

Les cinq murs de l'intelligence artificielle

Bertrand Braunschweig

DANS **FUTURIBLES** 2023/2 (N° 453), PAGES 5 À 24

ÉDITIONS **FUTURIBLES**

ISSN 0337-307X

DOI 10.3917/futur.453.0005

Article disponible en ligne à l'adresse

<https://www.cairn.info/revue-futuribles-2023-2-page-5.htm>



CAIRN.INFO
MATIÈRES À RÉFLEXION

Découvrir le sommaire de ce numéro, suivre la revue par email, s'abonner...

Flashez ce QR Code pour accéder à la page de ce numéro sur Cairn.info.



Distribution électronique Cairn.info pour Futuribles.

La reproduction ou représentation de cet article, notamment par photocopie, n'est autorisée que dans les limites des conditions générales d'utilisation du site ou, le cas échéant, des conditions générales de la licence souscrite par votre établissement. Toute autre reproduction ou représentation, en tout ou partie, sous quelque forme et de quelque manière que ce soit, est interdite sauf accord préalable et écrit de l'éditeur, en dehors des cas prévus par la législation en vigueur en France. Il est précisé que son stockage dans une base de données est également interdit.

Les cinq murs de l'intelligence artificielle

Par Bertrand Braunschweig ¹

Le lancement, fin novembre 2022, de ChatGPT, un agent conversationnel développé par OpenAI, a fait grand bruit et suscité nombre de reportages et commentaires sur les performances et avancées de l'intelligence artificielle (IA). Pour autant, si les progrès en la matière sont effectivement impressionnants, un certain nombre de limites demeurent, et l'avènement d'une IA totalement autonome et parfaite relève encore de la science-fiction. Ainsi Bertrand Braunschweig, responsable de la coordination scientifique du programme français Confiance.ai, pointe-t-il ici cinq grands « murs » sur lesquels l'IA pourrait bien buter si l'on ne s'y intéresse pas suffisamment. Il présente ces cinq murs, qui sont celui de la confiance envers l'IA, celui de sa consommation énergétique, celui de la sécurité des systèmes qu'elle régit, celui des interactions homme-machine, et enfin celui de l'inhumanité des machines. Puis il propose, en fin d'article, un certain nombre de pistes de travail et de recherche pour répondre aux défis posés par ces cinq murs : amélioration des architectures de réseaux, combinaison de modèles numériques et symboliques, interdisciplinarité renforcée, etc. S.D.

L'intelligence artificielle (IA) progresse à un rythme très rapide tant sur le plan de la recherche que sur celui des applications et pose des questions de société auxquelles toutes les réponses sont loin d'être données. Mais en avançant rapidement, elle fonce sur ce que j'appelle les cinq murs de l'IA, des murs sur lesquels elle est susceptible de se fracasser si l'on ne prend pas de précautions. N'importe lequel de ces cinq murs est en mesure de mettre un terme à sa progression, c'est pour cette raison qu'il est essentiel

1. Coordonnateur scientifique du programme Confiance.ai, programme industriel et académique qui constitue un des trois piliers du Grand Défi sur l'intelligence artificielle prouvable et certifiable. Une version longue de cet article, rédigée courant 2021, a été publiée en six épisodes sur le *blog Binaire* du journal *Le Monde* entre mars et avril 2022 (URL du premier épisode : <https://www.lemonde.fr/blog/binaire/2022/03/01/les-cinq-murs-de-lia/>). Consulté le 25 janvier 2023 ; les épisodes suivants sont sur le même site). Cet article en reprend les principales idées.

d'en connaître la nature et de chercher à apporter des réponses afin d'éviter un troisième hiver de l'IA, qui ferait suite aux deux premiers connus pendant les années 1970 et 1990, au cours desquels la recherche et le développement de l'IA se sont quasiment arrêtés faute de budget et d'intérêt de la communauté. Ces cinq murs sont ceux de la confiance, de l'énergie, de la sécurité, de l'interaction avec les humains et de l'inhumanité. Ils contiennent chacun un certain nombre de ramifications et sont bien évidemment en interaction.

