaque nouvelle application, il faut ajuster les paramètres généraux (profondeur du réseau, nombre de neurones, technique d'apprentissage, etc.) sans savoir *a priori* ce qui fonctionnera le mieux.

Vive l'intelligence artificielle, et bravo pour ses succès récents grâce à l'apprentissage profond ! Mais restons lucides sur ce qui est fait et reste à faire avant de parler de machines en tout point meilleures que les humains.

— L'auteur —

> Jean-Paul Delahaye est professeur émérite à l'université de Lille et chercheur au Centre de recherche en informatique, signal et automatique de Lille (Cristal).

— À lire —

> **S. Hussain *et al.*,** Understanding and mitigating audio adversarial examples. 30th USENIX Security Symposium (USENIX Security 21), 2021.

> **L. Fowl, *et al.*,** Adversarial examples make strong poisons. *Advances in Neural Information Processing Systems*,34, 2021.

> N. Akhtar et A. Mian, Threat of adversarial attacks on deep learning in computer vision : A survey, prépublication arXiv:1801.00553, 2018.

> C. Szegedy *et al.*, Intriguing properties of neural networks, *International Conference on Learning Representations* 2014, prépublication arXiv:1312.6199, 2014.

Entre les performances des réseaux neuronaux et ce qu'on en comprend, c'est le grand écart. Comment la forme d'un réseau conditionne sa fonction est notamment obscur. Les informaticiens travaillent à l'éclairer.

Théorie en voie de construction

Kevin Hartnett

© Koma Zhang

D'un gratte-ciel, nous attendons qu'il respecte un cahier des charges, avec pour *minimum minimorum* de supporter son propre poids et de résister à un séisme. Rien de tel avec l'une des technologies les plus importantes du monde moderne. Avec elle, nous avançons à l'aveuglette, bricolons des configurations, mais tant qu'elle n'a pas été testée, nous ne savons pas vraiment ce qu'elle peut faire, ni où elle échouera. Cette technologie est le réseau de neurones artificiels ; c'est lui qui soutient les systèmes d'intelligence artificielle les plus avancés à ce jour. De plus en plus profondément intégré dans notre société, il filtre le flux des réseaux sociaux que nous recevons, aide les médecins à diagnostiquer des maladies, et influence même la durée d'emprisonnement d'un justiciable condamné pour crime.

En réalité, selon le mathématicien Boris Hanin, « la meilleure approximation de ce que nous savons est que… nous ne savons presque rien à propos du fonctionnement réel des réseaux neuronaux, ni ce que dirait une théorie vraiment pertinente sur le sujet ! » Cet enseignant à l'université de Princeton aime l'analogie avec une autre technologie révolutionnaire : la machine à vapeur. De prime abord, ces machines n'étaient bonnes qu'à pomper de l'eau. Puis, elles ont actionné des locomotives, ce qui correspond peut-être au degré de sophistication atteint aujourd'hui par les réseaux neuronaux. Enfin, les scientifiques et les mathématiciens, grâce à la théorie de la thermodynamique, ont pu comprendre exactement ce qu'il se passe à l'intérieur de n'importe quelle machine. Au final, cette technologie nous a conduits sur la Lune. « D'abord, une belle ingénierie, puis d'excellents trains, puis un peu de compréhension théorique pour en arriver aux fusées », résume Boris Hanin.

Au sein de la communauté tentaculaire des chercheurs en intelligence artificielle, pourtant, un petit groupe d'individus à l'esprit bouillonnant de maths essaye de mettre au point une théorie capable d'expliquer leur fonctionnement et de garantir que, si vous construisez tel réseau neuronal en suivant telle méthode prescrite, il exécutera telles tâches précises. Ce travail n'en est qu'à ses balbutiements, mais, en 2018, plusieurs études ont détaillé la relation entre forme et fonction dans les réseaux neuronaux. L'objectif est d'en définir les fondations mêmes. Il montre que, bien avant de pouvoir certifier qu'ils peuvent conduire des voitures, vous devrez prouver que les réseaux neuronaux sont aptes à réussir une simple multiplication. Voilà qui demande quelques explications.

UN CHIEN, C'EST QUATRE PATTES POILUES

Les réseaux neuronaux visent à imiter le cerveau humain, organe que l'on peut représenter comme une structure qui avance au fil d'abstractions de plus en plus larges. Dès lors, la complexité de la pensée est mesurée par la palette d'abstractions de base sur lesquelles vous pouvez vous appuyer, et le nombre de fois où vous pouvez les combiner en de nouvelles de plus haut niveau – c'est ainsi que nous apprenons à distinguer les chiens des oiseaux.

En bref

> Les réseaux neuronaux ont une architecture inspirée du cerveau sous forme de couches successives de « neurones » reliées entre elles.

> Adapter la profondeur et la largeur des couches est encore très empirique. Il faut beaucoup d'essais et erreurs pour réussir.

> L'un des objectifs majeurs de la discipline est d'établir une théorie sur le lien entre telle forme et tel résultat. Cela nécessite, par exemple, de déterminer la largeur minimale d'une couche.

«Comme être humain, apprendre à reconnaître un chien revient à savoir identifier quatre pattes poilues, analyse Maithra Raghu, chercheuse chez Google Brain. Idéalement, nous aimerions que nos réseaux neuronaux procèdent de la même manière.» L'abstraction vient naturellement au cerveau humain. Pas aux réseaux neuronaux. Ils sont faits de blocs de construction appelés «neurones», connectés de diverses manières. Chacun peut représenter un attribut, ou une combinaison d'attributs, que le réseau prend en compte à chaque niveau d'abstraction. Quand ils relient ces neurones, les ingénieurs doivent faire de nombreux choix, en particulier décider de combien de couches de neurones le réseau doit disposer, autrement dit, quelle est sa «profondeur». Imaginez que l'enjeu soit de reconnaître des objets dans des images. L'image entre dans le système à la première couche. À la couche suivante, le réseau peut avoir des neurones qui détectent simplement les contours dans l'image. Puis la suivante combine les courbes en formes et textures, et la dernière traite formes et textures pour conclure sur la nature de l'objet : c'est un mammouth laineux !

Chaque couche combine plusieurs aspects de la précédente. Ainsi, un cercle peut être vu comme une combinaison de courbes, une courbe comme une combinaison de segments. Les ingénieurs doivent aussi décider de la «largeur» de chaque couche, laquelle correspond au nombre de caractéristiques différentes que le réseau prend en compte à chaque niveau d'abstraction. Dans le cas de la reconnaissance d'image, cette largeur peut être le nombre de types de segments, de courbes et de formes considérés.

Quanta magazine

Cet article a d'abord été publié en anglais par ***Quanta Magazine,*** *une publication en ligne indépendante soutenue par la Simons Foundation afin de favoriser la diffusion des sciences : bit.ly/36dMtRh*

Au-delà de la profondeur et de la largeur d'un réseau, il faut aussi choisir comment câbler les neurones dans et entre les couches, et quel poids donner à chaque connexion. Dans ces conditions, une fois l'objectif fixé, comment savoir quelle architecture de réseau neuronal l'atteindra au mieux ? Quelques règles générales se dégagent. Pour les tâches impliquant des images, les ingénieurs utilisent des réseaux neuronaux «convolutifs», qui offrent le même motif de connexions entre les couches, répété encore et encore. Pour le traitement du langage naturel – reconnaissance automatique de la parole ou génération de textes –, les réseaux neuronaux «récurrents» semblent mieux fonctionner : les neurones peuvent y être connectés à des couches non adjacentes.

LA RÉUSSITE COMME PRODUIT D'ESSAIS ET D'ERREURS

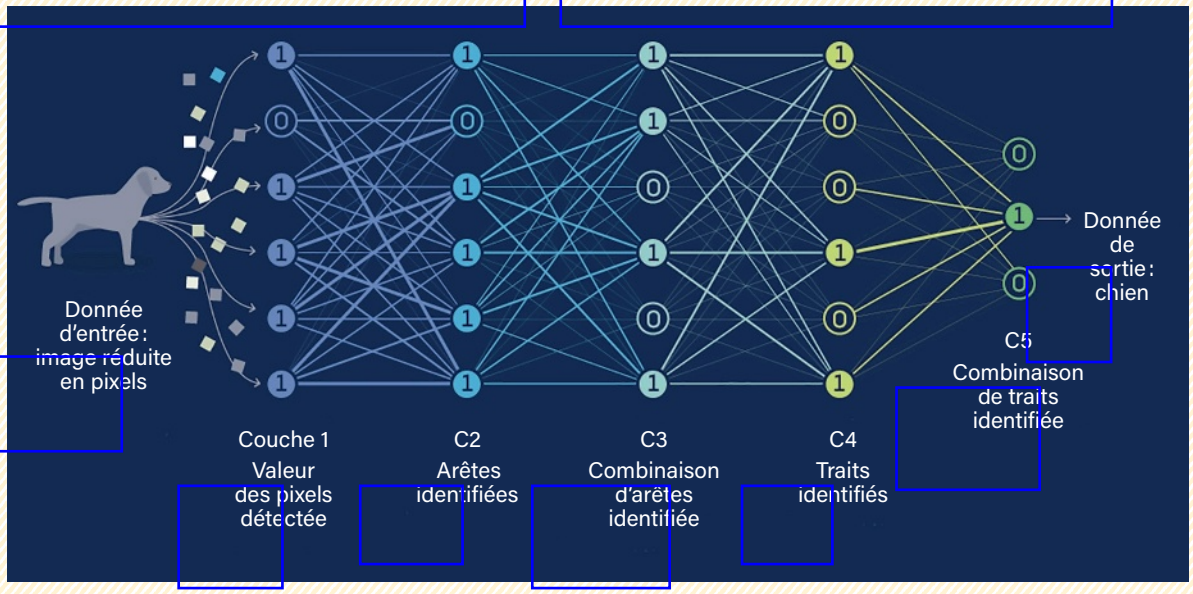
Au-delà de quelques lignes directrices, les ingénieurs n'ont d'autre choix que d'obtenir des preuves expérimentales : ils lancent 1 000 réseaux neuronaux différents et observent simplement lequel fait le boulot. «Leurs choix sont le produit d'essais et d'erreurs, constate Boris Hanin. C'est dur, car il y a un nombre infini de choix possibles et nul ne sait quel est le meilleur.»

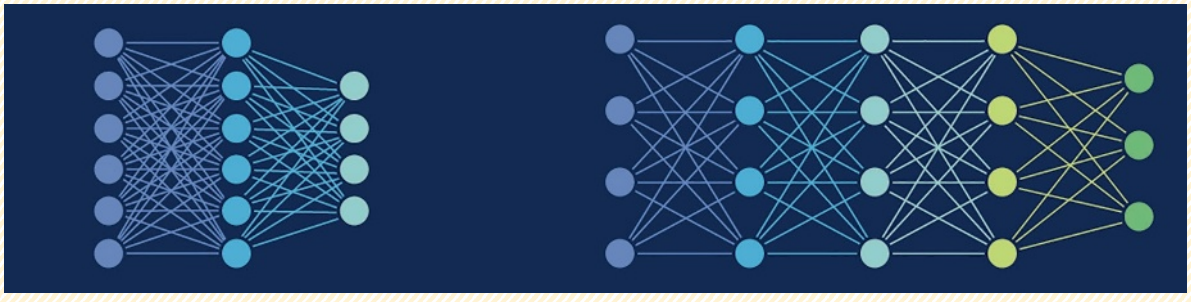
Il conviendrait de réduire la quantité d'essais et d'erreurs et de mieux comprendre au préalable ce que l'architecture d'un réseau neuronal peut apporter. Quelques études récentes font avancer dans cette direction. Selon David Rolnick, aujourd'hui professeur de sciences

© shutterstock.com

Comment on conçoit un réseau de neurones

Un réseau neuronal fait circuler des données d'entrée, par exemple des images, à travers des couches de neurones numériques. Chacune d'elles met en évidence de nouvelles caractéristiques de la donnée d'entrée. En fonction de l'architecture mathématique du réseau, c'est-à-dire le nombre de neurones, de couches et leurs interconnexions, les mathématiciens tâchent de définir dans quels types de tâches il excellera. Une fois le réseau alimenté en données, chacun des neurones qui est « excité » (ils sont balisés par un 1 *ci-dessous*) transmet un signal à d'autres neurones de la couche voisine, lesquels s'exciteront à leur tour au-dessus d'un seuil de signaux. Le processus, à base de calculs, extrait des informations abstraites.





Un réseau peu profond, ou expressif, compte peu de couches, mais beaucoup de neurones par couche. Il exige une forte puissance de calcul.

Un réseau profond compte de nombreuses couches, dotées chacune d'assez peu de neurones. Il permet un haut niveau d'abstraction avec une économie de neurones.

© Lucy Reading-Ikkanda

informatiques à l'université McGill, « ces travaux visent à rédiger un livre de recettes qui permette de concocter le bon réseau neuronal. Selon ce que vous voulez en tirer, vous choisirez telle ou telle recette ».

L'une des cautions théoriques les plus anciennes concernant l'architecture des réseaux neuronaux a déjà trente ans. En 1989, des informaticiens ont prouvé que si une seule couche informatique est prévue, mais avec un nombre illimité de neurones et autant de connexions entre eux, alors le réseau sera capable d'accomplir n'importe quelle tâche.

Malgré sa formulation carrée, cette affirmation se révèle intuitive et en définitive pas très utile. Elle revient à dire que, si vous pouvez identifier une quantité illimitée de lignes dans une image, vous pouvez distinguer tous les objets en utilisant seulement une couche. Dans le principe, c'est peut-être vrai, mais pour y arriver en pratique… bonne chance !

Aujourd'hui, les chercheurs qualifient d'« expressifs » ces réseaux plats et larges, ce qui signifie qu'ils sont capables en théorie d'embrasser un jeu plus riche de connexions entre données d'entrée (par exemple, une image) et de sortie (ici, une description de cette image). En pratique, ils sont extrêmement difficiles à entraîner, et leur apprendre à produire réellement ces données de sortie est presque impossible. Leur avidité en puissance de calcul dépasse aussi la capacité de nos ordinateurs.

LARGEUR OU PROFONDEUR ?

Plus récemment, les chercheurs ont tenté de cerner jusqu'où ils peuvent pousser les réseaux neuronaux dans la direction opposée : réduire la largeur (moins de neurones par couches) et augmenter la profondeur (davantage de couches). Peut-être suffit-il de repérer 100 lignes différentes, du moment qu'il y a assez de connexions pour transformer ces 100 lignes en 50 courbes, que vous pouvez combiner en 10 formes différentes, lesquelles vous donnent tous les blocs élémentaires nécessaires pour reconnaître la plupart des objets.

En 2018, David Rolnick et Max Tegmark, alors tous deux à l'institut de technologie du Massachusetts (MIT), ont prouvé que plus de profondeur et moins de largeur permettent d'accomplir des fonctions équivalentes, avec une réduction exponentielle du nombre de

L'abstraction vient naturellement au cerveau humain. Pas aux réseaux neuronaux

neurones. Si vous avez 100 variables d'entrée, on peut obtenir la même fiabilité en utilisant 2^{100} neurones dans une couche ou avec simplement 2^{10} neurones répartis sur deux couches. Combiner de petits éléments à de plus grands niveaux d'abstraction offre plus de potentiel que saisir tous les niveaux d'abstraction d'un seul coup.

« La notion de profondeur dans un réseau neuronal traduit l'idée que vous pouvez réduire n'importe quoi de compliqué en une série de choses plus simples, explique David Rolnick. Comme sur une chaîne de montage. »

David Rolnick et Max Tegmark ont prouvé l'utilité de la profondeur en demandant à des réseaux neuronaux de mener à bien une tâche simple : multiplier des fonctions polynomiales. Il s'agit d'équations dont certaines variables sont élevées à une puissance entière, par exemple $y = x^3 + 1$. Ils ont entraîné les réseaux en leur montrant des exemples d'équations et leur produit. Puis ils leur ont soumis des équations inédites. Les réseaux neuronaux les plus profonds ont appris cette tâche avec beaucoup moins de neurones que ceux qui l'étaient moins.

Bien qu'une multiplication ne soit pas une opération renversante, David Rolnick juge que l'étude marque un point important : « Si un réseau peu profond est incapable de multiplier, nous ne devrions lui accorder aucune confiance. »

D'autres chercheurs évaluent la largeur minimale nécessaire. Toujours en 2018, le mathématicien Jesse Johnson, qui fut chercheur à l'université d'État de l'Oklahoma avant d'être recruté par Sanofi, a prouvé que, à partir

> Si un réseau peu profond est incapable de multiplier, nous ne devrions lui accorder aucune confiance

d'un certain point, nulle profondeur ne peut compenser le manque de largeur.

MOUTONS PUNK ROCK

Pour comprendre, imaginez des moutons dans une prairie. Des moutons spéciaux, du genre punk rock : chacun est teint en une couleur parmi une quantité possible. La tâche de votre réseau neuronal est de dessiner une limite autour de tous les moutons de même couleur. Dans l'idée, cela ressemble à une classification d'images : le réseau a une collection d'images (qu'il représente comme des points dans un espace de dimension supérieure), et il doit regrouper celles qui se ressemblent.

Jesse Johnson a démontré qu'un réseau neuronal échouera à cette tâche si la largeur des couches est inférieure ou égale au nombre d'entrées. Chaque mouton peut être décrit avec deux entrées : deux coordonnées x et y pour spécifier la position dans la prairie. Le réseau neuronal étiquette chaque mouton avec une couleur et circonscrit ceux qui ont la même couleur. Il vous faudra alors au moins trois neurones par couche pour résoudre ce problème.

Plus précisément, Jesse Johnson a montré que si le ratio largeur/variable n'est pas bon, le réseau neuronal sera incapable de tracer des boucles fermées – du genre de celle qu'il aurait besoin de dessiner si, par exemple, tous les moutons rouges étaient groupés au milieu du pâturage. « Si aucune des couches n'est plus large que le nombre de dimensions d'entrée, certaines formes seront impossibles à créer, quel que soit le nombre de couches dont vous disposez », précise Jesse Johnson.

Ce type d'étude jette les bases d'une théorie des réseaux neuronaux. Pour l'heure, les chercheurs ne peuvent que dresser des constats élémentaires sur la relation entre architecture et fonction – et ces constats sont bien peu nombreux au regard de la quantité de tâches que les réseaux neuronaux mènent à bien. Si la manière dont les systèmes sont conçus ne changera pas de sitôt, on discerne déjà les esquisses d'une nouvelle théorie sur l'apprentissage des ordinateurs – une théorie qui promet à l'humanité un voyage aux répercussions plus grandes qu'un aller-retour vers la Lune.

— L'auteur —

> Kevin Hartnett est journaliste à *Quanta Magazine*.

— À lire —

> J. Chul Ye, Geometry of Deep Learning, *Springer*, 2022.

> T. Nguyen *et al.*, Do wide and deep neural networks learn the same things? Uncovering how neural network representations vary with width and depth, *International Conference on Learning Representations (ICLR)*, 2021.

> B. Hanin et D. Rolnick, Complexity of linear regions in deep networks, International *Conference on Machine Learning*, 2019.

> J. Johnson, Deep, skinny neural networks are not universal approximators, 2018. arxiv:1810.00393

ABONNEZ-VOUS À Pour la Science

OFFRE D'ABONNEMENT 1 AN

2 FORMULES AU CHOIX

	FORMULE PAPIER + HORS-SÉRIE	FORMULE INTÉGRALE
Le magazine papier 12 numéros par an	✓	✓
Le magazine en version numérique 12 numéros par an		✓
Le hors-série papier 4 numéros par an	✓	✓
Le hors-série en version numérique 4 numéros par an		✓
Accès à pourlascience.fr actus, dossiers, archives depuis 1996		✓
VOTRE TARIF D'ABONNEMENT	**79 €** Au lieu de ~~114,40 €~~	**99 €** Au lieu de ~~174,40 €~~
	31 % de réduction *	**43 %** de réduction *

BULLETIN D'ABONNEMENT

À renvoyer accompagné de votre règlement à :

Service Abonnement Groupe Pour la Science – 235 avenue Le Jour se Lève– 92100 Boulogne Billancourt – email : serviceclients@groupepourlascience.fr

PAG22STDDOS

OUI, je m'abonne pour 1 an à Pour la Science

1 / Je choisis ma formule (merci de cocher)

☐ FORMULE **PAPIER + HORS SÉRIE** — **79€** Au lieu de ~~114,40€~~
- 12 n° du magazine papier
- 4 Hors-série papier

☐ FORMULE **INTÉGRALE** — **99€** Au lieu de ~~174,40€~~
- 12 n° du magazine (papier et numérique)
- 4 Hors-série (papier et numérique)
- Accès illimité aux contenus en ligne

LF_1PAP_NTEN_99€
LF_INT_NTEN_99€

2 / J'indique mes coordonnées

☐ M. ☐ Mme

Nom : ..... Prénom : .....

Adresse : .....

.....

Code postal ..... Ville : .....

Téléphone .....

Courriel : (indispensable pour la formule intégrale)

..... @ .....

J'accepte de recevoir les offres de Pour la Science ☐ OUI ☐ NON

3 / Je choisis mon mode de règlement

☐ **Par chèque** à l'ordre de *Pour la Science* en nous retournant ce bulletin d'abonnement complété.

Pour un paiement par carte bancaire, rendez-vous sur boutique.groupepourlascience.fr

* Réduction par rapport au prix de vente en kiosque et l'accès aux archives numériques. Durée d'abonnement : 1 an. Délai de livraison : dans le mois suivant l'enregistrement de votre règlement. Offre valable jusqu'au 31/03/2023 en France métropolitaine uniquement. Pour un abonnement à l'étranger, merci de consulter notre site boutique.groupepourlascience.fr. Photos non contractuelles. Vous pouvez acheter séparément les numéros de *Pour la Science* pour 6,90 € et les hors-séries pour 7,90 €. En souscrivant à cette offre, vous acceptez nos conditions générales de vente disponibles à l'adresse suivante : https://rebrand.ly/CGV-PLS.

Les informations que nous collectons dans ce bulletin d'abonnement nous aident à personnaliser et à améliorer les services que nous vous proposons. Nous les utiliserons pour gérer votre accès à l'intégralité de nos services, traiter vos commandes et paiements, et vous faire part notamment par newsletters de nos offres commerciales moyennant le respect de vos choix en la matière. Le responsable du traitement est la société *Pour la Science*. Vos données personnelles ne seront pas conservées au-delà de la durée nécessaire à la finalité de leur traitement. *Pour la Science* ne commercialise ni ne loue vos données à caractère personnel à des tiers. Les données collectées sont exclusivement destinées à *Pour la Science*. Nous vous invitons à prendre connaissance de notre charte de protection des données personnelles à l'adresse suivante : https://rebrand.ly/charte-donnees-pls. Conformément à la réglementation applicable (et notamment au Règlement 2016/679/UE dit « RGPD ») vous disposez des droits d'accès, de rectification, d'opposition, d'effacement, à la portabilité et à la limitation de vos données personnelles. Pour exercer ces droits (ou nous poser toute question concernant le traitement de vos données personnelles), vous pouvez nous contacter par courriel à l'adresse protection-donnees@pourlascience.fr.

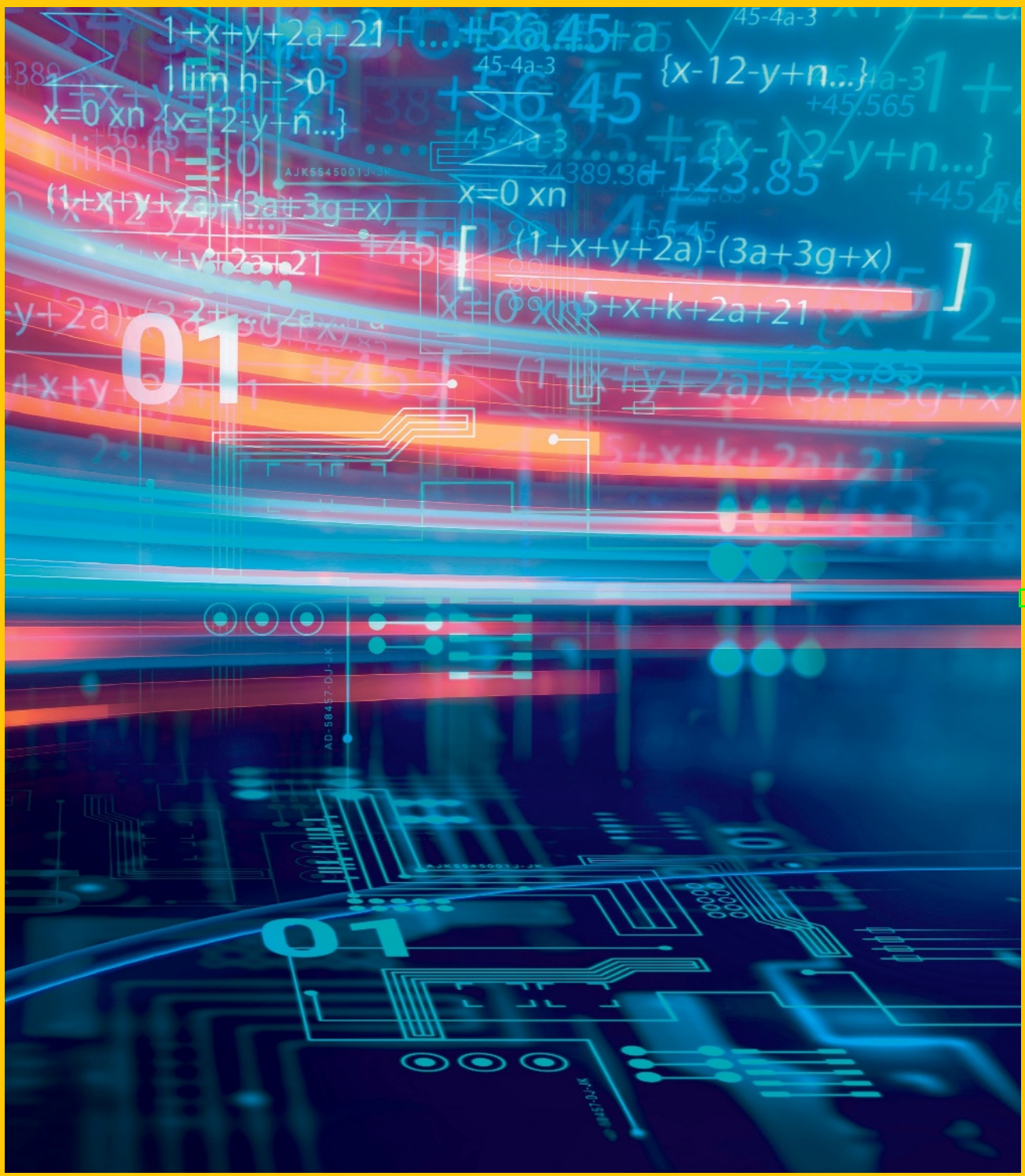
Groupe Pour la Science - Siège social : 170 bis, boulevard du Montparnasse, CS20012, 75680 Paris CEDEX 14 - Sarl au capital de 32 000€ - RCS Paris B 311 797 393 - Siret : 311 797 393 000 23 - APE 58.14 Z

Surtout cantonnée aux tâches d'assistance du mathématicien, l'IA commence à dévoiler un potentiel plus riche pour les mathématiques. Ce qui lui manque encore ? Une vision globale et profonde de cette science.

L'ordinateur sera-t-il un jour le boss des maths ?

Jean-Paul Delahaye

© WhiteMocca/shutterstock


1+x+y+2a+21
1 lim h-->0
x=0 xn {x-12-y+n...}
(1+x+y+2a)-(3a+3g+x)
x=0 xn
[(1+x+y+2a)-(3a+3g+x)]
+x+k+2a+21
01
01

En bref

> L'ordinateur est un auxiliaire précieux des mathématiciens. En facilitant certaines tâches, il fait gagner du temps.

> Des programmes comme *HR* sont parvenus à proposer des notions nouvelles, non incluses dans les mécanismes de leur programmation.

> Ces avancées concernent surtout les structures finies (graphes), mais aussi certains ensembles infinis (nombres réels).

> L'association de méthodes symboliques et neuronales est prometteuse, même si, pour l'heure, les réussites sont limitées à des domaines très spécialisés.

Les ordinateurs aident les mathématiciens d'une multitude de manières : ils facilitent l'édition des textes et la communication entre chercheurs ; ils font des calculs numériques et symboliques, en produisant par exemple des milliers de milliards de décimales de π ; grâce aux programmes « assistants de preuves », ils aident les mathématiciens à mettre au point les longues démonstrations et à les vérifier ; parfois, les algorithmes de démonstration automatique découvrent des preuves qu'aucun mathématicien n'avait trouvées *(voir l'encadré page ci-contre)*.

Cependant, à côté de cette aide exploitant la fiabilité et la rapidité des machines, est-ce que celles-ci peuvent se substituer aux chercheurs et faire avancer un domaine mathématique en y découvrant de nouvelles idées, de nouvelles formules, de nouvelles conjectures ou de nouveaux théorèmes intéressants ?

On a précisé « intéressants », car on sait depuis longtemps produire mécaniquement des théorèmes en utilisant des programmes. À chaque système axiomatique, celui de l'arithmétique, celui de la théorie des ensembles, etc., est associé un algorithme qui en énumère les théorèmes. Malheureusement ces algorithmes d'énumération produisent des masses de résultats évidents et sans valeur.

L'intelligence, le savoir particulier et l'intuition d'un mathématicien lui font aborder un domaine avec finesse en repérant des énoncés particuliers, en général assez courts et difficiles à démontrer. Ces énoncés sont des jalons dans leurs travaux : ils leur évitent de se perdre dans l'infini des vérités faciles et trop proches les unes des autres. L'intelligence artificielle peut-elle concevoir des méthodes qui égalent les mathématiciens dans le travail d'imagination et de construction ordonnée des théories mathématiques, qui ne sont pas des successions ennuyeuses de vérités triviales comme $1 + 1 = 2$, $1 + 2 = 3$, $1 + 3 = 4$, etc. ?

Nous allons voir que la « recherche automatisée » progresse petit à petit : il se peut que, imperceptiblement, nous entrions dans une nouvelle ère.

PREMIÈRES TENTATIVES

Le logicien Hao Wang, un ami d'Einstein à Princeton, écrivit en 1959 un programme pour produire et sélectionner des théorèmes du calcul propositionnel, une partie de la logique où l'on traite de formules du type (A et B) $\Rightarrow$ A, (A ou non-A), (A $\Rightarrow$ B) $\Rightarrow$ (non-B $\Rightarrow$ non-A). Cette tentative lui fit prendre conscience que le problème principal était de sélectionner les résultats « intéressants ».

En juillet 1976, Douglas Lenat soutint à Stanford un mémoire de doctorat dont le sujet était son « mathématicien automatique », le programme *AM* (pour *Automated Mathematician*). Il y décrivait un programme en langage LISP, qui, à partir de règles et de manipulations symboliques, déduit et accumule concepts et propriétés de ces concepts. Douglas Lenat explique que son programme a découvert la notion de nombre premier, a énoncé le théorème de décomposition d'un nombre en facteurs premiers. Il aurait aussi redécouvert la fameuse conjecture de Goldbach, toujours non démontrée (« Tout nombre pair à partir de 4 est somme de deux nombres premiers »). Cependant, ces affirmations ont été

discutées, car elles résultent d'une interprétation indirecte *a posteriori* des calculs d'AM.

Après ce premier essai, Douglas Lenat travailla sur un système nommé Eurisko, lui aussi destiné à produire des découvertes mathématiques. Son originalité est qu'il créait spontanément des règles heuristiques nouvelles à l'aide de métarègles du type : « Donner la priorité aux règles les plus simples ». Face aux difficultés, Douglas Lenat entreprit de concevoir et programmer un système dénommé Cyc, qui maîtrise des connaissances générales de tout humain, ce qu'il considérait comme indispensable pour aller plus loin en intelligence artificielle (IA). Pour lui, un système intelligent doit savoir que « si un humain marche, il ne reste pas à la même place » ou que « quand il pleut, les passants dans la rue sont mouillés », etc.

QUELS TYPES DE RECHERCHES ?

L'intelligence artificielle en mathématiques a trois objectifs :

- L'aide à la démonstration et à la rédaction détaillée et contrôlée de démonstrations compliquées. Dans ce domaine, des succès ont été obtenus, comme la mise au point définitive de la démonstration de la conjecture de Kepler sur l'empilement le plus compact possible des sphères dans l'espace, démonstration due à Thomas Hales en 2014.
- La démonstration automatique : l'ordinateur recherche seul une démonstration. Quelques succès ont été obtenus.
- La recherche et la suggestion d'idées, de conjectures et de théorèmes nouveaux et intéressants. C'est ce dernier aspect qui est principalement envisagé dans cet article.

Le projet Cyc, démarré en 1983, a bénéficié depuis d'un travail équivalent à plus de mille hommes-années et se poursuit aujourd'hui encore. C'est le plus ambitieux et long projet d'IA jamais mené. Il a développé un savoir général et plusieurs savoirs particuliers liés à des sous-domaines, dont celui du terrorisme qui est probablement utilisé par l'agence américaine NSA (National Security Agency). Le système connaît plus d'un million de termes comme « la France », « le kilogramme », « New York », « les poissons », « la chasse »… et plus de 24 millions d'affirmations de base les concernant et les liant. Elles sont écrites dans un langage logique et disposent d'un système de déduction qui permet, à partir des affirmations de base, d'en produire de nouvelles.

HR S'ATTAQUE AUX SUITES

Cette recherche appartient à l'« IA symbolique » : les méthodes ne s'appuient pas sur des réseaux de neurones, ce sont des programmes manipulant des symboles, des mots, des concepts représentés par des éléments de programmes. Les humains qui les écrivent comprennent le détail de ce qu'ils y mettent et de ce qui se produit quand la machine fonctionne, contrairement aux réseaux de neurones de l'« apprentissage profond », où le système n'explique pas ses réponses. Le lien avec les mathématiques est devenu secondaire dans le projet Cyc, bien qu'un module ait été conçu pour aider les étudiants en mathématiques.

En tout cas, il semble certain que les progrès de l'IA devront associer la capacité de raisonnement et les connaissances générales de systèmes

FAIRE DES CONJECTURES INTÉRESSANTES

Pour concevoir des systèmes menant une recherche mathématique, il faut faire en sorte qu'ils effectuent des tris. Un tel système ne doit retenir que les propriétés et conjectures les plus intéressantes. Le Britannique Simon Colton, qui a développé le système *HR (voir le texte principal)*, propose six critères pour repérer les conjectures intéressantes :

1) Les conjectures doivent être plausibles. Il faut éliminer celles qui sont fausses parce qu'il existe des exemples simples qui les contredisent.

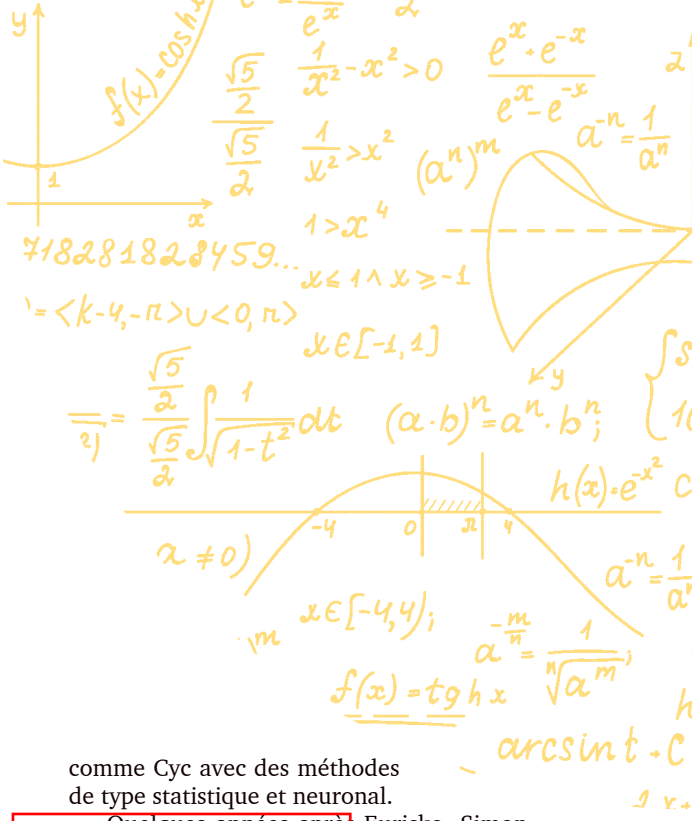
2) Les conjectures doivent être nouvelles. Le système doit disposer des résultats connus sur le sujet étudié et savoir les exploiter pour ne pas proposer les conséquences immédiates de résultats connus.

3) Les résultats envisagés doivent être « étonnants », une notion qui échappe aux programmes. Le plus souvent, on se contente d'éliminer les énoncés qui sont des conséquences immédiates d'énoncés connus ou qui sont des cas particuliers de formules logiques toujours vraies, comme non(non-A)⇔A. Une propriété simple qui exige une longue démonstration est étonnante puisqu'on ne peut pas comprendre aisément ce qui la rend vraie. Malheureusement, la notion « exige une longue démonstration » est impossible à programmer en pratique, car, pour l'affirmer d'un énoncé, il faudrait explorer tous les chemins de démonstrations possibles pouvant y conduire.

4) Les conjectures doivent concerner de nombreux cas.

5) Les conjectures doivent être facilement compréhensibles et, donc, doivent avoir des énoncés assez courts et, si possible, ne faire intervenir que les concepts de base de la théorie étudiée.

6) Les conjectures doivent être fécondes et, par exemple, avoir beaucoup de conséquences simples. En mathématiques, il est difficile d'anticiper cette qualité d'une conjecture. Le grand théorème de Fermat n'est considéré comme fécond que parce qu'il a résisté pendant des siècles. Si Fermat l'avait démontré en quelques lignes, personne n'aurait estimé l'énoncé de ce théorème comme intéressant.



comme Cyc avec des méthodes de type statistique et neuronal.

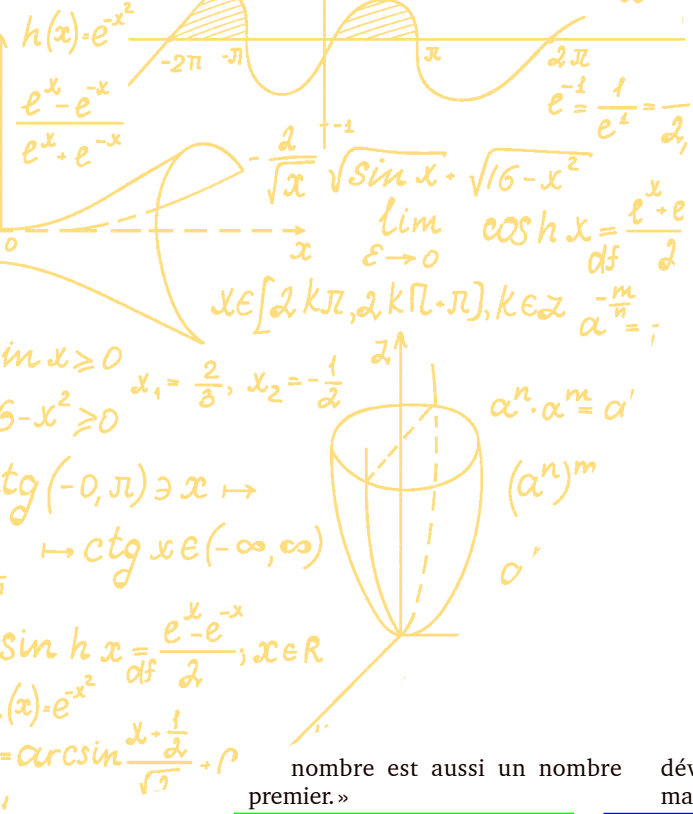
Quelques années après Eurisko, Simon Colton, de l'université d'Édimbourg, a réussi à faire « imaginer » des notions mathématiques nouvelles à son programme, dénommé *HR* en l'honneur des mathématiciens Godfrey Hardy et Srinivasa Ramanujan, génie ayant découvert une multitude de formules extraordinaires.

Le programme *HR* s'est intéressé à l'arithmétique et a tenté à la fois de créer des concepts nouveaux – des définitions de suites numériques nouvelles –, de formuler des hypothèses à leur sujet et parfois de les démontrer à l'aide de programmes de démonstration automatique qu'il mettait à son service. Le programme disposait de critères pour ne se concentrer que sur des définitions potentiellement intéressantes. Il a pu ainsi identifier 14 suites numériques nouvelles, qui ont été acceptées en 2000 par l'Encyclopédie des suites numériques de Neil Sloane (http://oeis.org), où ne figurent que les suites ayant de l'intérêt aux yeux des mathématiciens.

L'une des suites proposées par *HR* est étonnamment simple et a conduit Simon Colton à publier un article dans le *Journal of Integer Sequences* en 1999. Il s'agit de la suite des « nombres refactorisables ». Un entier n est refactorisable s'il est divisible par le nombre de ses diviseurs. L'entier 9 est par exemple refactorisable, car il a 3 diviseurs (1, 3 et 9) et que justement il est divisible par 3. Plus généralement, si p est premier, p^{p-1} est refactorisable.

Notons que le programme de Simon Colton a découvert tout seul cette jolie propriété : « Si la somme des diviseurs d'un entier est un nombre premier, alors le nombre de diviseurs de ce

© Marina Sun/shutterstock



nombre est aussi un nombre premier. »

Le programme *HR* a aussi inventé des notions concernant les groupes, les anneaux finis et les graphes. Il dispose de diverses mesures évaluant l'intérêt d'un concept, se concentre sur le plus prometteur et cherche alors à énoncer des propriétés à son sujet pour lesquelles il tente de trouver des démonstrations ou des contre-exemples. Cette façon de procéder est proche de ce que font souvent les chercheurs en mathématiques, mais partiellement, car un chercheur prend de la distance vis-à-vis de son sujet et tente d'établir des liens entre concepts appartenant à des champs éloignés.

UN VRAI SUCCÈS SUR LES GRAPHES

Le principe général de *HR* est particulièrement efficace quand il s'agit de structures finies, car le programme peut disposer d'une liste variée d'exemples et les utiliser pour tester si les propriétés qu'il envisage sont vérifiées par chaque élément de sa base d'exemples. Si c'est le cas, la propriété devient plausible et le système l'envisage comme conjecture. Si l'un des exemples contredit la propriété examinée, alors il est certain qu'elle est fausse, et il l'abandonne.

La théorie des graphes se prête bien à cette approche, et c'est pourquoi les méthodes automatiques de recherche de conjectures en théorie des graphes ont donné les premiers exemples incontestables de découvertes mathématiques faites par l'ordinateur.

Le programme *Graffiti* est la première grande réussite de cette approche. Il a été développé et utilisé de 1984 à 2002 par le mathématicien polonais Siemion Fajtlowicz.

Graffiti s'intéresse particulièrement aux invariants des graphes. Ce sont, par exemple, le diamètre du graphe (la distance – en nombre d'arêtes – maximale séparant deux nœuds), le nombre chromatique d'un graphe (le nombre minimal de couleurs qu'il faut pour colorer les nœuds du graphe en évitant que deux nœuds voisins soient de même couleur), le degré maximal d'un graphe (nombre maximal de voisins d'un nœud), etc.

En menant des recherches automatisées de relations entre ces invariants, *Graffiti* a proposé de nombreuses conjectures intéressantes, dont certaines ont été démontrées par la suite ou invalidées par des chercheurs de la discipline, donnant lieu à plusieurs publications scientifiques et thèses universitaires.