Ce texte se veut un outil de réflexion pour le lecteur, il est destiné à susciter des commentaires et réactions, que ce soit sur la réalité de ces murs, sur la complétude de mon analyse, ou sur la manière d'échapper à l'écrasement sur l'un de ces murs. Je précise cependant qu'il y a d'autres facteurs, non technologiques, qui mettent en cause l'avenir de l'IA et que je ne traite pas ici. Ainsi, par exemple, la pénurie de chercheurs, ingénieurs, techniciens capables de développer et de mettre en œuvre les technologies d'IA est très bien identifiée, elle se matérialise par les salaires élevés qui sont versés à celles et ceux qui affichent l'IA comme spécialité, et à la mise en place de nombreux programmes de formation qui, à terme, devraient permettre de revenir à une situation normale en la matière, l'offre rejoignant la demande. Il ne manque pas non plus de démarches gouvernementales, collectives, associatives et autres pour réglementer et gouverner l'IA² ; je n'aborderai pas ces aspects ici. Je recommande plutôt de s'intéresser aux travaux du Partenariat mondial sur l'intelligence artificielle (GPAI, *Global Partnership on Artificial Intelligence*, en anglais) qui rassemble nombre d'experts de disciplines, d'origines et de cultures différentes sur les sujets de société gravitant autour de l'IA.

Je reconnais également qu'il y a des avis différents à ce sujet. Le rapport d'étape 2021 de l'étude longitudinale sur 100 ans de Stanford³ examine les avancées de l'IA à ce jour et présente des défis pour le futur, très complémentaires de ceux que j'aborde ici. L'ouvrage *Human Compatible*⁴ de Stuart Russell, professeur à Berkeley et auteur du principal livre de cours sur l'IA depuis deux décen-

2. Notamment la proposition européenne en faveur d'un *Artificial Intelligence (AI) Act*, dont la version initiale a été publiée en avril 2021, et sa contrepartie *AI Liability Directive* de fin 2022, ou encore la proposition américaine pour un *AI Bill of Rights*, datant également de fin 2022.

3. LITTMAN Michael L. et alii, *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report*, Stanford University, septembre 2021. URL : https://ai100.stanford.edu/sites/g/files/sbiybj18871/files/media/file/AI100Report_MT_10.pdf. Consulté le 25 janvier 2023.

4. RUSSELL Stuart J., *Human Compatible: Artificial Intelligence and the Problem of Control*, New York : Viking, 2019.

nies, s'intéresse à la compatibilité entre machines et humains, sujet que je traite différemment quand je parle du mur de l'interaction.

La confiance

Si les personnes n'ont pas confiance envers les systèmes d'IA avec lesquels elles interagissent, elles les rejeteront. Il y a largement de quoi causer un troisième hiver de l'IA ! La confiance est une notion riche et multifactorielle : beaucoup de sociologues et de technologues se sont intéressés aux mécanismes de son établissement. Plusieurs organismes tentent de fournir des définitions de ce qu'est la confiance envers les systèmes d'intelligence artificielle ; ceci a été le sujet principal du groupe d'experts mobilisés par la Commission européenne (dont tous les travaux se font dans l'optique « *trustworthy AI* », c'est-à-dire d'une IA digne de confiance ⁵).

Je résumerai ici en disant que la confiance, en particulier envers les artefacts numériques dont l'IA fait partie, est une combinaison de facteurs technologiques et sociologiques. Technologiques, comme la capacité de vérifier la justesse d'une conclusion, la robustesse à des perturbations, le traitement de l'incertitude, etc. Sociologiques, comme la validation par des pairs, la réputation dans les réseaux sociaux, l'attribution d'un label par un tiers de confiance, etc. Les questions d'interaction avec les utilisateurs sont intermédiaires entre ces deux types de facteurs : transparence, explicabilité, qualité des interactions de manière plus générale.