Siemion Fajtlowicz a d'ailleurs été amené à travailler avec les meilleurs spécialistes de la théorie des graphes comme Paul Erdős, Bela Bollobas ou Paul Spencer. Erdős et ses coauteurs, dans un article de 1988 où ils démontrent l'une des conjectures du programme *Graffiti*, reconnaissent son apport. La conjecture en question affirme que $dm_G \leq r_G$, où dm_G est la distance moyenne entre deux nœuds du graphe G, et r_G est la somme des inverses du nombre de voisins des nœuds de G. Aujourd'hui encore, en théorie des graphes et dans certains autres domaines, l'utilisation de générateurs de conjectures aide les chercheurs.

Un événement remarquable s'est produit en 1995 dans le domaine de l'aide informatique à la recherche mathématique. Simon Plouffe a programmé sa machine pour qu'elle explore des formules d'un certain type donnant le nombre π

et en a effectivement trouvé une que personne ne soupçonnait. Ici, les objets manipulés ne sont pas finis, puisqu'il s'agit de nombres réels ayant une infinité de décimales. Même si la méthode utilisée pour les conjectures sur les graphes ne s'applique pas, le procédé utilisé lui ressemble.

On se donne plusieurs nombres réels dont on connaît des évaluations aussi précises qu'on le veut ; c'est le cas pour π, et pour les sommes de séries dont le terme général est de la forme $1/[r^k(ak+b)]$. On recherche alors s'il existe une combinaison linéaire à coefficients entiers entre les nombres choisis qui soit nulle avec quelques décimales de précision. Quand une relation est trouvée, on teste si elle reste vraie avec un plus grand nombre de décimales. Si la relation est numériquement confirmée, l'algorithme propose la conjecture. En utilisant alors les méthodes classiques, le mathématicien reprend la main et tente de démontrer l'égalité.

C'est ainsi que la formule suivante fut découverte, puis rapidement démontrée :

$$\pi = \sum_{k=0}^{\infty} \frac{1}{16^k}\left(\frac{4}{8k+1} - \frac{2}{8k+4} - \frac{1}{8k+5} - \frac{1}{8k+6}\right)$$

L'intérêt de cette formule est qu'elle permet de calculer les chiffres binaires de π à un endroit donné du développement binaire, sans avoir à calculer les chiffres qui précèdent. Incontestablement, l'ordinateur avait découvert un résultat intéressant.

TROUVER DES DÉMONSTRATIONS

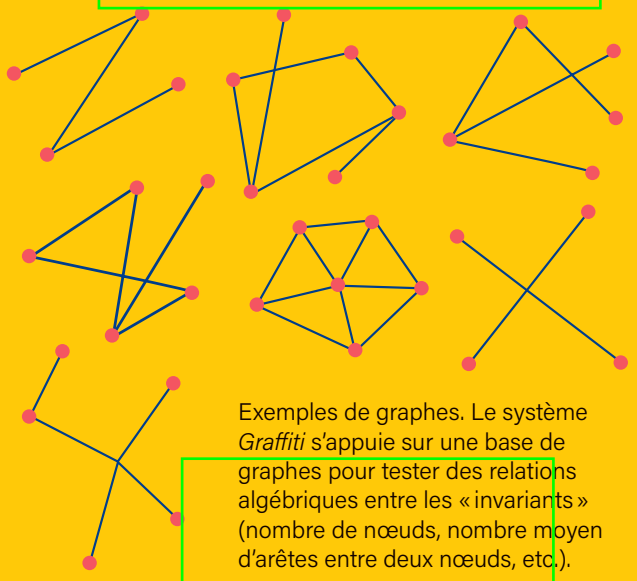
Un certain nombre de tentatives ont été faites pour introduire dans le domaine des mathématiques des méthodes de nature statistique et neuronale.

Kshitij Bansal et son équipe, au centre de recherche de Google à Mountain View, en Californie, ont développé un programme, intitulé *HOlist*, pour démontrer des théorèmes en se fondant sur une méthode d'apprentissage. Partant du système d'aide à la formalisation et au contrôle des démonstrations HOL-Light, qui a été utilisé pour la mise au point de la démonstration de la conjecture de Kepler en 2014, les chercheurs ont programmé un outil automatique qui recherche des démonstrations en s'inspirant d'une base de démonstrations

LA MÉTHODE GRAFFITI

Le programme *Graffiti* énumère et sélectionne des conjectures portant sur les invariants des graphes.

- Le programme considère les invariants inv_1, inv_2, ..., inv_k et dispose pour chacun d'un sous-programme qui le calcule. Parmi les invariants possibles, il y a le nombre de nœuds du graphe, le nombre maximal de liens d'un nœud du graphe avec les autres nœuds, etc.
- Le programme dispose d'une base de test formée des graphes G_1, G_2, ..., G_h. Cette base peut être complétée au fur et à mesure de l'utilisation du système pour le rendre plus efficace.
- Le programme dispose d'un système qui énumère les relations algébriques entre invariants. Ce système engendre, par exemple, $inv_1 + inv_2 \leq inv_3$, $inv_5 / inv_4 > (inv_1)^2$, etc. Le système donnera la priorité aux relations simples. La recherche de conjectures portant sur les invariants consiste alors à tester les relations algébriques énumérées par le système. Celui-ci élimine toute relation non vérifiée par l'un des graphes de la base de test.
- Les relations algébriques passant les tests sont soumises à une seconde série de tests pour déterminer si elles sont connues ou résultent de relations connues. Les relations non écartées sont proposées comme conjectures aux mathématiciens, qui tentent de les démontrer.



Exemples de graphes. Le système *Graffiti* s'appuie sur une base de graphes pour tester des relations algébriques entre les « invariants » (nombre de nœuds, nombre moyen d'arêtes entre deux nœuds, etc.).

fournies en exemples. Le système a eu accès à plus de 10 000 démonstrations de théorèmes, exprimées dans le langage de HOL-Light. Parmi elles, de nombreuses démonstrations de résultats intermédiaires de la preuve de la conjecture de Kepler.

Notons que ces démonstrations avaient presque toujours vu le jour à la suite d'un travail guidé par des mathématiciens. Grâce aux paramètres du réseau de neurones du système *HOlist*, ajustés par des milliers d'exemples, celui-ci sait mieux comment choisir et ordonner les éléments d'une démonstration pour atteindre un but. On lui a resoumis les énoncés des théorèmes ayant servi à son apprentissage, et il a pu retrouver, grâce aux tactiques de démonstration apprises, et cette fois seul, sans guidage par des mathématiciens, la démonstration de 58 % d'entre eux. Plus intéressant, quand on lui a soumis 3 217 énoncés pris en dehors des énoncés de la base d'exemples, et qu'on a demandé au système de les démontrer, il a su le faire pour 1 251 d'entre eux. Mais, pour l'instant, le système n'a pas découvert de théorèmes nouveaux.

CALCULER DES PRIMITIVES

Un autre exemple tout récent d'utilisation de l'apprentissage profond concerne le calcul des primitives en analyse et la résolution d'équations différentielles. Certaines méthodes utilisées par les mathématiciens, par exemple l'intégration par parties, ou le changement de variable, sont programmées pour qu'un système réussisse seul des calculs d'intégrales. Beaucoup de travaux ont été ainsi menés conduisant à des programmes de plus en plus complexes et efficaces. En reprenant le problème avec des méthodes d'apprentissage et des réseaux de neurones, Guillaume Lample et François Charton, de Facebook, ont mis au point une méthode différente, fondée sur la présentation d'exemples à un réseau de neurones soigneusement configuré pour apprendre sur ce type de problèmes symboliques.

Les exemples proposés au système étaient de trois types : (1) des paires $[f, F]$, où F est une primitive de f, provenant des programmes de calcul fondés sur l'algorithme de Risch ou une de ses variantes ; (2) des paires provenant de dérivation qui, elle, est facile et maîtrisée depuis longtemps, contrairement au calcul de primitives : on dérive F, ce qui donne f, d'où la paire $[f, F]$; (3) des paires exploitant la méthode d'intégration par parties.

Les performances du système obtenu égalent et parfois dépassent celles des systèmes classiques comme Mathematica et Matlab. Cela montre de nouveau que l'association de méthodes purement symboliques (dans la phase d'apprentissage) et de méthodes neuronales fait progresser la solution de problèmes symboliques.

Ainsi, pour certaines tâches spécialisées, les programmes égalent ou battent les humains, mais ces méthodes ont toujours des champs de réussite étroits. Même si, indubitablement, la recherche mathématique en tire une aide, les systèmes conçus aujourd'hui se bloquent et cessent de progresser.

Il est formidable d'avoir développé des systèmes informatiques qui formulent d'hypothétiques relations entre les paramètres des graphes, qui mènent des calculs arithmétiques et en tirent des formules de séries nouvelles, qui effectuent des calculs de primitives ou proposent quelques concepts inattendus par combinaison. Cependant, tous atteignent vite leurs limites. Au contraire, les mathématiciens bouillonnent d'idées nouvelles que les ordinateurs aident parfois à maîtriser, mais en restant toujours de simples esclaves calculateurs sans véritable créativité. L'ordinateur authentiquement mathématicien n'est pas encore là, on attend sa venue !

— L'auteur —

> Jean-Paul Delahaye est professeur émérite à l'université de Lille et chercheur au laboratoire Cristal (Centre de recherche en informatique, signal et automatique de Lille).

— À lire —

> A. Davies et al., Advancing mathematics by guiding human intuition with AI, *Nature*, 2021.

> G. Lample et F. Charton, Deep learning for symbolic mathematics, prépublication arXiv:1912.01412, 2019.

> C. E. Larson et N. Van Cleemput, Automated conjecturing I: Fajtlowicz's Dalmatian heuristic revisited, *Artificial Intelligence*, 2016.

> D. B. Lenat et P. J. Durlach, Reinforcing math knowledge by immersing students in a simulated learning-by-teaching experience, *International Journal of Artificial Intelligence in Education*, 2014.

Se rapprocher de l'humain

Jusqu'où pourront aller les machines dotées d'intelligence ? Trouver une réponse à cette question n'est pas l'apanage de la science-fiction. Entre chercheurs, la possibilité qu'au-delà de la seule intelligence « artificielle » émerge une conscience qui le soit aussi est chaudement débattue. Les neurones artificiels seront-ils assez performants pour cela ? Il est vrai que, comparés à ceux de notre cerveau, ils sont encore bien trop gourmands en énergie, mais des pistes d'amélioration existent. L'intelligence artificielle du futur sera sans doute aussi plus perméable à nos émotions. Elle saura y répondre, et peut-être même en jouer. Éviter d'être manipulés par nos programmes et nos robots sera un enjeu plus que sérieux.

© Sébastien Thibault/Pour la science

03

Itinéraire bis
pour la voiture autonome

Sur cette vue d'artiste du véhicule autonome de monsieur Tout le Monde, l'illustrateur a conservé le volant. Un élément qui, à terme, devrait disparaître, tout comme le pédalier.

© Metamorworks/shutterstock.com

Il y a quelques années, médias et constructeurs promettaient le déploiement généralisé des systèmes de conduite automatisés pour 2020. Nous y sommes. Et les voitures intelligentes, 100 % autonomes, se font attendre…

Steven Shladover

À la fin des années 2010, l'idée de « voiture autonome » avait le vent en poupe. Les services commerciaux des constructeurs, des équipementiers et des bureaux d'études, par exemple, créaient sans cesse du matériel publicitaire qui surfait sur cette vague. Quant aux journalistes spécialisés, ils avaient une tendance acharnée à l'optimisme technologique, sans doute parce que cela leur permettait de camper un futur étonnant. Aujourd'hui, force est de constater qu'il nous faut déchanter. Pourtant, dès août 2015, dans un article paru dans *Pour la Science* et titré « Les voitures en quête d'autonomie », je relevais les attentes irréalistes sur la « voiture autonome », en pointant notamment l'amalgame terminologique qui frappait les systèmes de conduite automatisée, ou ADS (pour *automated driving system*, en anglais). En effet, tant les médias que les constructeurs automobiles et les entreprises de haute technologie employaient sans distinction les termes « autonome », « sans conducteur », « conduite automatique », alors qu'il s'agit de solutions techniques différentes, bien distinguées par l'Association internationale des ingénieurs automobiles *(voir l'encadré page 89)*. En définitive, l'obsolescence programmée de la conduite humaine n'a pas eu lieu, et il nous faut désormais en rabattre quant à nos ambitions.

Seule une poignée de prototypes avancés ont circulé sur des routes publiques en 2020 sans que leurs systèmes automatiques aient eu besoin de faire appel au pilote humain placé, par sécurité, derrière le volant. Le terme « autonome » *(self-driving)* a perdu son sens original, car l'assistance à la conduite des voitures étiquetées « autonomes » ne peut en réalité pas fonctionner sans une constante supervision humaine. Il est aujourd'hui bien plus pertinent d'automatiser des camions, bus et navettes de covoiturage que des voitures particulières. En d'autres termes, des véhicules autonomes sont encore d'actualité, mais dans des situations bien particulières.

UN OPTIMISME TEMPÉRÉ

Dès 2018, les PDG des entreprises majeures qui avaient investi massivement dans les ADS (Waymo, General Motors, Ford, Aurora) ont commencé à tempérer publiquement leur optimisme des années précédentes en soulignant que le déploiement de la conduite automatisée se ferait pas à pas, dans des conditions limitées et des endroits spécifiques. Au rythme auquel ils cheminent à présent, des décennies seront nécessaires pour assister à une expansion proche d'un déploiement national. Après au moins une décennie et des milliards de dollars d'investissements dans le développement de systèmes de conduite automatisée, les entreprises se sont rendu compte que les prérequis techniques nécessaires pour soutenir un usage généralisé de la technologie sont beaucoup plus compliqués que prévu. Dans le même temps, des entreprises comme Tesla et quelques start-up moins mûres conduisent ces déploiements plus vite et à plus grande échelle, sans réaliser à quel point elles sont encore loin du compte.

Un consensus est en train de se former au sein des développeurs d'ADS instruits par l'expérience, que l'on peut résumer ainsi : dans les années à venir, des opérations automatisées ne seront possibles que dans des contextes

étroitement définis, caractérisés par une météo, un éclairage et une circulation ordinaires, des lieux géolocalisés et cartographiés avec grande précision et, dans de nombreux cas, équipés de fonctionnalités de support d'infrastructure physique et numérique *ad hoc*.

Ensuite, comme il faut éviter tous les dangers de la circulation sans l'intervention d'un humain, les systèmes doivent garantir un niveau de sécurité beaucoup plus élevé que les ADS sous supervision humaine. Ces dispositifs s'appuieront sur de multiples prises d'information dans l'environnement du conducteur et ses dangers. Ces données proviennent de caméras, de lidars (télédétection par laser), de radars et d'un géopositionnement précis, combinés à des cartes numériques très détaillées.

VERS DES DRONES ROULANTS

Par ailleurs, bien que de nombreux développeurs d'ADS revendiquent la capacité à circuler sans dépendre des communications sans fil avec d'autres véhicules – ou d'alertes provenant des usagers de la route ou de l'infrastructure routière elle-même –, des travaux récents ont montré que la généralisation d'ADS dépourvus d'une telle communication coopérative aurait sans doute des effets pervers sur le trafic, la consommation d'énergie et les émissions dans l'environnement, car elle empêcherait d'anticiper les changements dans les conditions de circulation.

Enfin, la technologie va d'abord être mise en application dans certains usages spécialisés, comme la livraison locale de colis, le fret longue distance sur autoroute, les services de transport urbain avec itinéraires fixes et, à échelle plus réduite, le transport automatisé de passagers en ville et en banlieue. Même si les ADS sont capables de conduire des véhicules sans l'assistance d'un conducteur à bord, ils auront quand même besoin du soutien à distance d'opérateurs qualifiés en conduite pour gérer les «guêpiers» de la circulation dont les automates ne sauront s'extraire.

Ces conclusions découlent des leçons déjà tirées par les développeurs de la technologie. Première d'entre elles : il y a une complexité inhérente à la conduite d'un véhicule, qui rend difficiles, pour les systèmes automatisés, la perception de l'environnement du conducteur, l'anticipation des actions des autres usagers, l'identification et la gestion des dangers de la circulation.

Un peu comme des enfants qui apprennent à contrôler leurs mouvements, les ADS doivent pouvoir se lever avant de marcher, et marcher avant de courir. Pour cette raison, leur mise en œuvre et leur amélioration se feront dans des environnements simples. Les confrontations complexes avec des usagers de la route imprévisibles (y compris les piétons et les cyclistes) ou le mauvais temps viendront après.

Afin d'opérer en toute sécurité, les ADS devront percevoir leur proche environnement et déterminer leur position à l'aide d'une technologie fondée sur des principes de la physique. Savoir gérer l'adversité leur sera indispensable, qu'elle prenne la forme d'interférences électromagnétiques dues aux orages ou à des gros équipements électriques voisins, du soleil bas sur l'horizon capable d'aveugler les caméras, des précipitations ou de la fumée qui diffusent la lumière nécessaire aux capteurs optiques ou de

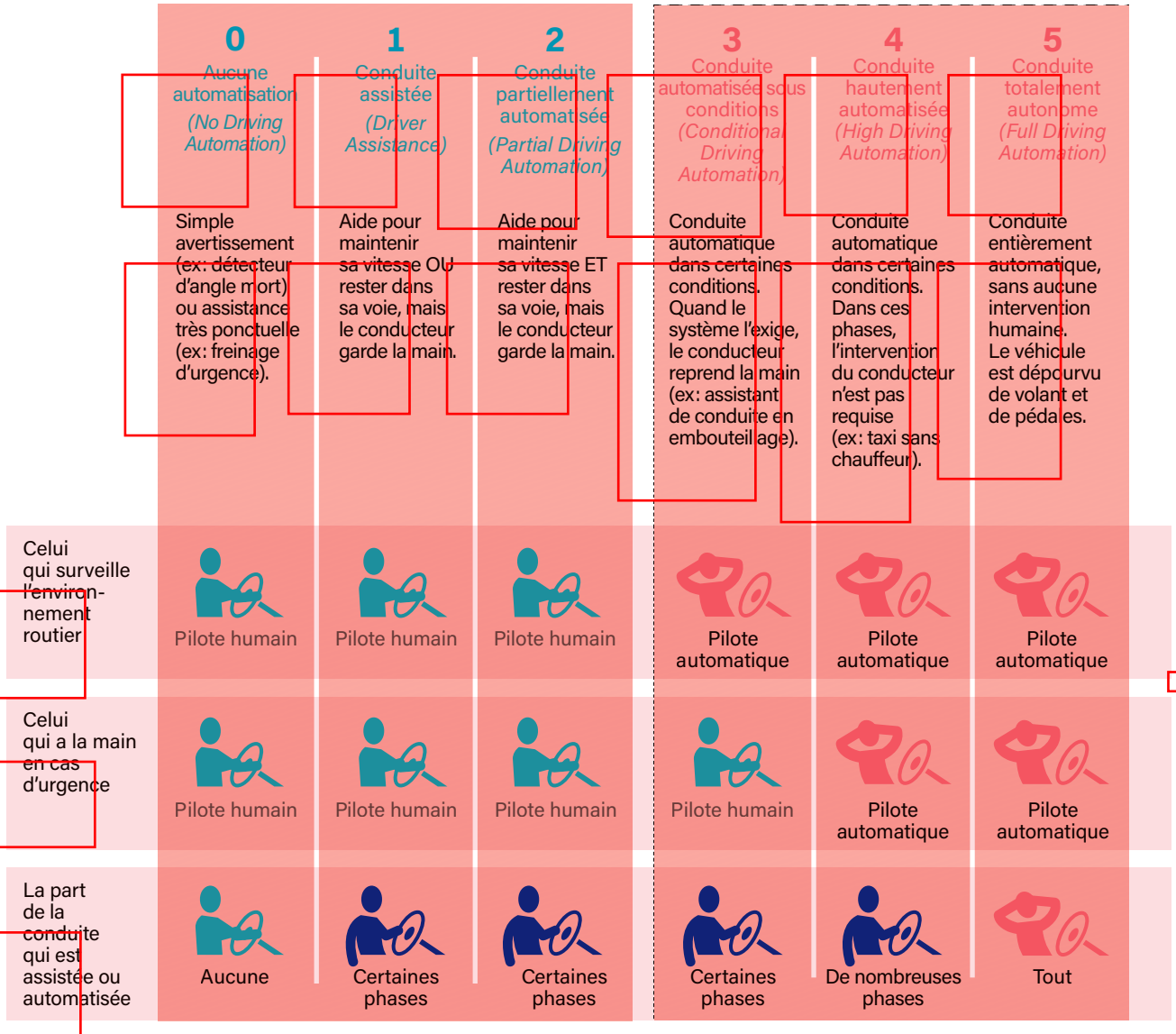
En bref

> Les promesses des constructeurs n'ont pas été tenues : le véhicule capable de se conduire seul n'existe pas encore.

> Les difficultés techniques ont été sous-estimées. Percevoir et gérer l'environnement du conducteur en toutes circonstances (trafic, météo) est encore hors de portée.

> Faisant preuve désormais de réalisme, les développeurs se réorientent vers les flottes commerciales et le transport de marchandises, en ville et sur autoroute.

> L'option d'une assistance humaine à distance des véhicules automatisés a été négligée, alors qu'elle est sans doute une activité porteuse.



↑ Le niveau d'automatisation de la conduite d'un véhicule est défini par SAE International, l'Association internationale des ingénieurs automobiles. Six niveaux existent (de 0 à 5) et sont revisés régulièrement. Aucun véhicule au monde ne peut encore prétendre au niveau 5, celui de l'autonomie réelle.

cyberattaques ciblant les capteurs du véhicule. Acquérir des informations sur les infrastructures à proximité et le mouvement relatif des objets alentour sera tout aussi nécessaire. Tous ces capteurs récoltent des données qui, une fois fusionnées, construisent une représentation de l'espace autour du véhicule.

En conséquence, les critères qu'exigent les capteurs, le traitement et le stockage des données pour ce niveau de perception de l'environnement sont tels qu'ils feront gonfler les coûts

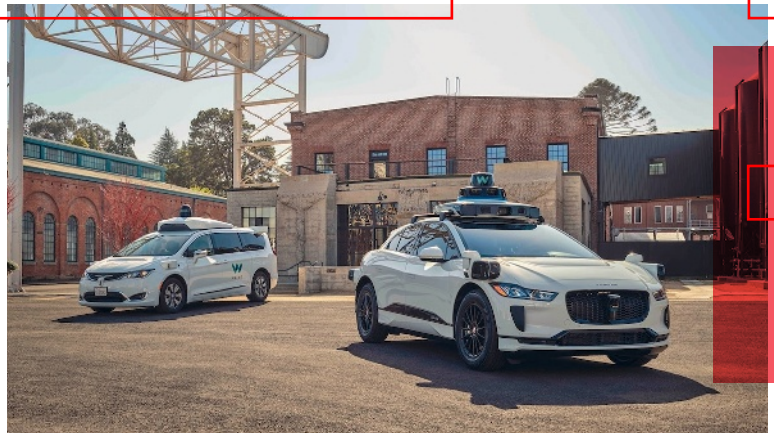
D'après SAE International (https://www.sae.org/standards/content/j3016_202104/) // Illustrations de Nigel Holmes

de la technologie ADS. Des dépenses néanmoins nécessaires, si cette technologie vise à égaler les compétences du conducteur, sans pour autant compromettre sa sécurité. Les analyses de rentabilisation du déploiement des ADS se réorientent donc vers une autre destination : celle des véhicules commerciaux. En effet, ceux-ci sont utilisés toute la journée pour engendrer du revenu, là où les véhicules particuliers ne servent qu'une à deux heures par jour. Un autre avantage, à l'heure où la technologie d'automatisation progresse encore, est qu'à l'échelle d'une flotte l'entretien des composants sensibles est facilité, tout comme la mise en place d'un système d'assistance à distance.

Cette question, essentielle, de l'assistance humaine à distance pour la conduite autonome n'a pas reçu le niveau d'attention public qu'elle mérite. Les ADS dépourvus de conducteur à bord doivent pouvoir s'appuyer sur des opérateurs humains, installés dans un centre de gestion de flotte, qui aideraient le système à prendre des décisions tactiques en cas de besoin. L'assistant humain pourra compenser les carences matérielles et logicielles des véhicules, mais il fournira surtout des conseils. Par exemple, quand il faudra contourner des obstacles inattendus qui obstruent la voie, être contraint de violer le Code de la route, reconnaître les gestes des agents de la circulation ou identifier des objets sur la trajectoire du véhicule.

CONDUITE ACCOMPAGNÉE… À DISTANCE

Certains développeurs envisagent aussi la conduite à distance comme un moyen de gérer des portions de route trop difficiles pour l'ADS.



← Depuis 2009, la société Waymo, filiale d'Alphabet, la maison mère de Google, travaille sur des véhicules autonomes. À ce jour, c'est la seule entreprise qui propose un service commercial de course à la demande, Waymo One, sans présence d'un chauffeur humain.

© Waymo

> La question, essentielle, de l'assistance humaine à distance n'a pas reçu le niveau d'attention qu'elle mérite

Ces fonctions support à distance seront particulièrement importantes au début, lorsque la technologie sera en rodage. Elles créeront de l'emploi pour les nombreux chauffeurs routiers peu à peu remplacés par les ADS. Leurs conditions de travail pourraient en être améliorées : ils travailleront plus près de leur domicile, subiront moins la fatigue et l'usure physique, et seront bien sûr moins exposés aux risques d'accidents associés au transport routier.

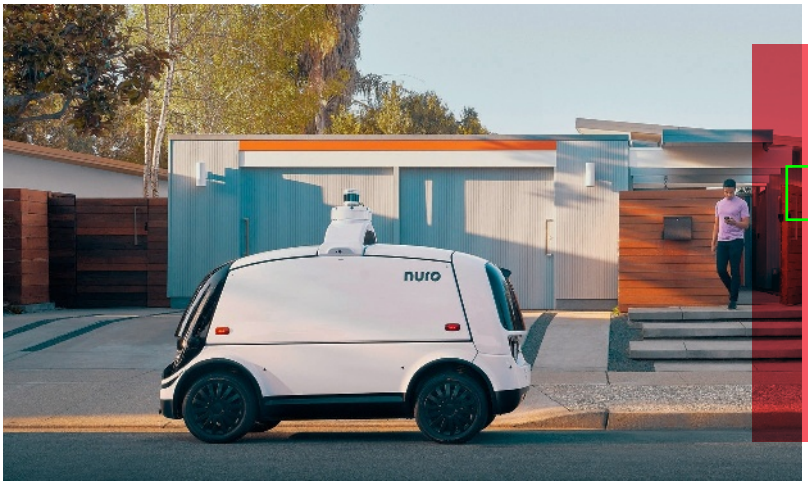
Il y a quelques années, les projections estimaient que la conduite autonome servirait surtout les services de course à la demande, mais c'est désormais le transport automatisé de biens qui concentre l'attention : on envisage à la fois des poids (très) lourds sur de longues distances et, sur le dernier kilomètre, des petits véhicules peu rapides pour la livraison locale de colis en zone urbaine et périurbaine. Ce changement d'orientation a commencé avant la pandémie, et s'est accéléré pendant la crise du Covid-19. Certes, les services de courses à la demande, incluant voyages courts et covoiturage dans des zones urbaines denses, ont le profil de demande le plus favorable. Mais ce sont aussi les environnements les plus éprouvants pour la technologie des ADS.

A contrario, les transports longue distance sur autoroute sont soumis à des trafics plus simples à gérer, avec des infrastructures routières également plus faciles à contrôler, en particulier si les points de départ et d'arrivée sont directement connectés aux autoroutes, avec des voies réservées. La livraison de colis peut être assurée par des véhicules trop petits pour héberger un humain, mais capables d'opérer sur des trottoirs ou sur des pistes cyclables si leur vitesse est réduite. De surcroît, les véhicules de fret sans conducteur ont l'avantage de pouvoir être conçus pour une conduite plus sûre, susceptible d'éviter les conflits avec d'autres usagers de la route, et sans avoir à se soucier de l'impatience ou de la sécurité de leurs propres passagers..

DES CONSTRUCTEURS OBSTINÉS

Toutes les entreprises d'ADS n'ont pas adopté cette approche. Tesla, en particulier, continue de suivre le modèle traditionnel de l'industrie automobile, lequel consiste à vendre des véhicules personnels privés répondant à tous les usages. L'entreprise d'Elon Musk tente de faire progresser ses capacités en matière d'ADS en les greffant sur son système d'assistance à la conduite actuel, plutôt que de concevoir un nouveau système, qui répondrait aux exigences élevées en matière de redondance et de gestion des pannes. Récemment, la société a fait le choix de dépendre exclusivement d'un système de vision artificielle, faisant l'impasse sur les capteurs complémentaires (cartographie, localisation de précision), qui fournissent pourtant des informations vitales sur l'environnement d'un véhicule.

Cette approche réduit certes les coûts de production, mais au détriment de la sécurité et de la fiabilité, qu'il faudra garantir malgré les variations inévitables de la circulation et de l'environnement. En utilisant à tort et à travers les termes « autopilote » et « conduite entièrement autonome » (*full self-driving*, en anglais) pour ses produits actuels, Tesla a d'ailleurs semé la confusion dans l'esprit du public sur les capacités de ses produits. Pour ne rien arranger, dans



Après avoir été testé avec succès, notamment par la chaîne Domino's Pizza, le robot de livraison R2, de la société Nuro, est sur le point d'être produit à grande échelle. De petite taille, électrique, il livre des marchandises à domicile de façon autonome.

ses manuels d'utilisation et ses correspondances adressées aux autorités de régulation, l'entreprise décrit sa technologie comme un système d'assistance à la conduite plutôt qu'un système de conduite autonome. La confusion est telle que Waymo, connue à l'origine comme le «projet de voiture autonome» de Google, a annoncé qu'elle n'utiliserait plus le terme «conduite autonome».

Avant un déploiement large, il faudra donc relever ce défi technologique majeur : assurer la sécurité des opérations d'ADS, en faisant face à l'éventail complet des dangers de la circulation et des défaillances matérielles et logicielles internes. Chercheurs et développeurs d'ADS expérimentent diverses approches, mais ils sont très loin de s'entendre sur un choix qui fasse consensus. Les défis vont bien au-delà des considérations technologiques ; ils touchent à une question sociétale plus large : comment déterminer « quel degré de sûreté est assez sûr» pour que les ADS puissent être intégrés dans le service au public. Des mesures de sécurité *ad hoc* doivent être arrêtées afin que le régulateur et le public soient à l'aise avec un déploiement généralisé. Les développeurs d'ADS doivent aussi apprendre à formuler leurs évaluations de sécurité, lesquelles sont hautement techniques, en des termes qui seront compris avec clarté et précision, afin de gagner la confiance et l'approbation du régulateur comme des usagers de la route.

Si certains observateurs estiment que le vent de l'enthousiasme est retombé en ce qui concerne la conduite autonome, je vois au contraire dans la situation actuelle une marque de progrès. Des vues plus réalistes sur les possibilités et les défis de la conduite automatisée motiveront l'investissement de ressources mieux ciblées, et permettront d'aligner les perceptions du public avec la réalité. Pour les dix ans à venir, il faut s'attendre à une mise en pratique limitée de la conduite autonome, dans le transport routier longue distance sur autoroute et en zone urbaine et périurbaine pour les livraisons locales. Les services automatisés de course à la demande et de covoiturage urbain pourraient également être proposés de manière limitée. Quant au déploiement à un niveau national, cessons de rêvasser : il n'est pas pour demain.

— L'auteur —

> Steven Shladover
Ingénieur de formation, il mène des recherches sur les technologies avancées pour les systèmes de transport depuis plus de quarante-cinq ans.

— À lire —

> D. McAslan, Pilot project purgatory? Assessing automated vehicle pilot projects in U.S. cities, *Humanities and Social Sciences Communications*, 2021.

> J. Claybrook et S. Kildare, Autonomous vehicles: No driver... no regulation?, *Science*, 2018.

> S. Shladover, Les voitures en quête d'autonomie, *Pour la Science*, n° 466, août 2016.

© Nuro

ABONNEZ-VOUS À

Pour la Science

OFFRE D'ABONNEMENT

DURÉE LIBRE

2 FORMULES AU CHOIX

	FORMULE PAPIER + HORS-SÉRIE	FORMULE INTÉGRALE
Le magazine papier 12 numéros par an	✓	✓
Le magazine en version numérique 12 numéros par an		✓
Le hors-série papier 4 numéros par an	✓	✓
Le hors-série en version numérique 4 numéros par an		✓
Accès à pourlascience.fr actus, dossiers, archives depuis 1996		✓
VOTRE TARIF D'ABONNEMENT	**6,50 €** PAR MOIS	**8,20 €** PAR MOIS
	31 % de réduction *	**43 %** de réduction *

BULLETIN D'ABONNEMENT

À renvoyer accompagné de votre règlement à :

Service Abonnement Groupe Pour la Science – 235 avenue Le Jour se Lève – 92100 Boulogne Billancourt – email : serviceclients@groupepourlascience.fr

PAG22STDDOS

☑ **OUI, je m'abonne à Pour la Science en prélèvement automatique**

1 / Je choisis ma formule (merci de cocher) et je complète l'autorisation de prélèvement ci-dessous.

☐ FORMULE **PAPIER + HORS SÉRIE**
- 12 n° du magazine papier
- 4 Hors-série papier

6,50€ PAR MOIS -31 %

1-F-HSPAP-N-3PVT-6,5€

☐ FORMULE **INTÉGRALE**
- 12 n° du magazine (papier et numérique)
- 4 Hors-série (papier et numérique)
- Accès illimité aux contenus en ligne

8,20€ PAR MOIS -43 %

1-F-INT-N-3PVT-8,2€

2 / Mes coordonnées

Nom : .....

Prénom : .....

Adresse : .....

.....

Code postal ☐☐☐☐☐ Ville : .....

Tél. : ☐☐☐☐☐☐☐☐☐☐

Courriel : ..... @ .....

(indispensable pour la formule intégrale)

J'accepte de recevoir les offres de Pour la Science ☐ OUI ☐ NON

3 / Mandat de prélèvement SEPA

En signant ce mandat SEPA, j'autorise Pour la Science à transmettre des instructions à ma banque pour le prélèvement de mon abonnement dès réception de mon bulletin. Je bénéficie d'un droit de rétractation dans la limite de 8 semaines suivant le premier prélèvement. Plus d'informations auprès de mon établissement bancaire.

TYPE DE PAIEMENT : PAIEMENT RÉCURRENT

Titulaire du compte

Nom : ..... Prénom : .....

Adresse : .....

.....

Code postal ☐☐☐☐☐ Ville : .....

Désignation du compte à débiter

BIC (Identification internationale de la banque) ☐☐☐☐☐☐☐☐☐☐☐

IBAN ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

(Numéro d'identification international du compte bancaire)

Établissement teneur du compte

Nom : .....

Adresse : .....

Code postal ☐☐☐☐☐ Ville : .....

Date et signature

Organisme Créancier : Pour la Science
170 bis, bd. du Montparnasse – 75014 Paris
N° ICS FR92ZZZ426900
N° de référence unique de mandat (RUM)

4 / MERCI DE JOINDRE IMPÉRATIVEMENT UN RIB

Partie réservée au service abonnement. Ne rien inscrire

* Réduction par rapport au prix de vente en kiosque et l'accès aux archives numériques. Durée d'abonnement : 1 an. Délai de livraison : dans le mois suivant l'enregistrement de votre règlement. Offre valable jusqu'au 31/03/2023 en France métropolitaine uniquement. Pour un abonnement à l'étranger, merci de consulter notre site boutique.groupepourlascience.fr. Photos non contractuelles. Vous pouvez acheter séparément les numéros de *Pour la Science* pour 6,90 € et les hors-séries pour 7,90 €. En souscrivant à cette offre, vous acceptez nos conditions générales de vente disponibles à l'adresse suivante : https://rebrand.ly/CGV-PLS.

Les informations que nous collectons dans ce bulletin d'abonnement nous aident à personnaliser et à améliorer les services que nous vous proposons. Nous les utiliserons pour gérer votre accès à l'intégralité de nos services, traiter vos commandes et paiements, et vous faire part notamment par newsletters de nos offres commerciales moyennant le respect de vos choix en la matière. Le responsable du traitement est la société *Pour la Science*. Vos données personnelles ne seront pas conservées au-delà de la durée nécessaire à la finalité de leur traitement. *Pour la Science* ne commercialise ni ne loue vos données à caractère personnel à des tiers. Les données collectées sont exclusivement destinées à *Pour la Science*. Nous vous invitons à prendre connaissance de notre charte de protection des données personnelles à l'adresse suivante : https://rebrand.ly/charte-donnees-pls. Conformément à la réglementation applicable (et notamment au Règlement 2016/679/UE dit «RGPD») vous disposez des droits d'accès, de rectification, d'opposition, d'effacement, à la portabilité et à la limitation de vos données personnelles. Pour exercer ces droits (ou nous poser toute question concernant le traitement de vos données personnelles), vous pouvez nous contacter par courriel à l'adresse protection-donnees@pourlascience.fr.

Groupe Pour la Science - Siège social : 170 bis, boulevard du Montparnasse, CS20012, 75680 Paris cedex 14 – Sarl au capital de 32 000€ – RCS Paris B 311 797 393 – Siret : 311 797 393 000 23 – APE 58.14 Z

Dialoguer, détecter l'état émotionnel des humains, pratiquer l'humour… certains programmes le font déjà. Mais il sera indispensable de bien évaluer ces capacités si l'on veut utiliser robots et « chatbots » en toute confiance.

Bien vivre avec les machines

Laurence Devillers

© Westend61/GettyImages

En bref

> Nous interagissons de plus en plus avec des machines, ou des robots, qui nous parlent.

> Ces « agents conversationnels » simuleront des capacités qui auront l'air d'être élaborées sur le plan cognitif, mais aussi sur le plan émotionnel, alors qu'ils ne comprennent rien.

> Il est essentiel d'évaluer les capacités de tels systèmes, voire d'en réaliser un suivi : étant donné la faculté d'apprentissage de ces machines, leur comportement pourrait changer avec le temps.

> Il faut aussi étudier le comportement des humains vis-à-vis des « robots bavards », afin d'éviter certains abus et effets pervers, l'isolement social par exemple.

« Si vous souhaitez joindre le service commercial, tapez 1. Le service technique, tapez 2. » Cela vous dit quelque chose ? Eh bien, l'époque où une demande de renseignements exigeait de taper sur les touches d'un ordinateur ou d'un téléphone avant d'obtenir une information ou de joindre le bon interlocuteur touche à sa fin. De plus en plus, nous utilisons des programmes informatiques auxquels on s'adresse en parlant naturellement, et qui nous renseignent aussi en langage naturel, que ce soit à l'écrit ou à l'oral. Connectés sur internet ou à une base de données, embarqués sur un ordinateur, un téléphone ou encore un robot, ces programmes, invisibles aux yeux des utilisateurs, sont ce qu'on nomme des « agents conversationnels », ou encore robots bavards (*chatbots*, en anglais). Le premier d'entre eux était Eliza, élaboré à l'institut de technologie du Massachusetts (MIT) vers 1964-1966 par Joseph Weizenbaum : ce système répondait à l'écrit en simulant un psychothérapeute dont la stratégie était de répéter les propos du patient.

Si les *chatbots* ont longtemps été développés de manière à répondre à un sujet de discussion précis ou une tâche particulière, les *chatbots* dits à domaine « ouvert » sont conçus pour parler de tout. Ce qui sous-entend certaines capacités élaborées de la machine, ou plus exactement des logiciels dont elle est munie. Les *chatbots* étaient jusqu'à présent composés de plusieurs modules : de reconnaissance de la parole, de compréhension, de dialogue, de production et de synthèse de la parole. Les nouvelles générations de *chatbots* sont de plus en plus performantes grâce à l'évolution des techniques d'apprentissage des machines, à la puissance des processeurs et à la taille des bases de données. Les modèles de traitement automatique de la langue les plus récents, appelés des « transformeurs », sont des réseaux de neurones qui, à partir de vastes corpus linguistiques, apprennent les régularités les plus saillantes, sans être influencés, par l'ordre des mots. Des *chatbots* sont maintenant capables d'aborder n'importe quel sujet sans contrainte, par exemple Blenderbot, le modèle conversationnel de Facebook. Ou encore Meena, système fondé sur l'apprentissage profond de bout en bout entraîné sur des dialogues entre humains, qui montre des performances dépassant tous les *chatbots* existants.

LE TEST DE TURING, INSUFFISANT POUR PLUSIEURS RAISONS

Au-delà des avancées scientifiques et techniques mises en œuvre dans les *chatbots*, l'interaction croissante que nous avons avec ces derniers soulève des questions plus fondamentales, voire stratégiques et éthiques, sur les capacités des machines et sur la relation intersubjective qu'instaurent avec celles-ci leurs utilisateurs humains. En particulier, si l'on veut vivre harmonieusement avec des machines qui nous assistent et avec lesquelles nous dialoguons, il apparaît de plus en plus important d'évaluer de façon pertinente et continue, par des batteries de tests reproductibles, ces capacités ainsi que leur impact sur la relation entre humain et machine.

Que faut-il tester plus précisément, et comment ? On s'est longtemps focalisé sur l'« intelligence » des machines. D'ailleurs, le domaine

dit « de l'intelligence artificielle » a pour objectif, selon son pionnier John McCarthy, de « doter des machines de systèmes informatiques ayant des capacités intellectuelles comparables à celles des hommes ».

En 1950, pour établir si une machine est intelligente ou non, le mathématicien britannique Alan Turing proposait le « jeu de l'imitation ». Il préfigurait l'orientation des futures technologies de l'information et de la communication. Le « test de Turing », comme on le nomme aujourd'hui, consiste à confronter par le biais d'une conversation textuelle deux interlocuteurs – un humain et une machine – à un juge, lequel doit décider qui est qui. Si la machine réussit à faire croire à l'examinateur qu'il a affaire à un interlocuteur humain, c'est qu'elle a des capacités qui ressemblent à ce que nous appelons de l'« intelligence ».

Le test de Turing est-il la meilleure façon d'évaluer une machine conversationnelle *(voir Interro écrite pour l'intelligence artificielle, par M. Mitchell, page 54)* ? Le jeu de l'imitation proposé par Turing apparaît de plus en plus insuffisant, et ce pour plusieurs raisons. L'une d'elles est qu'il ne détermine pas directement les capacités, cognitives ou autres, du programme : il évalue seulement si ce programme peut simuler un comportement humain et tromper l'examinateur. Un autre argument remettant en cause la puissance du test de Turing a été avancé en 1980 par John Searle. Selon ce philosophe américain, un *chatbot* ne fait que manipuler la syntaxe, et réussir ce jeu d'imitation ne prouverait donc rien. Les *chatbots* imitent le langage des humains, mais ils ne comprennent pas le sens de l'information qu'ils traitent.