Les facteurs sociologiques ne sont pas propres à l'IA : dans un réseau de confiance entre humains, la transmission de la confiance ne fait pas nécessairement appel aux facteurs technologiques. En revanche, la base technologique de la confiance en l'IA est bien spécifique et pose de nombreux défis : on ne sait pas bien, aujourd'hui, prouver que les conclusions d'un système entraîné par apprentissage sur une base de données sont les bonnes, qu'elles sont robustes à des petites variations, qu'elles ne sont pas entachées de biais, etc. Il existe de nombreux programmes de R&D à ce sujet, dont un des plus importants est l'initiative *Confiance.ai* ⁶, centrée

5. « Ethics Guidelines for Trustworthy AI » ; « Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-assessment » Commission européenne, respectivement 8 avril 2019. URL : <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> ; et 17 juillet 2020. URL : <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>. Consultés le 27 janvier 2023.

6. Site Internet <http://www.confiance.ai/>. Voir le livre blanc *Towards the Engineering of Trustworthy AI Applications for Critical Systems: The Confidence.ai Program*, octobre 2022. URL : <https://www.confiance.ai/>

sur les systèmes stratégiques (transport, défense, énergie, industrie) et portée par de grands groupes industriels dans le cadre du Grand Défi sur la fiabilisation et la certification de l'IA.

Tant que cette question restera ouverte, le risque pour l'IA de se heurter au mur de la confiance sera majeur. Il le sera encore plus pour les systèmes à risques (au sens de la Commission européenne dans sa proposition de réglementation de l'IA ⁷).

Le mur de l'énergie

Les chercheurs de l'université de Stanford publient annuellement l'édition de l'*AI Index*, qui mesure la progression des technologies d'IA dans le monde. L'édition de fin 2019 présentait pour la première fois l'évolution des besoins de calcul des applications de l'IA qui ont suivi la loi de Moore (doublement tous les 18 mois) de 1960 à 2012. Depuis, ces besoins doublent tous les trois mois et demi ! La demande du plus gros système d'IA connu à l'époque (et qui a donc doublé plusieurs fois depuis) était de 1 860 pétaFLOPS*jours (un pétaFLOPS pendant un jour) ⁸, soit plus de deux mois de calcul s'il utilisait la totalité de la machine Jean Zay (supercalculateur pour l'IA et le calcul scientifique mis à disposition des chercheurs français et européens depuis 2020), et une consommation électrique de près de trois gigawattheures. Pis encore, si le rythme actuel se poursuit, la demande sera encore multipliée par un facteur 1 000 dans trois ans, et un million dans six ans !

Le mur de l'énergie est bien identifié par certains chercheurs en apprentissage profond. L'article fondateur d'Emma Strubell *et alii* ⁹ établissait que l'entraînement d'un grand réseau de neurones de traitement de la langue naturelle de type « *transformer* », avec optimisation de l'architecture du réseau, consommait autant d'énergie que cinq voitures particulières pendant toute leur durée de vie (tableau 1).

[ai/wp-content/uploads/2022/10/LivreBlanc-Confiance.ai-Octobre2022.pdf](#). Consulté le 27 janvier 2023.

7. « Règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'Union », Commission européenne, COM(2021) 206 final, 21 avril 2021. URL : <https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:52021PC0206&from=EN>. Consulté le 27 janvier 2023.

8. FLOPS : *F*loating-*p*oint *O*perations *P*er *S*econd / opérations en virgule flottante par seconde. Péta-FLOPS = 10¹⁵ FLOPS.

9. STRUBELL Emma, GANESH Ananya et MCCALLUM Andrew, « Energy and Policy Considerations for Deep Learning in NLP », ArXiv, Cornell University, juin 2019. URL : <https://arxiv.org/pdf/1906.02243.pdf>. Consulté le 27 janvier 2023.