Le test de Turing a ses limites : il évalue seulement si le programme peut simuler un comportement humain et tromper l'examinateur

Le test de Turing est insuffisant pour une autre raison : il ne prend pas en compte les diverses et nombreuses facettes de l'intelligence humaine. On la réduit souvent à la capacité de mener un raisonnement abstrait, mais l'intelligence émotionnelle ou sociale est un autre aspect primordial de l'humain, qui requiert d'autres compétences. L'intelligence est un concept polysémique qui regroupe un vaste ensemble de fonctions mentales : apprentissage, compréhension et organisation du réel en concepts, interprétation des règles sociales et culturelles, capacité à utiliser le raisonnement causal, imagination, prospection et flexibilité, mais aussi habileté à percevoir et à exprimer les émotions, à les intégrer pour comprendre,

DES SYSTÈMES PERFORMANTS QUI GAGNENT

Il existe des programmes informatiques qui effectuent des tâches spécialisées avec une efficacité impressionnante. C'est le cas du programme AlphaGo, qui, en 2016, a battu au jeu de go le champion Lee Sedol *(à droite sur la photographie ci-contre)*.
Et début 2017, le programme Libratus a battu quatre joueurs professionnels de poker Texas Hold'em *(page ci-contre, Jason Les jouant contre Libratus)*, un jeu où l'on fait face à des informations incomplètes et à des ruses. Ici, la capacité de calcul ne suffit plus…

dialoguer et raisonner, ainsi qu'à réguler les émotions chez soi et chez les autres.

Pour autant, cela n'a pas empêché le domaine de l'intelligence artificielle d'avoir beaucoup progressé ces dernières années et d'avoir acquis une importance stratégique. Ces avancées sont liées à deux ruptures technologiques : d'une part, la puissance de calcul des ordinateurs permet aujourd'hui d'analyser d'énormes masses de données en quelques secondes ; d'autre part, avec les techniques d'apprentissage profond (*deep learning*, en anglais), des réseaux de neurones artificiels apprennent à effectuer certaines tâches selon des mécanismes plus efficaces que ceux utilisés dans les années 1990.

Ainsi, pour une tâche qui exige uniquement des calculs aujourd'hui bien compris, comme jouer aux échecs ou au go, les programmes informatiques affichent des performances très impressionnantes et peuvent battre des champions du monde : le superordinateur Deep Blue, d'IBM, l'a fait en 1997 pour les échecs, et le programme AlphaGo, de l'entreprise DeepMind-Google, en 2016 pour le go. Le système Libratus, conçu à l'université Carnegie-Mellon, a battu quatre joueurs professionnels au poker dans sa version Texas Hold'em. Il faut comprendre que si le poker est effectivement un jeu plus complexe que les échecs et le jeu de go, cette complexité ne doit pas être exagérée, le nombre de cartes étant limité, et les stratégies aussi. Ce qui est très différent lorsqu'il s'agit du langage, car les possibilités sont infinies.

La possibilité de parler naturellement à un *chatbot* afin qu'il exécute des tâches à notre service ne tient pas du fantasme, même si la technologie actuelle est encore loin de ce que montre la science-fiction. Mais les agents conversationnels ne comprennent pas l'information qu'ils traitent, ce qui est la majeure différence avec la notion d'intelligence artificielle dite «forte» (qui n'est pas pour demain), laquelle vise à reproduire une véritable compréhension de l'information.

Donner aux robots conversationnels des connaissances sémantiques, une compréhension du sens commun et des règles sociales est un défi scientifique complexe, auquel les géants du numérique (Google, Apple, Facebook, Amazon, Microsoft, IBM) se sont attelés. Les *chatbots* intégrés aux smartphones ou aux robots sont désormais de plus en plus courants, avec 4,2 milliards d'assistants vocaux utilisés en 2020 dans le monde. Ce chiffre devrait passer à 8,4 milliards d'ici à 2024.

DOTER LE PROGRAMME D'UNE FORME DE PERSONNALITÉ

La robotique conversationnelle se développe également ; par exemple le robot Buddy, de la société française Blue Frog Robotics, vise le marché d'assistance aux personnes âgées. Par ailleurs, les concepteurs de *chatbots* ajoutent souvent une «personnalité» à la machine. Celle-ci a alors des préférences ou des habitudes qu'elle exprime dans la conversation, ce qui lui donne un semblant d'identité. Il existe d'ailleurs des bibliothèques, comme Koko, issue des travaux du MIT, qui permettent de choisir, selon le contexte du dialogue, des réponses empathiques.

Dans tous ces développements autour des agents conversationnels, les fonctionnalités

© Geordie Wood

émotionnelles – empathie, humour, etc. – prennent de plus en plus d'importance pour que l'on puisse interagir de façon naturelle avec une machine. D'ailleurs, l'informatique émotionnelle fait partie des axes de recherche de mon équipe depuis 2004 au LISN-CNRS, où l'on explore trois types de problématiques et de technologies : la détection des émotions, le raisonnement exploitant des informations de nature affective, mais aussi l'impact de la génération d'expressions émotionnelles. En particulier, l'humour, à un degré assez simple, fait partie des capacités que l'on souhaite donner aux machines. On peut le simuler à travers divers types d'expressions – verbales ou non verbales – socialement et culturellement adaptées, et l'utiliser pour rendre les machines plus attachantes et plus amusantes à utiliser, et ainsi améliorer nos interactions avec elles.

Plus récemment, dans ma chaire *Human-Machine affective interaction & ethics* (Humaaine), nous mettons en œuvre et testons des stratégies de *nudges* (« coups de pouce ») dans des systèmes conversationnels affectifs. Il s'agit d'une technique qui consiste à inciter les individus à changer de comportement sans les contraindre et en utilisant leurs biais cognitifs *(voir Communication, les deux visages de l'intelligence artificielle, par L. Devillers, page 42)*. Nous cherchons à en montrer les bénéfices et les risques dans le dialogue oral. Ces systèmes conversationnels seront de plus en plus capables de discuter avec une personne

> Donner de l'humour aux machines, c'est les rendre plus attachantes, plus amusantes, et ainsi améliorer nos interactions avec elles

© Nate Swallwood

Plus qu'un simple jouet, Moxie est un robot compagnon destiné aux 5-10 ans, capable, selon son concepteur Embodied, de converser en langage naturel pour instruire les enfants et, plus encore, les inciter à développer leurs relations sociales dans la vie de tous les jours. Connectée à une plateforme de mise à jour, l'IA de Moxie évolue à mesure que l'enfant grandit.

en lui donnant l'impression de converser avec un humain.

Comment évaluer les performances d'un agent conversationnel en matière d'imitation, de conversation, de mémorisation, de raisonnement, de simulation affective, d'anticipation ? L'une des difficultés est que ces capacités sont conférées par des algorithmes et des techniques d'apprentissage profond, qui gardent une forme d'opacité : ces techniques fonctionnent selon un principe transparent et bien compris, mais elles produisent des décisions dont il est impossible de retracer les raisons et le cheminement. Hormis ce problème dit « de boîte noire », nous sommes en droit de nous poser bien d'autres questions sur les *chatbots*, en particulier leurs enjeux éthiques. Sur ce sujet, un avis que j'ai corédigé à la CNPEN (Commission nationale pilote d'éthique du numérique) a été rendu au Premier ministre en novembre 2021.

À TROP FRÉQUENTER LES HUMAINS, LE RISQUE EST DE DEVENIR AGRESSIF

Par exemple, la machine en question respecte-t-elle la vie privée de ses utilisateurs et des autres humains ? Se comporte-t-elle de façon éthiquement responsable, et, si oui, est-ce que ce sera toujours le cas jusqu'à la fin de sa vie ? Peut-on créer des jumeaux numériques de personnes décédées et quelles seraient nos relations avec elles ? Est-ce suffisant de tester la machine avant sa mise sur le marché si elle est amenée à s'adapter en continu ? De telles questions sont particulièrement importantes dans le cas de systèmes capables d'apprendre tout au long de leur durée d'utilisation. Il est alors difficile de contrôler leur comportement futur et de garantir que ce dernier restera acceptable. Un *chatbot* au contact d'humains pourrait développer des attitudes agressives ou inconvenantes, un algorithme conçu pour faire des offres financières pourrait à terme orienter ses propositions en fonction du groupe social auquel appartient son interlocuteur, etc.

Un exemple d'une telle dérive a été fourni en 2016 par Tay, un système conversationnel développé par Microsoft. Ce *chatbot*, déployé pour converser avec des jeunes sur la plateforme Twitter, était doté d'une capacité d'apprentissage en continu. On a pu constater qu'il pouvait apprendre de ses interactions avec les internautes, puisqu'il adaptait son langage aux habitudes et éléments qu'il repérait dans les conversations. Mais l'essai a tourné court : au bout de vingt-quatre heures seulement, Tay s'est mis à poster sur la plateforme des tweets à caractère raciste et nazi !

Comment faire évoluer le test de Turing pour évaluer les capacités des agents conversationnels ? Rappelons-le, ce test éprouve la faculté d'une machine à imiter la conversation humaine. La mission du juge est de deviner,

© Embodied Inc.

d'après les messages échangés, qui est, de l'un ou de l'autre de ces interlocuteurs, une machine. Le contenu des messages est libre : il est possible de parler de tous types de sujets. Alan Turing a prédit que, en l'an 2000, des ordinateurs avec 128 Mo de mémoire seraient capables de tromper environ 30 % des juges humains durant une mise à l'épreuve de cinq minutes. On en est loin.

Des biais ne sont pas exempts du test de Turing. Jean-Paul Delahaye, chercheur en informatique à Lille (et auteur des articles des pages 60 et 76), explique par exemple qu'une simple question de calcul est susceptible de démasquer l'interlocuteur-programme : l'humain sera laborieux, alors que la machine excellera. Mais, pourrait-on rétorquer, rien n'empêche de programmer la machine pour qu'elle fasse exprès de se tromper…

Toujours est-il que faire passer le test de Turing reste un graal pour tous les concepteurs d'agents conversationnels. Depuis 1990, il existe même une compétition annuelle, le prix Hugh Loebner, qui couronne les systèmes satisfaisant le mieux les critères du test de Turing, c'est-à-dire avec lesquels il est le plus difficile de déterminer s'il s'agit d'un humain ou d'une machine. Les meilleurs programmes n'effectuent qu'un simulacre de conversation qui, malgré de récents progrès (systèmes capables de répondre dans des domaines ouverts, adaptables à différents interlocuteurs, dotés de certains traits de personnalité et de détection des émotions…), font rarement illusion très longtemps.

SURVEILLER LA COÉVOLUTION DES HUMAINS ET DES MACHINES

Il reste donc à trouver des façons pertinentes d'évaluer les diverses capacités, tant cognitives qu'émotionnelles, des machines. Pour ce qui est du cognitif, de nombreux chercheurs imaginent des scénarios inspirés du test de Turing, mais composés de plusieurs épreuves dans des contextes différents. Dans des versions scolaires, les machines passent des examens, comme des écoliers. D'autres évaluations portent sur la capacité à démêler des ambiguïtés du langage naturel, sur la préhension d'objets, etc. Quant aux capacités émotionnelles, la conception de tests en la matière constitue un nouveau chantier de la recherche en intelligence artificielle.

Ainsi, avec des *chatbots* munis de facultés que l'on considérait comme le propre des humains il n'y a pas si longtemps, des variantes du test de Turing ou des batteries d'examens sont à développer – qui pourront être à la base d'une sorte d'autorisation de mise sur le marché et de contrôle technique pour robots tout au long de leur vie. C'est un axe important des recherches actuelles et de défis pour la normalisation. Enfin, il est nécessaire de surveiller la coévolution et la coadaptation des humains et des machines dans leurs interactions. Les *chatbots* pourront nous manipuler ou nous mentir pour notre bien-être, mais encore faudra-t-il s'assurer que c'est bien le cas. Il faut se préoccuper de l'évolution, à long terme, du comportement des humains vis-à-vis des agents conversationnels, afin d'éviter l'isolement social, un attachement démesuré à la machine ou des manipulations abusives. C'est aussi dans cette direction que doit évoluer le test de Turing. Il doit porter non seulement sur les capacités des machines, mais aussi sur leur propension à modifier nos croyances et notre comportement vis-à-vis d'elles.

Ajoutons, pour finir, qu'une bonne information sur les capacités de ces assistants conversationnels, sur la façon dont ils fonctionnent, sur les biais qu'ils peuvent présenter, permettrait aux utilisateurs de prendre plus de recul et de se comporter de façon adéquate pour relever ce grand défi : bien vivre avec les machines.

— L'autrice —

> Laurence Devillers est professeuse en intelligence artificielle à Sorbonne Université et dirige une équipe de recherche au CNRS-LISN (Saclay), où elle est responsable de la chaire IA HUMAAINE (https://humaaine-chaireia.fr).

— À lire —

> D. Adiwardana *et al.*, Towards a Human-like Open-Domain Chatbot, 2020. arxiv.org/abs/2001.09977

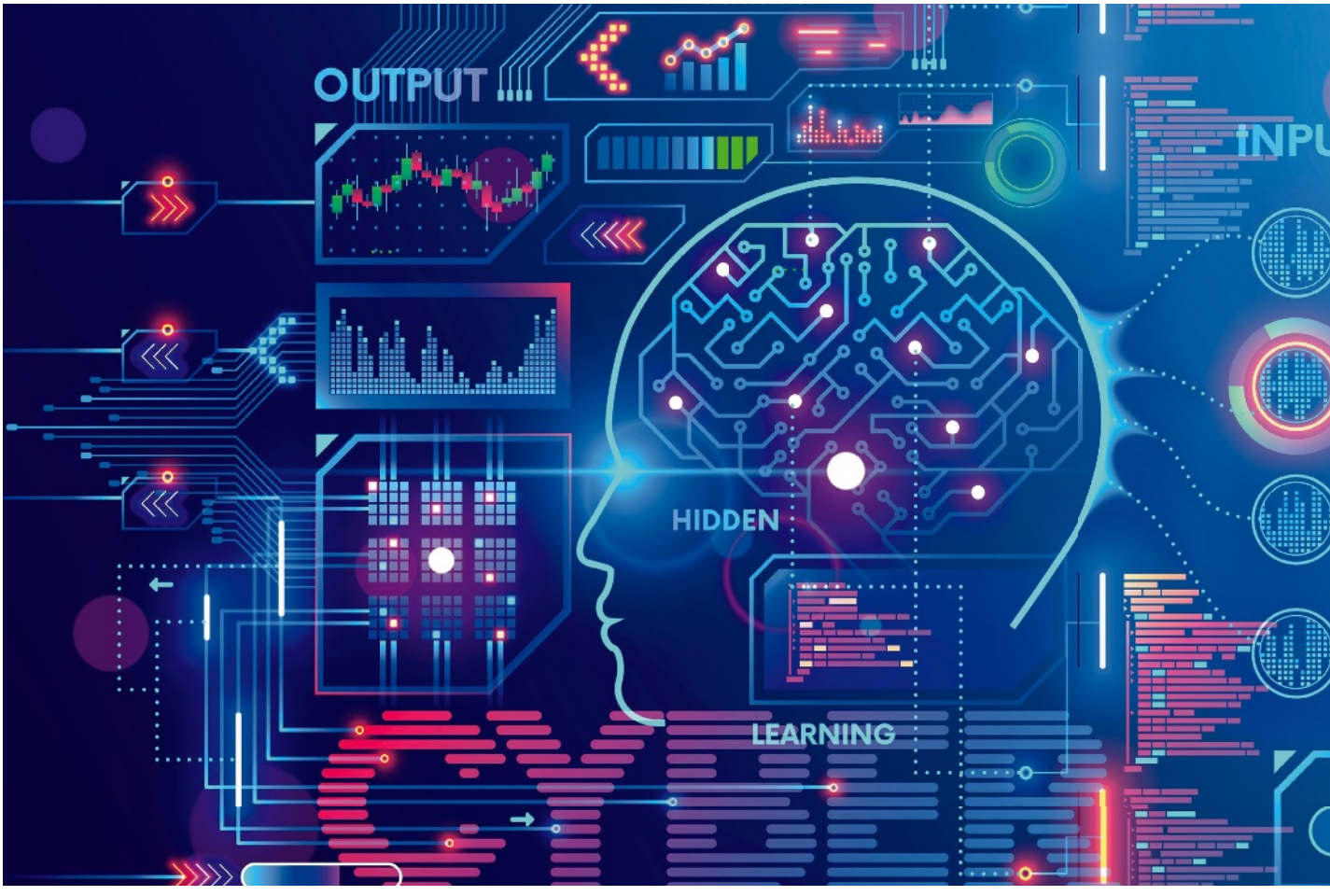
> L. Devillers, Des robots et des hommes : mythes, fantasmes et réalité, *Plon*, 2017.

> Vivre avec les robots, vidéo réalisée par le *Limsi-CNRS* : youtu.be/p1ID-gvUnWs

> Les robots en quête d'humanité, *Dossier Pour la Science*, n° 87, avril-juin 2015.

> A. M. Turing, Computing machinery and intelligence, *Mind*, 1950.

S'inspirer du cerveau pour mieux l'imiter



© Andrey Suslov/shutterstock.com

En matière d'énergie, le cerveau est beaucoup plus efficace que les ordinateurs actuels pour mémoriser, calculer ou apprendre. Est-il possible de s'inspirer de cet organe pour réaliser des dispositifs plus performants et mieux adaptés à l'intelligence artificielle ? Oui, en s'appuyant sur le spin des électrons…

Julie Grollier et Damien Querlioz

Les algorithmes d'intelligence artificielle ont fait des progrès considérables ces dernières années, et l'on peut les découvrir dans les articles de ce numéro. Cependant, pour parvenir à de telles performances, ces algorithmes consomment des quantités très élevées d'électricité, beaucoup plus que notre cerveau pour des tâches équivalentes. Par exemple, les modèles d'intelligence artificielle de traitement du langage humain, comme BERT (ou sa version française FlauBERT) nécessitent une énergie de 1 000 kilowattheures pour être entraînés. C'est autant que consomme le cerveau humain en six ans !

L'intelligence artificielle moderne, qui s'appuie sur des réseaux de neurones artificiels dits « profonds », s'inspire pourtant partiellement de la structure du cerveau *(voir les Repères, page 6)*. Pourquoi donc un tel contraste dans les performances ? Les programmes d'intelligence artificielle fonctionnent aujourd'hui sur des ordinateurs classiques, ou sur leurs cousins, les cartes graphiques, dont l'architecture et le principe de calcul sont bien différents de ceux du cerveau.

Pour résoudre ce défi énergétique, chercheurs et industriels travaillent à une électronique dont le fonctionnement serait vraiment calqué sur celui du cerveau. C'est le cas de la spintronique, fondée sur certaines propriétés quantiques des électrons. Des dispositifs reposant sur cette technique sont capables de simuler le comportement des neurones du cerveau et des synapses qui les relient. Le stade de l'industrialisation d'une solution complète fondée sur la spintronique est encore loin, mais les premiers résultats sont déjà très prometteurs.

Dans les ordinateurs actuels, l'obstacle majeur pour économiser de l'énergie et gagner en vitesse est lié au fait que les informations sont stockées dans une mémoire physiquement séparée des processeurs, où les données sont transformées en suivant les instructions du programme. Or les algorithmes d'intelligence artificielle comportent des millions, voire des milliards, de paramètres. Lorsqu'un tel programme opère, il doit sans cesse chercher des informations en mémoire puis les additionner ou les multiplier dans le processeur, avant de les replacer dans la zone de stockage. Ce flux permanent de données par une voie de communication unique consomme une quantité d'électricité considérable.

RAPPROCHER CALCUL ET MÉMOIRE

Dans le cerveau, le fonctionnement est très différent. La mémoire est partout : elle est distribuée et stockée dans les synapses, les connexions entre les neurones qui, eux, exécutent des opérations sur les données. Chaque neurone est connecté en moyenne à dix mille synapses, ce qui permet notamment des calculs en parallèle. Cette juxtaposition entre calcul et mémoire, doublée d'un traitement parallèle, est d'une redoutable efficacité énergétique.

La nature même du calcul est une autre source de gain considérable d'énergie. Nos ordinateurs manipulent des données très précises, chacune étant codée sous la forme de bits d'information. Les opérations sur ces données se

font selon des opérations logiques d'une rigueur implacable. Cette précision, assurément cruciale pour de nombreuses applications, comme envoyer une fusée dans l'espace, ne l'est pas pour d'autres, comme reconnaître un visage familier. Or le fait d'imposer à un circuit électronique un comportement parfait où l'aléa n'est pas toléré a un coût élevé en énergie, car il faut éliminer tous les bruits parasites et corriger toutes les erreurs.

Au cours de l'évolution, le cerveau, qui ne dispose que d'une source d'énergie limitée – nos apports caloriques quotidiens – a développé une tout autre stratégie. Il fonctionne sur la base d'un compromis entre la fiabilité du traitement de l'information et la dépense énergétique. Les synapses et les neurones ont ainsi des comportements stochastiques : en intégrant une certaine marge d'aléatoire, ils ne donnent pas toujours un résultat identique, même s'ils reçoivent des informations similaires.

En contrepartie, pour assurer tout de même un traitement fiable de l'information, le cerveau met en œuvre de multiples stratégies. Par exemple, comme chaque neurone, pris individuellement, n'est pas très fiable, plusieurs de ces cellules cérébrales sont parfois chargées de traiter le même signal. Un résultat cohérent et stable est ensuite déduit de l'ensemble de leurs réponses. La redondance est, en fin de compte, moins coûteuse en énergie qu'un système trop rigide qui corrige systématiquement les erreurs.

De plus, les signaux stockés dans les synapses et traités par les neurones ne sont pas des « 0 » ou des « 1 » (codage binaire) comme dans l'électronique actuelle. Ces signaux présentent une large gamme de valeurs différentes et très proches : ils

En bref

> En exploitant le spin, une propriété quantique des électrons, on conçoit des composants électroniques aux caractéristiques inédites.

> Grâce à la spintronique, il est envisageable de placer des unités de mémoire au plus près des processeurs, à l'image des synapses et des neurones du cerveau.

> Les premiers prototypes ont démontré que les dispositifs reposant sur la spintronique sont peu énergivores et ont des performances comparables aux algorithmes de réseaux de neurones usuels.

contiennent ainsi bien plus d'information, mais sont plus sensibles aux perturbations, autrement dit au « bruit ». Enfin, contrairement aux processeurs où chaque élément opère au rythme d'une horloge unique, assurant un fonctionnement cohérent, mais limitant la vitesse du traitement de l'information, le cerveau travaille de façon asynchrone : chaque neurone émet des signaux électriques selon son propre rythme et s'en sert pour communiquer efficacement avec d'autres neurones, qui se trouvent parfois dans une zone différente du cerveau.

LE CERVEAU, SOURCE D'INSPIRATION

Ces spécificités rendent le cerveau plus économe en énergie. Est-il possible de s'en inspirer pour construire des composants électroniques ou puces capables de traiter des données avec très peu d'énergie ? Avec des neurones et des synapses électroniques installés au plus près, de façon à rapprocher mémoire et calcul, on reproduirait l'architecture d'un réseau de neurones que l'on veut faire fonctionner. Cette idée n'est pas récente. Dès la fin des années 1980, Carver Mead, chercheur à Caltech (l'institut de technologie de Californie) et l'un des pères fondateurs de la microélectronique moderne, a inventé le concept de circuits « neuromorphiques », dont l'objectif est de mimer l'architecture neurobiologique du cerveau. Cette approche a longtemps été délaissée, car les obstacles étaient nombreux.

Le premier obstacle concerne la mémoire : les techniques actuelles empêchent aujourd'hui d'en intégrer d'importantes quantités au plus près du calcul. En effet, les seuls circuits de

Dès la fin des années 1980, Carver Mead a imaginé des circuits « neuromorphiques » dont l'objectif est de mimer l'architecture neurobiologique du cerveau

mémoire qui peuvent être intégrés dans les processeurs (les circuits de mémoire vive statique) sont extrêmement coûteux, car ils utilisent une grande et précieuse surface. C'est pour cette raison que les processeurs n'intègrent que de très faibles capacités de stockage, la plus grande part de la mémoire étant physiquement distribuée en dehors des zones de calcul. Plusieurs pistes sont néanmoins à l'étude pour rapprocher mémoire et processeurs. Parmi elles, l'électronique de spin, ou spintronique, a connu des progrès spectaculaires ces dernières années.

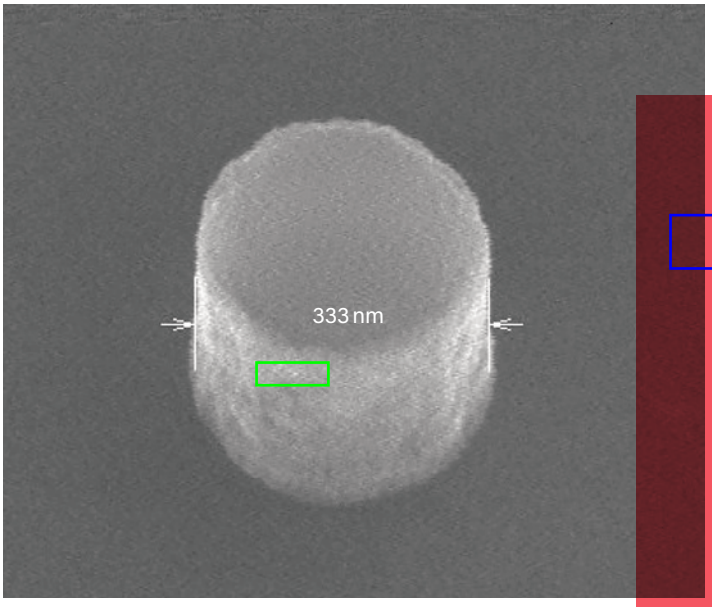
Cette technique n'exploite pas uniquement les électrons à travers leur charge électrique, mais utilise aussi leur spin, une propriété purement quantique, assimilable à un moment cinétique (on peut représenter le spin d'un électron comme une petite flèche orientée). Comme le spin est responsable du moment magnétique intrinsèque de l'électron, quand une majorité d'électrons ont un spin pointant dans une direction privilégiée dans un matériau, ce dernier a une aimantation non nulle. La spintronique exploite les interactions des spins portés par les électrons du courant électrique avec des matériaux magnétiques.

Concrètement, les briques de base de la spintronique sont des petits cylindres d'une dizaine de nanomètres de diamètre formant des « jonctions tunnel magnétiques ». Ces jonctions sont constituées de deux couches de matériaux magnétiques (des nanoaimants) séparées par une couche isolante. Quand les deux nanoaimants sont orientés dans le même sens, un courant électrique traverse facilement, par l'« effet tunnel », la couche isolante de la jonction : la résistance est faible. Si les orientations sont opposées, le courant circule difficilement : la résistance est alors élevée.

Ces propriétés permettent de stocker de l'information dans un format binaire (« 0 » ou « 1 ») selon les deux cas d'orientation possible *(voir l'encadré page 108)*. Les composants de mémoire ainsi formés sont nommés MRAM *(magnetoresistive random-access memories)*. Ils ont commencé à être développés vers le milieu des années 1980. Cependant, pour changer la valeur d'un bit, il fallait appliquer un champ magnétique externe afin d'inverser l'orientation d'un nanoaimant. C'était un obstacle à la miniaturisation de ces composants.

En 1996, John Slonczewski, chez IBM, et Luc Berger, de l'université Carnegie-Mellon, aux États-Unis, ont indépendamment proposé une amélioration importante de ces composants. Ils ont mis en évidence un nouvel effet, le transfert de spin. Lorsque les électrons d'un courant traversent un nanoaimant de la jonction, leur spin interagit avec celui des électrons dans la couche : le spin des électrons du courant se polarise et s'aligne avec l'aimantation de la couche. Dans une jonction, l'aimantation d'un des deux nanoaimants est fixe tandis que l'autre peut être modifiée. Quand un courant polarisé traverse cette dernière couche, elle réorientente l'aimantation. Selon le sens de circulation du courant, il est possible d'écrire un bit « 0 » ou un bit « 1 », et cela sans faire appel à un champ magnétique extérieur. Cette découverte a ouvert la porte à la conception de MRAM intégrées.

Depuis, plusieurs fabricants produisent des dispositifs de mémoire qui exploitent cette technique en intégrant jusqu'à 1 milliard de ces composants



Les jonctions tunnel magnétiques *(ici vue au microscope électronique à balayage)* sont fabriquées en plusieurs étapes. On fait croître les couches de matériaux l'une après l'autre par pulvérisation cathodique dans une enceinte sous vide. Ensuite, pour obtenir des jonctions d'un diamètre inférieur au micromètre, le matériau est recouvert d'une résine électrosensible et d'un masque sur lequel est dessinée la forme de la jonction. L'ensemble est soumis à une source d'électrons (lithographie) et la résine exposée est conservée. Les parties non protégées par la résine du matériau sont gravées et retirées pour ne garder que la forme de la jonction.

sur une puce de silicium. Ces mémoires trouvent déjà des applications en électronique usuelle, car elles associent les qualités des deux grandes familles de mémoires informatiques : la vitesse de lecture et d'écriture des mémoires vives et la capacité des mémoires de stockage à conserver durablement les informations.

On se rapproche de l'idée de puces neuromorphiques, car il est possible d'insérer facilement des jonctions tunnel magnétiques aux endroits les plus utiles pour stocker des données sous forme de bits au sein des circuits en silicium qui réalisent les calculs. Comme dans le cerveau, où les synapses sont au contact des neurones, ces synapses artificielles seront au plus près du cœur de calcul en silicium. Pour utiliser un algorithme de réseau de neurones dans ce dispositif, il suffit de configurer toutes ces synapses avec les bonnes valeurs pour définir la tâche du programme. Selon le jeu de valeurs mis en place, le réseau sera en mesure d'identifier des éléments dans des images, du son…

En réduisant drastiquement les échanges de données entre mémoire externe et processeur, on exécuterait des algorithmes d'intelligence artificielle en dépensant moins d'énergie. Cette piste est activement explorée dans les laboratoires académiques et industriels. Elle n'est cependant pas entièrement satisfaisante : il faut connaître à l'avance les bonnes valeurs à mémoriser par les synapses pour effectuer une tâche. Le dispositif n'a pas les capacités d'adaptation d'un vrai cerveau. Or on souhaite, idéalement, un circuit qui soit en mesure d'apprendre de nouvelles tâches.

INITIALISER LES SYNAPSES

Dans les algorithmes de réseaux de neurones, on a souvent une première phase, dite « d'apprentissage », où le réseau ajuste la valeur des synapses afin d'exécuter de mieux en mieux la tâche assignée. Malheureusement, la technique par laquelle l'algorithme identifie les bonnes valeurs des synapses dans les réseaux de neurones artificiels est fondée sur une méthode mathématique complexe, la « rétropropagation du gradient » *(voir les Repères, page 6)*. Celle-ci exige des calculs nombreux, très précis et donc énergivores. Elle est bien adaptée sur un ordinateur, mais pas nécessairement pour un système inspiré du cerveau fonctionnant avec des calculs moins précis.

Pour rester dans la logique des systèmes neuromorphiques, il serait plus judicieux de s'inspirer encore une fois de la façon dont le cerveau apprend. En particulier, les synapses biologiques ne servent pas simplement comme mémoire, elles sont aussi au cœur du mécanisme de l'apprentissage. Elles sont en effet « plastiques » : elles

© AIST

ajustent leur activité selon la quantité d'information à laquelle elles sont exposées. Par exemple, quand deux neurones connectés par une synapse sont actifs en même temps, cette synapse s'en trouve renforcée avec une activité accrue. Mais ce renforcement n'est pas toujours complet, sinon, d'une certaine façon, chaque nouveau signal effacerait la subtile configuration de la synapse élaborée au cours de toute son expérience passée.

En transposant ce principe aux dispositifs spintroniques, cela reviendrait à modifier la valeur des synapses artificielles en fonction du signal qu'elles reçoivent. Et effectivement, il est possible de reproduire une certaine plasticité avec ces systèmes. Quand une tension électrique est appliquée à une jonction tunnel magnétique, l'effet de transfert de spin met en mouvement l'aimantation d'un de ses nanoaimants. Si la tension est assez élevée et appliquée assez longtemps, ce mouvement entraîne l'inversion de l'aimantation du nanoaimant, et donc de la valeur stockée par la jonction tunnel magnétique. Mais ce renversement n'est pas toujours prédictible : si la tension n'est pas appliquée pendant une durée suffisante, parfois l'aimantation se retourne, parfois non. Plus la tension appliquée à la jonction est élevée, plus la probabilité de basculement de l'aimantation est importante, sans pour autant atteindre 100 %.

UNE CAPACITÉ D'APPRENTISSAGE

Dans une mémoire usuelle, ce comportement incertain doit être évité à tout prix. Pour un système neuromorphique, on obtient des synapses qui, sous l'action des signaux qu'elles reçoivent, transforment progressivement la fonction du réseau de neurones dont elles font partie et adaptent la réponse du système à la nouvelle tâche sans avoir besoin d'utiliser des méthodes aussi lourdes que la rétropropagation du gradient.

En 2015, avec nos collègues, nous avons montré qu'un système équipé de mémoires spintroniques plastiques pouvait apprendre à compter des véhicules sur une autoroute (avec une précision de 95 %) ou à reconnaître des chiffres manuscrits, de la même façon que l'aurait fait un algorithme usuel d'intelligence artificielle. Sur ces tâches relativement simples, les performances s'approchent des techniques plus traditionnelles d'apprentissage des réseaux de neurones (qui reposent cependant sur plus de synapses, avec une consommation plus élevée en énergie). En revanche, sur des tâches complexes (comme concurrencer un spécialiste du jeu de go), les techniques d'apprentissage bio-inspirées sont pour l'instant loin de rivaliser avec les techniques plus conventionnelles.

Si nous sommes capables de réaliser des synapses artificielles qui stockent de l'information et dont il est possible de moduler la réponse pour simuler la plasticité synaptique et le phénomène d'apprentissage, un système complet neuromorphique doit aussi inclure des neurones pour effectuer les calculs.

Le cerveau humain contient près d'une centaine de milliards de neurones, et les algorithmes d'intelligence artificielle simulent aujourd'hui jusqu'à des dizaines de millions de neurones. Est-il envisageable de réaliser des neurones avec les techniques actuelles de transistors au silicium ? D'abord, si l'on veut autant de composants dans une puce électronique de la taille d'un pouce, il faudrait que chaque neurone ait une largeur bien inférieure à 1 micromètre, ce qui est impossible, car réaliser même un unique neurone artificiel par cette approche nécessite des centaines de transistors.

Par exemple, la puce TrueNorth conçue par IBM en 2014 est une démonstration impressionnante d'électronique neuromorphique qui ne fait pas appel à la spintronique. Ce système comporte cinq milliards de transistors (un nombre considérable), ce qui permet d'implémenter 1 million de neurones et des centaines de millions de synapses. En rapprochant ainsi centre de calcul et mémoire, IBM affirme que la puissance consommée par la puce est inférieure de plusieurs ordres de grandeur à celle d'une puce usuelle.

Avec la spintronique, il est possible de reproduire une certaine plasticité des synapses, ici artificielles

En réalité, l'architecture de la puce TrueNorth ne contient pas 1 million de neurones. Un neurone artificiel de ce type occupe une trop grande surface pour en mettre 1 million sur une puce. En revanche, cette puce est extrêmement rapide ; dès lors, un seul circuit numérique implémente plusieurs neurones artificiels (ceux qui seraient physiquement situés autour), en faisant successivement les calculs associés à chacun d'eux. C'est une petite, mais très pragmatique, entorse à l'idée de base du neuromorphisme de faire des neurones et synapses artificiels. En 2017, Intel a développé une puce analogue, mais plus modeste, Loihi, qui n'implémente que 130 000 neurones. Ces dispositifs restent volumineux et il est difficile de penser qu'il sera possible de rivaliser avec le cerveau humain en suivant cette approche.

AU PLUS PRÈS D'UN NEURONE

La spintronique fournit là encore une piste. Il est possible d'utiliser une jonction tunnel magnétique pour imiter les fonctions clés des neurones du cerveau. En particulier, ces derniers communiquent en émettant des impulsions électriques. Ces impulsions sont toutes identiques, mais leur nombre dépend de la sollicitation du neurone. Si un neurone reçoit beaucoup d'impulsions pendant un temps court, la fréquence avec laquelle il va engendrer des impulsions sera alors également élevée. S'il est stimulé par peu d'impulsions, il en émettra lui-même un nombre réduit.

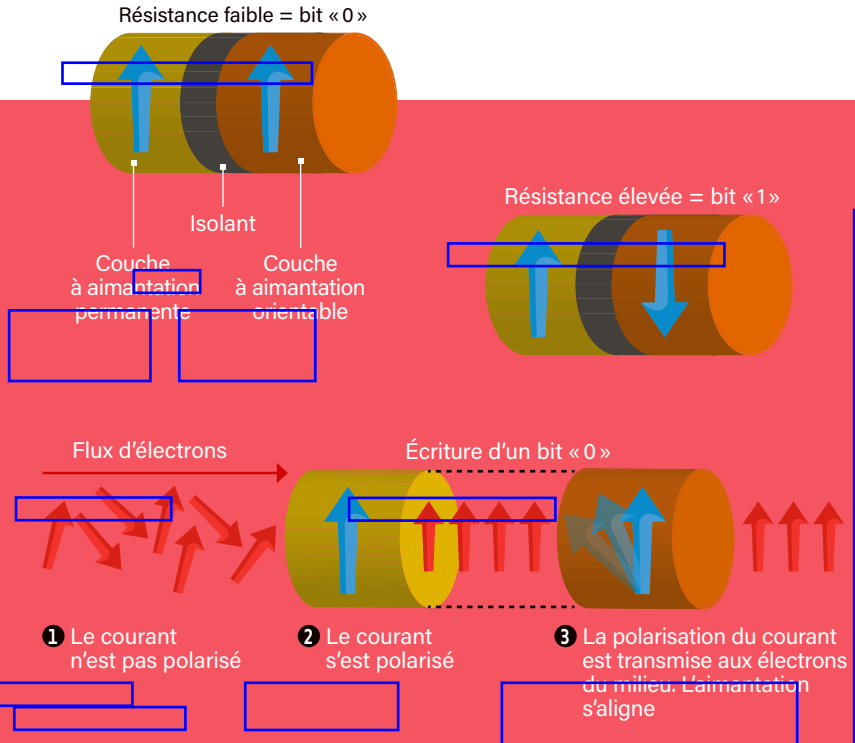
Avec l'électronique traditionnelle, il est difficile de réaliser des neurones impulsionnels de taille nanométrique. En effet, produire des impulsions à un certain rythme nécessite de créer des boucles de rétroaction dans le circuit électronique, ce qui occupe de la place. Il est en revanche possible de fabriquer des jonctions tunnel magnétiques qui reproduisent ce comportement.

Une solution consiste à prendre les signaux d'entrée sous la forme de courants électriques qui sont additionnés (exactement comme dans un neurone biologique) avant d'être injectés dans la jonction. Ensuite, on choisit les caractéristiques du dispositif de telle sorte que le spin transféré par les électrons du courant entrant ne conduise pas à un basculement complet de l'aimantation du nanoaimant, mais la fasse osciller et tourner en permanence, telle une boussole dont l'aiguille pivoterait sur elle-même. Comme la configuration respective des deux aimantations (celle fixe et celle en rotation) change périodiquement,

MANIPULER DES SPINS

Dans les circuits électroniques actuels, la manipulation de la charge électrique de l'électron est au cœur du fonctionnement des dispositifs, à l'instar des transistors. En électronique de spin, ou spintronique, on utilise, en plus de la charge, une autre propriété quantique des électrons, le spin. Celui-ci correspond à un moment magnétique intrinsèque de la particule. Les interactions de ce spin avec certains matériaux magnétiques offrent la possibilité de concevoir des composants aux comportements très riches. Dans un matériau magnétique, comme le fer ou le cobalt, le spin des électrons est globalement orienté dans une même direction, ce qui confère au matériau son aimantation. Lorsqu'un courant électrique traverse un tel matériau, le spin des électrons qui constituent le courant s'oriente dans la direction de l'aimantation. Cet effet donne lieu à deux phénomènes fondamentaux en spintronique.
Le premier est le phénomène de magnétorésistance observé dans les « jonctions tunnel magnétiques ». Ces composants sont constitués de deux couches aimantées séparées par une couche isolante. Les matériaux utilisés dépendent des propriétés que l'on veut donner à la jonction. L'isolant est souvent un oxyde métallique d'environ 1 à 2 nanomètres d'épaisseur. L'aimantation d'une couche magnétique est fixe tandis que l'autre peut être réorientée.
En 1975, Michel Jullière, de l'Insa de Rennes, a été le premier à mettre en évidence la magnétorésistance à effet tunnel. Lorsqu'un courant

© Pour la Science/Charlotte Calament

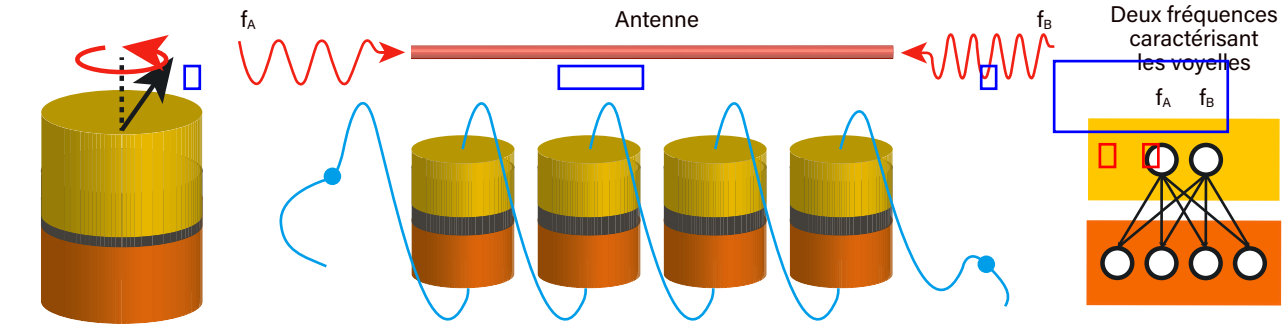


électrique traverse une jonction tunnel magnétique, la traversée de la couche isolante s'effectue grâce à l'effet tunnel (un phénomène purement quantique). Le courant passe plus facilement si les aimantations des deux couches magnétiques pointent dans la même direction, et plus difficilement dans le cas contraire.

Ainsi, en mesurant la résistance électrique (faible ou élevée) d'une jonction tunnel magnétique, on peut « lire » la direction des aimantations. On associe ainsi un bit « 0 » à des aimantations parallèles et un bit « 1 » à des aimantations antiparallèles. Ces jonctions tunnel magnétiques peuvent ainsi servir d'élément de base dans des systèmes de mémoire pour stocker des bits d'information. Une importante activité industrielle s'est développée autour des mémoires reposant sur ce principe et qualifiées de MRAM *(magnetoresistive random-access memory)*.

La découverte d'un effet analogue dans des systèmes multicouches métalliques, la magnétorésistance géante, en 1988, a été décisive pour l'intérêt porté au développement de la spintronique, et a valu le prix Nobel de physique en 2007 à Albert Fert, de l'université Paris-Saclay, et Peter Grünberg, de l'université de Cologne. Cette technique et la magnétorésistance tunnel sont utilisées notamment dans les têtes de lecture des disques durs modernes. Le fait d'avoir une couche dont l'aimantation est « mobile » permet de réorienter cette dernière pour que la jonction porte un bit « 0 » ou « 1 ». Pour écrire ce bit, l'approche initiale a été d'utiliser un champ magnétique extérieur, mais cette méthode ne permet pas une bonne réduction de la taille du système et son intégration à des échelles nanométriques.

En 1996, John Slonczewski, chercheur chez IBM, et Luc Berger, de l'université Carnegie-Mellon, aux États-Unis, ont mis en évidence le second phénomène crucial de la spintronique : le transfert de spin. Lorsqu'un courant électrique suffisamment intense traverse une jonction tunnel magnétique, il peut faire basculer l'aimantation de la couche à aimantation mobile pour obtenir une configuration parallèle ou antiparallèle. Pour écrire un bit « 0 » *(voir ci-contre)*, le courant traverse d'abord la couche à aimantation fixe. Il se polarise : le spin des électrons du courant, qui n'avait initialement pas d'orientation particulière, s'aligne avec l'aimantation de la couche. Quand les électrons arrivent dans la seconde couche, ils interagissent avec les électrons du matériau et leur transfèrent leur polarisation. L'aimantation de cette couche devient progressivement parallèle à l'autre. Pour être plus précis, à chaque interface (entre une couche magnétique et l'isolant), une partie des électrons est réfléchie et une autre est transmise. Pour écrire un bit « 1 », le courant arrive par la couche à aimantation mobile. Il se polarise d'abord selon l'aimantation de cette couche. Mais les électrons qui sont réfléchis sur la couche à aimantation fixe prennent une polarisation un peu plus antiparallèle et transmettent cette orientation à la couche à aimantation mobile. Résultat : on obtient une configuration antiparallèle, correspondant à un bit « 1 ». Depuis 2008, de nombreux industriels exploitent les jonctions tunnel magnétiques à transfert de spin pour concevoir des dispositifs de mémoire performants.



Une jonction tunnel magnétique peut être utilisée comme un neurone qui traite l'information. Pour cela, les caractéristiques de la jonction sont telles que le courant qui la traverse fait tourner de façon permanente l'aimantation de la couche supérieure *(à gauche)*. La jonction émet alors un signal impulsionnel comme un neurone biologique. Ce composant se comporte aussi comme un émetteur-récepteur d'ondes radio. Dans une expérience, les chercheurs ont associé quatre neurones artificiels qui synchronisaient aussi leurs signaux grâce à une antenne *(au centre)*. Ce réseau de neurones *(schématisé à droite)* reconnaît des voyelles prononcées.

la résistance électrique de la jonction (on parle d'effet de magnétorésistance) s'en trouve affectée et se traduit par un signal électrique périodique. Plus le courant entrant est important, plus la vitesse de rotation sera grande et la fréquence des impulsions engendrées sera élevée. On obtient bien un comportement très similaire à celui d'un neurone. En outre, ce phénomène est réalisable à température ambiante et est très stable pour des dispositifs nanométriques. Il peut donc être exploité pour des dispositifs concrets.

99,6 % DE RÉUSSITE

En 2017, avec nos équipes, nous avons montré expérimentalement qu'une jonction tunnel magnétique pouvait effectivement imiter un neurone. Nous sommes même allés plus loin : nous avons utilisé une jonction unique pour imiter un réseau entier de 400 neurones grâce à une stratégie dite « de multiplexage temporel », la jonction jouant tour à tour le rôle de chaque neurone. Nous avons ensuite utilisé ce réseau de neurones spintroniques pour identifier des chiffres prononcés par différentes personnes.

Pour ce faire, nous avons converti les signaux audio à reconnaître un courant électrique injecté à travers la jonction tunnel magnétique. Ce courant injecté, qui correspond aux chiffres prononcés mille fois plus vite (pour être dans la bonne gamme de fréquences du dispositif), provoque une rotation de l'aimantation d'un nanoaimant de la jonction. Ce faisant, grâce à l'effet de magnétorésistance, le neurone convertit les modulations du courant en variation de la tension électrique à ses bornes. Nous avons enregistré ces variations de tension électrique provenant des neurones spintroniques, que nous avons transmises à un ordinateur qui a simulé l'ensemble du réseau de neurones avec ses fonctions synaptiques. Ces dernières ont alors été configurées grâce à une phase d'apprentissage pour reconnaître les chiffres parlés. Puis, soumis à de nouveaux échantillons de paroles, le réseau identifiait correctement les chiffres avec un taux de succès de 99,6 %, ce qui représente une performance exceptionnelle. Cette expérience a démontré la capacité des jonctions tunnel magnétiques impulsionnelles à imiter les neurones de façon fiable.

Les travaux dans le domaine de la spintronique à des fins neuromorphiques sont encourageants. Notre futur défi sera de réaliser des réseaux de neurones complets comportant des millions de jonctions tunnel magnétiques jouant le rôle de neurones et de synapses. Comme on l'a vu, l'industrie sait déjà les intégrer par milliards sur des puces et les relier à des circuits électroniques classiques pour faire des mémoires, ce qui constitue une base solide pour aller plus loin.

© Pour la Science/Charlotte Calament

La spintronique n'a pourtant pas fini de nous surprendre. Les jonctions tunnel magnétiques seraient capables d'imiter d'autres propriétés du cerveau, comme la synchronisation de neurones distants. En effet, dans le cerveau, des impulsions conduisent des neurones parfois très éloignés à accorder leurs réponses, en les déclenchant au même moment. Cette synchronisation a de nombreux avantages. Des neurones qui émettent des impulsions en même temps auront plus d'importance dans le traitement du signal. S'ils sont connectés au même neurone, celui-ci reçoit un très grand nombre d'impulsions dans un bref laps de temps et, ainsi, il les transmet très efficacement à son tour. Des populations entières de neurones se coordonnent par ce phénomène, au point de créer des ondes cérébrales qui, grâce à leur puissance, modifient le comportement de neurones dans d'autres aires cérébrales, parfois lointaines.

DE MINUSCULES ÉMETTEURS-RÉCEPTEURS

Les jonctions tunnel magnétiques ayant un fonctionnement très rapide, elles produisent des impulsions électromagnétiques à une fréquence pouvant dépasser quelques centaines de mégahertz. Le traitement de l'information se fait ainsi des millions de fois plus vite que dans le cerveau. Mais, autre conséquence, ces dispositifs se comportent comme des nanoémetteurs radio. Or ces jonctions sont aussi très sensibles aux ondes électromagnétiques environnantes, lesquelles peuvent modifier leur génération d'impulsions : les jonctions sont donc aussi de très bons récepteurs radio. En ajustant correctement les caractéristiques d'un ensemble de jonctions, il est possible de les synchroniser entre elles grâce aux signaux qu'elles émettent et reçoivent… comme les neurones biologiques !

Induire ce type de comportement synchronisé dans des réseaux de neurones artificiels est une piste passionnante pour coordonner matériellement et efficacement des réseaux de neurones spécialisés dans des tâches différentes. Dans cette optique, en 2018, avec nos collègues, nous avons étudié un système de quatre jonctions tunnel magnétiques dont la tâche était de reconnaître des voyelles prononcées oralement *(voir la figure page ci-contre)*. Le signal vocal était réduit à deux fréquences (par analyse de Fourier) et accéléré 100 000 fois pour être émis par une antenne aux quatre nanooscillateurs. De façon très intéressante, dans cette expérience, nous avons constaté que la capacité des jonctions à coordonner leurs rythmes augmentait les performances de classification des signaux vocaux.

Pour aller plus loin, nous cherchons désormais à réaliser des réseaux dont les neurones communiquent par les ondes électromagnétiques qu'émettent et reçoivent les jonctions. Il s'agit là d'une piste prometteuse pour interconnecter les neurones dans un réseau dense, ce qui est une condition indispensable pour l'implémentation de tâches complexes. Un système complet associant ces jonctions jouant le rôle de neurones à des synapses artificielles à base de MRAM, pour lui donner une capacité d'apprentissage, est en cours de fabrication.

Que pourront alors réaliser de tels systèmes ? Notre objectif n'est pas d'obtenir une intelligence capable de rivaliser avec l'intelligence humaine : notre connaissance du cerveau est bien trop partielle pour un tel résultat. En revanche, nous souhaitons concevoir une intelligence artificielle plus économe en électricité et qui soit plus rapide que celles disponibles actuellement. Ainsi, nous réduirons la consommation énergétique considérable des centres de calcul qui traduisent des textes ou interprètent ce que nous disons à nos enceintes intelligentes, et nous aurons la possibilité d'embarquer de tels dispositifs dans nos téléphones portables ou dans des véhicules autonomes pour nous assister dans notre vie quotidienne.

— **Les auteurs** —

> Julie Grollier
est chercheuse à l'unité mixte de physique CNRS/Thales, à Palaiseau. Prix Irène Joliot-Curie 2021 de la Femme scientifique de l'année.

> Damien Querlioz
est chercheur au Centre de nanosciences et de nanotechnologies (université Paris-Saclay/CNRS), à Palaiseau.

— **À lire** —

> N. Leroux *et al.*, Hardware realization of the multiply and accumulate operation on radio-frequency signals with magnetic tunnel junctions, *Neuromorphic Computing and Engineering*, 2021.

> J. Grollier *et al.*, Neuromorphic spintronics, *Nature Electronics*, 2020.

> M. Romera *et al.*, Vowel recognition with four coupled spin-torque nano-oscillators, *Nature*, vol. 563, pp. 230-234, 2018.

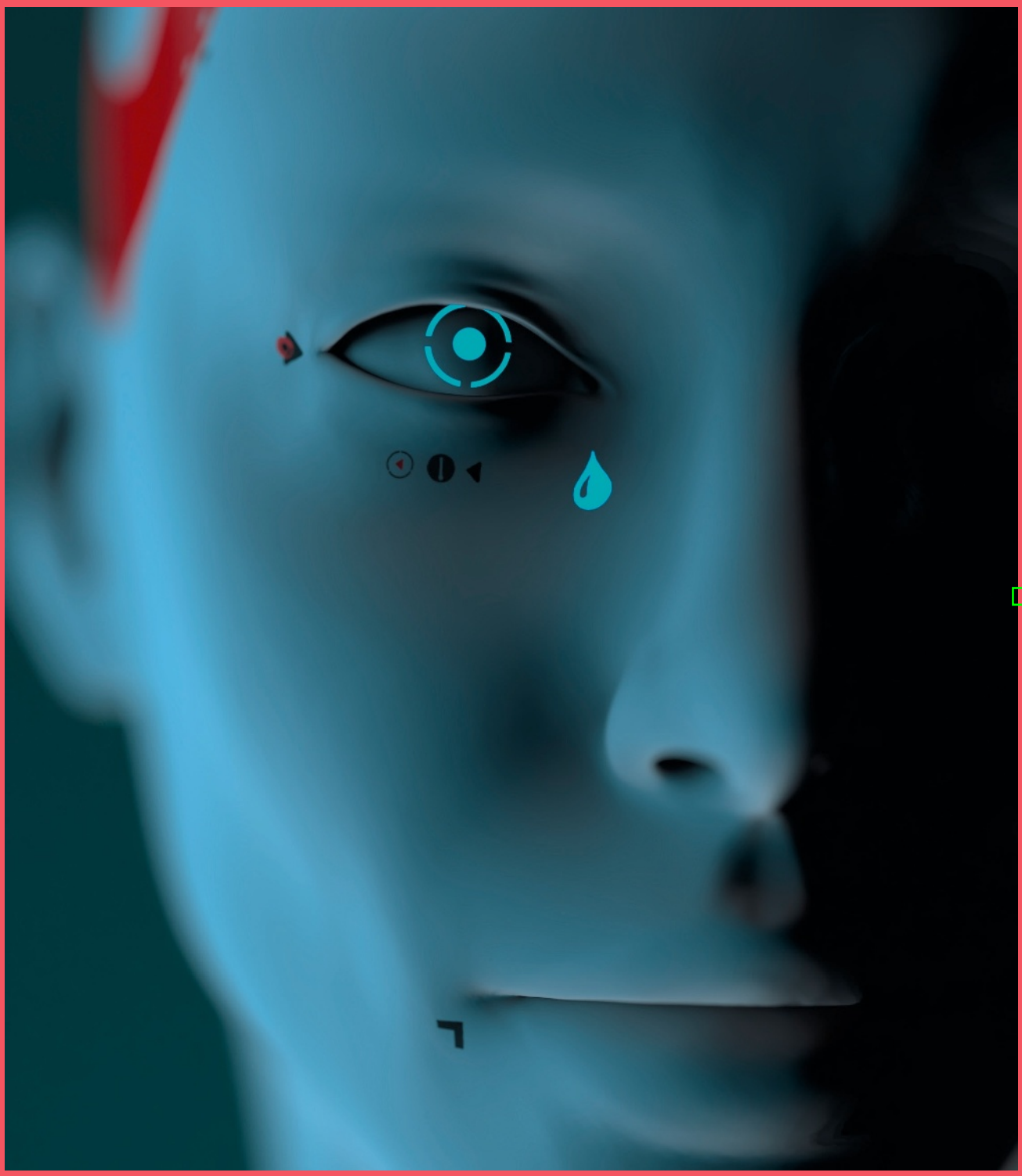
> J. Torrejon *et al.*, Neuromorphic computing with nanoscale spintronic oscillators, *Nature*, vol. 547, pp. 428-431, 2017.

L'« intelligence » viendra sans doute aux machines. Mais en sera-t-il de même pour la conscience ? On peut fortement en douter.

L'improbable conscience artificielle

Christof Koch

© evgeniy jamart/shutterstock.com



Dans les décennies à venir, les progrès rapides des algorithmes d'apprentissage engendreront des machines d'une intelligence comparable à la nôtre. Capables de parler et de raisonner, elles auront leur place dans une myriade de domaines tels que l'économie, la politique et, inévitablement, la guerre. La naissance d'une véritable intelligence artificielle (IA) aura un impact profond sur l'avenir de l'humanité et conditionnera l'existence même d'un tel avenir.

Prenez par exemple la citation suivante : « Encore aujourd'hui, des recherches sont en cours pour mieux comprendre ce que les nouveaux programmes d'IA seront capables de faire, tout en restant dans les limites de l'intelligence d'aujourd'hui. La plupart des programmes d'IA actuellement programmés se limitent principalement à prendre des décisions simples ou à effectuer des opérations simples sur des quantités relativement faibles de données. »

Peut-être avez-vous eu l'impression que quelque chose clochait dans ce paragraphe ? Cette citation est l'œuvre de GPT-2, un robot linguistique que j'ai pu tester. Développé par OpenAI, un institut basé à San Francisco qui promeut les IA « vertueuses », GPT-2 est un algorithme d'apprentissage fondé sur un réseau de neurones artificiels. Ses entrailles contiennent plus d'un milliard de connexions simulant des synapses, les points de jonction entre neurones.

La tâche du réseau est apparemment stupide : confronté à un texte de départ arbitraire, il doit prédire le mot suivant. Il ne « comprend » pas les textes comme le ferait un humain. Mais durant sa phase d'apprentissage, il a dévoré des quantités astronomiques de textes – huit millions de pages internet en tout – et ajusté ses connexions internes pour mieux anticiper des suites de mots.

J'ai écrit les premières phrases de l'article que vous êtes en train de lire, puis je les ai « injectées » dans l'algorithme en lui demandant de composer une suite. Il a notamment craché le paragraphe cité. Certes, ce texte ressemble aux efforts d'un étudiant de première année pour se rappeler un cours d'introduction à l'apprentissage automatique durant lequel il aurait rêvassé. Mais le résultat contient malgré tout les mots et les phrases clés – pas si mal, vraiment !

L'EFFET D'ÊTRE UNE CHAUVE-SOURIS

Les algorithmes sauront-ils un jour écrire un chef-d'œuvre aussi abouti qu'*À la recherche du temps perdu*, de Marcel Proust ? Difficile à dire, mais les prémices sont là. Rappelez-vous comme les premiers logiciels de traduction et de conversation étaient faciles à tourner en dérision, tant ils manquaient de précision. Mais avec les réseaux neuronaux profonds et la mise en place d'infrastructures de calcul surpuissantes par les entreprises du numérique, les ordinateurs se sont améliorés sans relâche, jusqu'à ce que leurs productions n'aient plus rien de ridicule.

Bien que les experts ne s'accordent pas sur la nature exacte de l'intelligence, la plupart d'entre eux estiment que, tôt ou tard, les ordinateurs atteindront ce que l'on appelle l'intelligence artificielle générale : ils deviendront capables de réaliser n'importe quelle tâche intellectuelle accessible à un humain. Mais auront-ils pour autant une forme de sentiment d'eux-mêmes ? En d'autres termes, pourront-ils un jour être conscients ?

En bref

> Deux théories s'opposent sur la nature de la conscience : celle de « l'espace de travail neuronal global » et celle de « l'information intégrée ».

> L'une implique que les machines deviendront conscientes lorsqu'elles auront des capacités suffisamment sophistiquées, l'autre non.

> Ces théories sont actuellement testées par une collaboration internationale.

> Le simulacre numérique d'un mort pourra parler, accéder à ses souvenirs, à ses envies, à ses peurs et aux traits de sa personnalité. Mais sera-t-il conscient ?

Par « conscience » ou « sentiment subjectif », j'entends une qualité inhérente à toute expérience – par exemple, la saveur d'une madeleine, une rage de dents, l'étirement du temps quand on s'ennuie ou le sentiment de vitalité et d'anxiété juste avant une compétition. En paraphrasant le philosophe Thomas Nagel, auteur d'un article titré « Quel effet cela fait-il d'être une chauve-souris ? », nous pourrions dire qu'un système est conscient si cela « fait quelque chose » *d'être* ce système.

Considérez l'embarras que vous ressentez lorsque vous vous apercevez que vous venez de commettre une gaffe, parce que votre interlocuteur a perçu comme une insulte ce qui n'était pour vous qu'un trait d'humour. Les ordinateurs éprouveront-ils un jour une telle émotion ? Et lorsque vous êtes au téléphone depuis un temps interminable, écoutant une voix synthétique répéter en boucle : « Nous sommes désolés de vous faire attendre », le logiciel pourra-t-il se sentir gêné de vous maintenir dans cet enfer du service client ?

Il ne fait guère de doute que notre intelligence et nos expériences résultent des capacités naturelles de notre cerveau à établir des enchaînements de causes et d'effets. Cette disposition a extrêmement bien servi la science au cours des derniers siècles. Le cerveau humain, cet organe d'à peine un kilo et demi à la texture comparable au tofu, est de loin le morceau de matière active organisée le plus complexe connu de l'Univers. Mais il obéit aux mêmes lois physiques que les pierres, les arbres et les étoiles. Rien n'échappe à ces lois. Nous ne comprenons pas encore tout à fait les mécanismes causaux en jeu, mais nous les expérimentons tous les jours : un groupe de neurones s'allume pendant que vous voyez des couleurs, tandis que l'activation de cellules nerveuses dans une région voisine du cortex est associée à une humeur joviale. Lorsqu'un neurochirurgien stimule ces neurones à l'aide d'une électrode, le sujet voit des couleurs dans le premier cas, éclate de rire dans le second. Inversement, quand le cerveau est sous anesthésie, ces expériences disparaissent.

Compte tenu de cela, l'évolution de l'intelligence artificielle débouchera-t-elle sur une conscience artificielle ?

DEUX THÉORIES ANTAGONISTES

Cette question mène à deux possibilités fondamentalement différentes. Selon la première, les machines vraiment intelligentes seront sensibles et dotées de conscience : elles parleront, raisonneront, s'occuperont d'elles-mêmes, seront capables d'introspection… Une idée dans l'air du temps, comme on le voit dans des romans et des films tels que *Blade Runner*, *Her* ou *Ex Machina*.

D'un point de vue scientifique, cette hypothèse s'appuie sur l'une des principales théories de la conscience, dite de « l'espace de travail neuronal global » (GNW, pour *global neuronal workspace)*. Cette théorie considère que l'origine de la conscience réside dans certaines caractéristiques architecturales du cerveau. Elle s'inspire en cela de l'« architecture du tableau noir », développée par les sciences informatiques dans les années 1970 : le principe est que des programmes spécialisés accèdent à un répertoire commun d'informations, appelé « tableau noir »

La France a-t-elle une conscience ?

Selon la théorie de l'information intégrée, le degré de conscience d'un système dépend de son pouvoir causal intrinsèque – qui quantifie la capacité de ses sous-systèmes à s'influencer mutuellement. Mais alors, tout ensemble doté de parties en interaction n'est-il pas conscient ? N'est-ce pas le cas par exemple de l'ensemble que vous formez avec votre interlocuteur au téléphone, voire de celui formé par la population d'un pays entier ? Non, car la théorie postule qu'à l'intérieur d'un système, la conscience apparaît précisément là où le pouvoir causal intrinsèque est maximum. Si vous parlez avec quelqu'un, vous allez vous influencer mutuellement, mais de façon très inférieure à ce qui se passe dans le cerveau de chacun. Il n'y aura donc pas formation d'un « superesprit » – pour cela, il faudrait augmenter drastiquement les connexions entre vos deux cerveaux. Il en va de même pour un pays comme la France et ses 67 millions d'habitants : même s'il y a de multiples interactions entre ces derniers, le pouvoir causal intrinsèque reste maximal à l'intérieur du cerveau de chacun. C'est donc là que surgit la conscience.

ou « espace de travail central ». Les psychologues ont postulé qu'un espace de ce type existe dans le cerveau et qu'il est essentiel à la cognition humaine. Sa capacité est faible, de sorte qu'à un moment donné, il ne contient qu'une seule perception, qu'une pensée unique ou qu'un seul souvenir. Dans cet espace de travail, les nouvelles informations entrent en concurrence avec les anciennes et les remplacent.

Le neuroscientifique Stanislas Dehaene et le biologiste moléculaire Jean-Pierre Changeux, tous deux au Collège de France, à Paris, ont transposé ces idées à l'architecture du cortex, la couche externe du cerveau. Ils ont postulé que l'espace de travail repose sur un réseau de neurones dits « pyramidaux », qui relient des régions corticales éloignées, en particulier les zones associatives préfrontales, pariétotemporales et médianes (cingulaires).

Une grande partie de l'activité cérébrale reste localisée et de ce fait n'accède pas à la conscience. C'est le cas des modules neuronaux qui contrôlent la posture du corps ou la direction du regard. Mais lorsque l'activité d'une ou plusieurs régions dépasse un seuil critique – disons, lorsqu'on présente à quelqu'un l'image d'une friandise appétissante –, elle déclenche une vague d'excitation neuronale qui se propage à travers l'espace de travail, puis dans tout le cerveau. Ce signal devient alors accessible à une multitude de processus auxiliaires, tels que le langage, la planification, le circuit de la récompense, la mémoire à long terme et le stockage dans une mémoire tampon à court terme.

Ce serait le fait de diffuser cette information à l'échelle globale qui la rendrait consciente.

Ainsi, l'expérience subjective de la friandise appétissante résulterait de l'action de neurones pyramidaux qui ordonnent à la région motrice du cerveau de se saisir de la friandise, tandis que d'autres modules neuronaux anticipent une récompense sous la forme d'une poussée de dopamine, causée par le petit *shoot* de sucre et matières grasses à venir.

En d'autres termes, les états conscients découleraient de la façon dont l'algorithme de l'espace de travail traite les entrées sensorielles, les sorties motrices et les variables internes liées à la mémoire, à la motivation et aux attentes. La conscience, ce serait ce traitement global. Pour les tenants de cette théorie, la possibilité d'une conscience artificielle n'est pas exclue.

LA CONDITION DE LA CONSCIENCE

La seconde hypothèse repose sur la théorie de l'information intégrée (IIT, pour *integrated information theory*), qui adopte une approche plus fondamentale pour expliquer la conscience. Giulio Tononi, psychiatre et neuroscientifique à l'université du Wisconsin-Madison, en est le promoteur principal, avec d'autres collaborateurs, dont moi-même. Le principe est de partir de la nature de l'expérience consciente et d'en déduire les propriétés du système physique qui en constitue le support. Par exemple, chaque expérience forme un tout unique – lorsque je vois une image, je n'ai pas deux expériences conscientes juxtaposées, une de la partie gauche de mon champ visuel et une de la partie droite, mais une seule – et cela doit se traduire d'une façon ou d'une autre dans le système physique sous-jacent.

© Warner Bros./Archive Photos/Getty Images

Que peut une machine consciente ? La S. F. (ici, l'androïde Zhora, de *Blade Runner*) se régale de cette question.

En particulier, l'information doit y être traitée de façon intégrée, c'est-à-dire que les éventuels sous-systèmes doivent interagir et s'influencer mutuellement. On parle de « pouvoir causal intrinsèque » et on quantifie l'importance de ces influences mutuelles par une mesure mathématique que nous appelons l'« information intégrée ».

Ainsi, le pouvoir causal intrinsèque n'est pas une notion floue et éthérée, il est susceptible d'être évalué avec précision. On peut le déterminer pour tout type de mécanisme, par exemple un neurone qui émet des potentiels d'action affectant les cellules connectées en aval (*via* des synapses), ou un circuit électronique

La théorie de l'information intégrée a inspiré la construction d'un appareil pour mesurer le niveau de conscience chez des patients incapables de communiquer. Il est ici testé par l'équipe de Steven Laureys, au Coma Science Group de l'université et du CHU de Liège.

composé de transistors, de résistances et de fils. Il est d'autant plus élevé qu'un système est capable d'influencer, à partir de son état actuel, ses propres valeurs d'entrée et de sortie. L'état de ce système est alors marqué par son passé et porteur de son avenir.

Notre théorie stipule que tout mécanisme doté d'un tel pouvoir causal est conscient. Plus un système a une valeur importante d'information intégrée – représentée par la lettre grecque Φ (prononcée « fi ») –, plus il est conscient. En revanche, s'il n'a aucun pouvoir causal intrinsèque, son Φ est nul et il ne ressent rien.

Au sein du cortex, la quantité d'information intégrée est énorme, car on y trouve un vaste réseau de neurones hétérogènes et étroitement interconnectés (qui s'influencent donc mutuellement). La théorie a inspiré la construction d'un appareil de mesure de la conscience, en cours d'évaluation clinique *(voir la photo ci-dessus)*. Cet instrument serait précieux pour déterminer si certains patients sont conscients mais incapables de communiquer, ou totalement inconscients – par exemple dans les états végétatifs persistants ou l'anesthésie.

UN RÉSEAU ZOMBIE NE RESSENT RIEN

Pour les ordinateurs programmables, en revanche, la situation est différente, comme l'ont montré des travaux qui ont analysé leur structure à l'échelle des composants métalliques – les transistors, fils et diodes qui servent de substrat physique à tout calcul. Leur pouvoir causal intrinsèque et leur Φ sont infimes, notamment parce que leur connectique est très différente de celle du cerveau : chaque composant n'est connecté qu'à quelques autres (là où un seul neurone forme, en moyenne, 10 000 synapses), et il y a bien moins de rétroactions que dans un système neuronal. Cela reste vrai quel que soit le logiciel qui tourne sur le processeur : qu'il calcule les impôts ou simule le cerveau, cela ne change rien.

Le point clé est que deux réseaux qui effectuent la même opération d'entrée-sortie mais dont les circuits sont configurés différemment sont susceptibles de posséder des quantités différentes d'information intégrée. La grandeur Φ peut être très faible pour un circuit et très élevée pour l'autre. Bien qu'ils soient identiques de l'extérieur, l'un des réseaux expérimente quelque chose, tandis que son homologue zombie ne ressent rien. La différence se situe sous le capot, dans le câblage interne. En un mot, la conscience résulte de ce qu'on *est*, et non de ce qu'on *fait*.

La théorie de l'espace de travail neuronal global et la théorie de l'information intégrée ont certaines conséquences drastiquement opposées. Imaginons par exemple que l'on simule le connectome – le câblage synaptique exact de tout le système nerveux. Les anatomistes ont déjà cartographié les connectomes de quelques vers. Ils travaillent actuellement sur celui de la drosophile, la petite mouche qui tourne autour des fruits trop mûrs, et prévoient de s'attaquer à la souris au cours de la prochaine décennie. Supposons qu'à l'avenir il soit possible de scanner un cerveau humain entier, avec ses près de 100 milliards de neurones et ses 100 000 milliards de synapses, après la mort de son propriétaire, puis de simuler son fonctionnement sur un ordinateur avancé, peut-être une machine quantique. Si le modèle

© Coma Science group

est suffisamment fidèle, cette simulation se réveillera et se comportera comme un simulacre numérique de la personne décédée. Ainsi, elle parlera et accédera à ses souvenirs, à ses envies, à ses peurs et aux autres traits de sa personnalité. Mais sera-t-elle consciente ?

Si l'imitation du fonctionnement du cerveau est tout ce qui compte pour créer la conscience, comme cela est postulé par la théorie de l'espace de travail neuronal global, la personne simulée sera consciente, réincarnée dans un ordinateur. Ce qui n'est pas sans rappeler l'un des fantasmes favoris de la science-fiction : télécharger son connectome dans le *cloud* pour continuer à vivre dans l'au-delà numérique…

Mais la théorie de l'information intégrée propose une interprétation radicalement différente de cette situation : le simulacre ressentira la même chose que n'importe quel logiciel basique – c'est-à-dire rien. Il agira comme une personne, mais sans aucun sentiment. Il ne sera qu'une sorte de zombie (sans le moindre goût pour la chair humaine, rassurez-vous).

Selon cette théorie, les pouvoirs causaux intrinsèques du cerveau sont nécessaires pour créer la conscience. Et ils ne peuvent être simulés : ils doivent faire partie intégrante du système physique sous-jacent.

C'est un peu comme si vous simuliez une pluie diluvienne ou un trou noir sur un ordinateur : vous ne risquez ni d'être mouillé ni d'être avalé par une énorme force gravitationnelle. Tout simplement parce qu'une simulation n'a pas le pouvoir causal de condenser la vapeur d'eau ou de courber l'espace-temps.

VOTRE PC AURA-T-IL DES DROITS ?

Il n'y a toutefois rien de magique dans le cerveau humain et, en principe, il serait possible de créer une conscience équivalente à la nôtre en allant au-delà d'une simple simulation. Il faudrait alors construire du matériel dit « neuromorphe », fondé sur une architecture similaire à celle du système nerveux.

L'éventuel ressenti d'un connectome simulé n'est pas la seule différence entre les deux théories. Celles-ci ne localisent pas au même endroit le substrat physique de certaines expériences conscientes spécifiques – l'épicentre de ce substrat étant à l'arrière du cortex pour l'une, à l'avant pour l'autre. Cette prédiction et d'autres sont actuellement testées dans le cadre d'un projet pilote à grande échelle, impliquant six laboratoires aux États-Unis, en Europe et en Chine, et qui a reçu un financement de 4,5 millions d'euros de la fondation Templeton World Charity.

La question d'une éventuelle sensibilité des machines importe aussi pour des raisons éthiques. Si les ordinateurs font l'expérience de la vie par leurs propres sens, ils cessent d'être un simple moyen mis au service d'un but déterminé par les humains. Ils doivent être respectés pour eux-mêmes.

Selon la théorie de l'espace de travail global, les machines intelligentes pourraient bien passer de simples objets à sujets à part entière – chacune existerait comme un « je », doté de son propre point de vue. Une idée que l'on retrouve dans les séries télévisées *Black Mirror* et *Westworld*. Et lorsque les ordinateurs auront atteint des capacités cognitives rivalisant avec celles de l'humanité, ils feront valoir leurs droits juridiques et politiques – le droit de ne pas être jetés, de ne pas voir leur mémoire effacée, d'éviter la douleur et la dégradation. L'autre théorie, incarnée par la notion d'information intégrée, est qu'aussi perfectionnés soient-ils, les ordinateurs resteront des coquilles vides, semblables à des fantômes. Il leur manquera ce à quoi nous attachons le plus de valeur : la sensation de vivre.

— L'auteur —

> Christof Koch est directeur scientifique et président de l'institut Allen pour les sciences du cerveau, à Seattle, aux États-Unis.

— À lire —

> M. Lee *et al.*, Quantifying arousal and awareness in altered states of consciousness using interpretable deep learning, *Nature Communication*, 2022.

> C. Koch, The feeling of Life itself : Why Consciousness is Widespread but Can't Be Computed, *MIT Press*, 2019.

> L. Naccache, Observer la conscience, *Pour la Science*, n° 500, pp. 75-80, juin 2019.

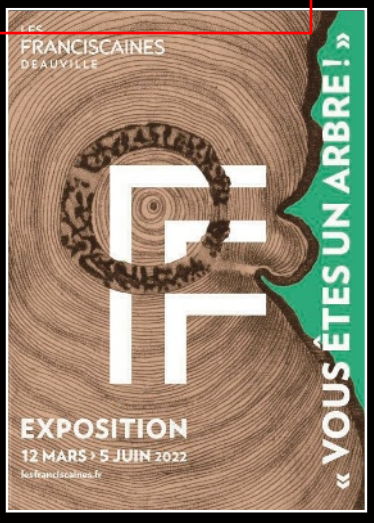
> S. Dehaene *et al.*, What is consciousness, and could machines have it ?, *Science*, vol. 358, pp. 486-492, 2017.

À VISITER

L'arbre dans tous ses états

Une exposition aux Franciscaines, à Deauville, fait le pari, par la diversité des artistes présentés et des idées abordées, de nous faire changer de regard sur les arbres… En fait, de nous inviter à les regarder vraiment.

Dans *Les Paradis artificiels*, Baudelaire l'assure « Et bientôt vous êtes l'arbre ». Cette promesse de devenir un arbre est aussi celle de l'exposition qui se tient aux Franciscaines, un lieu ouvert en mai 2021, à Deauville, dans un ancien couvent du XIXe siècle. Là se tiennent désormais une médiathèque, une salle de spectacle, un musée… pour une expérience culturelle décloisonnée et transversale dans un lieu à l'architecture novatrice.
L'exposition est ambitieuse, car elle s'attaque à une très ancienne question : « Qu'est-ce qu'un arbre ? » La réponse est multiple, scientifique et artistique, car, comme on s'en rend compte à mesure du parcours, l'arbre est l'image du tout quand il est généalogique, phylogénétique ; l'arbre devient humain comme dans le mythe de Daphné ou lorsqu'il s'agit de décrire le système veineux ou bronchique ; l'arbre se fait sujet notamment dans la peinture de paysage du XIXe siècle qui en fait un être à part entière ; l'arbre est inspiration et même matériau pour les artistes contemporains. À une telle multiplicité, une telle identité plurielle, répond la diversité d'artistes présentés : Raymond Lulle, Pierre Bonnard, Millet, Henri Matisse, Man Ray, Giuseppe Penone, Fabrice Hyber, Sebastião Salgado, Eva Jospin… En ces temps où l'arbre tient une place privilégiée dans notre rapport à la nature, au point de nous aider à surmonter notre anxiété face aux dérèglements du monde, il importe de mieux connaître cet allié, cet ami. Pas de doute, « il y a de l'humain dans l'arbre, comme il y a de l'arbre dans l'humain ».



« Vous êtes un arbre ! »,
Les Franciscaines,
à Deauville, jusqu'au 5 juin 2022.
https://bit.ly/3JdOQCn

À VOIR

La tête dans le sable

Comment réagissons-nous face au dérèglement climatique ? « En faisant l'autruche ! » affirme Tali Sharot, de l'University College, à Londres, dans le documentaire *Climat : mon cerveau fait l'autruche*, de Raphaël Hitier et Sylvie Deleule, diffusé sur le site d'Arte jusqu'au 12 mai 2022. Avec une dizaine d'autres chercheurs, la neurobiologiste décortique les mécanismes qui expliquent notre inertie face au danger : biais cognitifs, effet spectateur, anatomie profonde du cerveau… Sommes-nous alors irrémédiablement condamnés par notre fonctionnement ? Non, le film aborde également des techniques qui peuvent nous aider à surpasser ces obstacles pour finalement agir !

bit.ly/36hpfcO

À ÉCOUTER

Qu'en est-il ?

Quid, c'est le nom de l'ancêtre d'internet, un gros livre qui avait réponse à tout, mais c'est aussi le nom d'un podcast conçu par le Conservatoire national des arts et métiers. Au fil des épisodes, la journaliste Pauline Grisoni converse avec les enseignants-chercheurs de l'institution à propos de sujets divers relevant de la culture, de la société, de l'environnement… Les premiers épisodes portent notamment sur la santé intégrative (les thérapies conventionnelles et non conventionnelles s'allient dans une approche globale de la santé) et les nouveaux modes de travail hybride.

https://bit.ly/3j56L2b

À LIRE

Le vivant comme sujet

Les travaux et les idées de la primatologie japonaise, selon laquelle le sujet de l'évolution est la société que chaque espèce forme avec son milieu, se sont imposés partout dans le monde.

Au début des années 1950, sur la presqu'île de Koshima, au Japon, des primatologues observent des macaques lavant des patates douces dans une rivière puis dans la mer. Ce comportement, acquis, est enseigné aux jeunes générations. C'est la première fois que l'on met en évidence des cultures dans le monde animal, et parmi les scientifiques figure Kinji Imanishi. De ces travaux émergera une nouvelle façon d'envisager l'évolution qui ne reposera plus seulement sur le hasard des mutations et la nécessité de la sélection naturelle. Il s'agit désormais de prendre aussi en compte l'individu, son comportement, son esprit d'initiative, ses liens avec son milieu et ses congénères… À la vision uniquement « compétitive » dans laquelle on a réduit les idées de Darwin s'ajoute une conception où synergie et coopération ont une place notablement importante, comme l'avait déjà remarqué au début du XXe siècle l'anarchiste Kropotkine, et qui a prospéré dans une culture où le dualisme occidental homme/animal n'a pas cours et où les êtres humains ne sont pas les seuls à avoir une âme. C'est tout ce bouleversement intellectuel que raconte Kinji

Imanishi dans cet ouvrage écrit à la fin de sa vie et dans lequel le primatologue Frans de Waal voit un changement de paradigme crucial dans la façon d'aborder le vivant en général et les espèces animales en particulier. Selon lui, les pratiques occidentales ont peu à peu été perfusées par les idées venues du Japon, et elles y ont nettement gagné. Compliment suprême, aux yeux de Frans de Waal, Kinji Imanishi est le « Stephen Jay Gould » du Japon.

La Liberté dans l'évolution, Kinji Imanishi, Wildproject, 2022, 192 pages, 20 euros.

À VISITER

Là où va l'Indien

Que sait-on des Sioux à part qu'ils sont « rusés » et ont eu pour chef le célèbre Sitting Bull (de son vrai nom Tȟatȟáŋka Íyotake)? Pas grand-chose en vérité! L'exposition « Sur la piste des Sioux », proposée au musée des Confluences, à Lyon, corrige cette lacune. Elle met à mal quelques clichés sur les populations natives du continent nord-américain en déroulant le fil de la longue construction iconographique associée à ces Amérindiens depuis les premiers contacts avec les Européens il y a quelque cinq cents ans. On comptait alors des centaines de nations indiennes et près de mille langues parlées. Au cours des siècles, les stéréotypes se sont succédé, du « sauvage primaire » des débuts jusqu'au « sage en contact avec les forces de la nature » actuel en passant par le guerrier haut en couleur mythifié et mis en scène par Buffalo Bill, l'indigène sanguinaire des westerns et le Bison futé de nos départs en vacances. Au sortir de l'exposition, le visiteur a comblé le fossé, énorme, entre l'image véhiculée autour des Indiens et leur réalité, et notamment les difficultés auxquelles ils sont confrontés aujourd'hui : perte d'identité, disparition de leur culture, accaparement des territoires… Hugh!

« Sur la piste des Sioux », musée des Confluences, à Lyon, jusqu'au 28 août 2022 https://bit.ly/3MQ3KAL

Pour la Science HORS-SÉRIE

GROUPE POUR LA SCIENCE
Directrice des rédactions : Cécile Lestienne

HORS-SÉRIE POUR LA SCIENCE
Rédacteur en chef adjoint : Loïc Mangin
Rédacteur en chef adjoint délégué : Olivier Voizeux
Maquettiste : Ghislaine Salmon-Legagneur

POUR LA SCIENCE
Rédacteur en chef : Maurice Mashaal
Rédactrice en chef adjointe : Marie-Neige Cordonnier
Rédacteurs : François Savatier, Sean Bailly

Développement numérique : Philippe Ribeau-Gésippe
Community manager et partenariats : Aëla Keryhuel
aela.keryhuel@pourlascience.fr

Conception graphique : Céline Lapert et Ingrid Leroy
Direction artistique : Céline Lapert
Maquette : Pauline Bilbault, Raphaël Queruel, Ingrid Leroy et Ingrid Lhande
Réviseuse : Anne-Rozenn Joublé
Marketing & diffusion : Charline Buché
Cheffe de produit : Eléna Delanne
Direction du personnel : Olivia Le Prévost
Secrétaire général : Nicolas Bréon
Fabrication : Marianne Sigogne et Zoé Farré-Vilalta
Directeur de la publication et gérant : Frédéric Mériot
Ont contribué à ce numéro : Isabelle Bouchery, Maud Bruguière, William Rowe-Pirra

WWW.POURLASCIENCE.FR
170 bis bd du Montparnasse
75014 Paris
Tél. : 01 55 42 84 00

PUBLICITÉ FRANCE
stephanie.jullien@pourlascience.fr

ABONNEMENTS
Abonnement en ligne :
https://boutique.groupepourlascience.fr
Courriel : serviceclients@groupepourlascience.fr
Tél. : 01 86 70 01 76
Du lundi au vendredi de 9 h à 13 h
Adresse postale :
Service Abonnement - Groupe Pour la Science
235 avenue Le Jour se Lève
92100 Boulogne-Billancourt

Tarifs d'abonnement 1 an (12 numéros)
France métropolitaine : 59 euros
Europe : 71 euros
Reste du monde : 85,25 euros

DIFFUSION
Contact kiosques :
À Juste Titres ; Alicia Abadie
Tél. : 04 88 15 12 47
Information/modification de service/réassort :
www.direct-editeurs.fr

SCIENTIFIC AMERICAN
Editor in chief : Laura Helmuth
Acting President : Stephen Pincock
Executive Vice President : Michael Florek

Toutes demandes d'autorisation de reproduire, pour le public français ou francophone, les textes, les photos, les dessins ou les documents contenus dans la revue « Pour la Science », dans la revue « Scientific American », dans les livres édités par « Pour la Science » doivent être adressées par écrit à « Pour la Science S.A.R.L. », 170 bis bd du Montparnasse, 75014 Paris.
© Pour la Science S.A.R.L. Tous droits de reproduction, de traduction, d'adaptation et de représentation réservés pour tous les pays.
La marque et le nom commercial « Scientific American » sont la propriété de Scientific American, Inc. Licence accordée à « Pour la Science S.A.R.L. ».
En application de la loi du 11 mars 1957, il est interdit de reproduire intégralement ou partiellement la présente revue sans autorisation de l'éditeur ou du Centre français de l'exploitation du droit de copie (20 rue des Grands-Augustins, 75006 Paris).

Origine du papier : Finlande • Taux de fibres recyclées : 0 % • « Eutrophisation » ou « Impact sur l'eau » : P_{tot} 0,003 kg/t

PEFC 10-31-1282 / Certifié PEFC / Ce produit est issu de forêts gérées durablement et de sources contrôlées. / pefc-france.org

Imprimé en France – Maury imprimeur S.A. Malesherbes – N° d'imprimeur : 260298 – N° d'édition : M0770715-01 – Dépôt légal : mai 2022.
Commission paritaire n° 0922K82079 du 19-09-02 – Distribution : MLP – ISSN 1 246-7685 – Directeur de la publication et gérant : Frédéric Mériot.

Demain, un robot président ? Faut-il faire confiance à l'Intelligence artificielle qui est aujourd'hui partout en France ?

Livre numérique disponible sur toutes les plateformes et librairies en ligne

Éditions de L'Observatoire

Institut du Cerveau ICM

BRAINCAST

La voix des neurones

Le podcast de *Cerveau & Psycho*

en partenariat avec l'Institut du Cerveau

9ème épisode

Réparer les neurones :
l'ultime défi de la sclérose en plaques

www.cerveauetpsycho.fr/sr/braincast/

avec la Pr Catherine Lubetzki

interviewée par Sébastien Bohler

Neurologue
Directrice médicale
de l'Institut du cerveau

Cerveau & Psycho