Tableau 1. Estimation des émissions de CO₂ issues de l'entraînement d'un modèle courant d'IA de traitement du langage naturel (NLP), comparées à celles d'autres consommations familières (en livres* d'équivalent CO₂)

Vol New York-San Francisco, 1 passager	1 984
Une année de vie humaine (moyenne)	11 023
Une année de vie humaine d'un Américain (moyenne)	36 156
Véhicule particulier à moteur thermique, sur toute sa durée de vie	126 000
Modèle d'entraînement NLP - pipeline (analyse, SRL**)	39
<i>Avec réglage et expérimentation</i>	<i>78 468</i>
Gros modèle d'entraînement <i>Transformer</i>	192
<i>Avec optimisation de l'architecture du réseau neuronal</i>	<i>626 155</i>

*Une livre équivaut à 453,6 grammes.

***Safe Reinforcement Learning* (apprentissage par renforcement).

Source : STRUBELL Emma, GANESH Ananya et MCCALLUM Andrew, *op. cit.*

L'article de Neil Thompson *et alii*¹⁰ allait plus loin en concluant que « les limites de calcul de l'apprentissage profond seront bientôt contraignantes pour toute une série d'applications, ce qui rendra impossible l'atteinte d'importantes étapes de référence si les trajectoires actuelles se maintiennent ». Et ceci est souligné par les chiffres donnés par l'*AI Index* qui insiste sur le facteur exponentiel correspondant. Fin 2021, Neil Thomson *et alii* ont complété cette analyse¹¹ sur l'exemple du traitement d'images (ImageNet) et abouti à estimer à 9 ce facteur entre la réduction du taux d'erreur et le besoin en calcul et données, ce qui signifie qu'une division par 2 du taux d'erreur nécessite 500 fois plus de calcul, et une division par 4 demanderait 250 000 fois plus.

Le mur de la consommation énergétique liée aux besoins de calcul intensif des applications de l'IA, qui sont basées sur l'apprentissage profond et consommatrices de très grandes quantités de données, en arrêtera inévitablement la croissance exponentielle, à terme relativement rapproché, si l'on ne fait rien pour y remédier.

10. THOMPSON Neil C. *et alii*, « The Computational Limits of Deep Learning », ArXiv, Cornell University, juillet 2020, mis à jour en juillet 2022. URL : <https://arxiv.org/pdf/2007.05558.pdf>. Consulté le 27 janvier 2023.

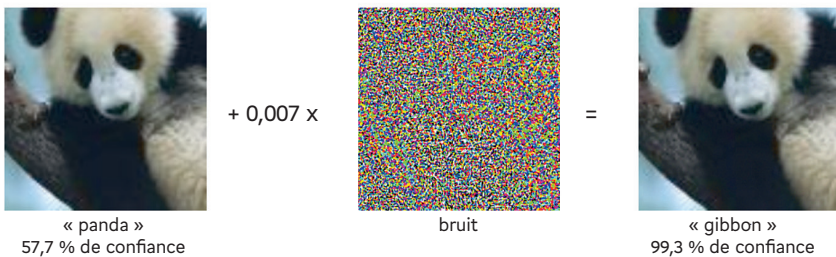
11. THOMPSON Neil C. *et alii*, « Deep Learning's Diminishing Returns: The Cost of Improvement Is Becoming Unsustainable », *IEEE Spectrum*, 24 septembre 2021. URL : <https://spectrum.ieee.org/deep-learning-computational-cost>. Consulté le 27 janvier 2023.

Le mur de la sécurité

Les questions de sécurité des systèmes d'information ne sont pas propres à l'IA, mais les systèmes d'IA ont certaines particularités qui les rendent sensibles à des problèmes de sécurité d'un autre genre, et tout aussi importants. Si les systèmes d'IA sont, comme tous les systèmes numériques, susceptibles d'être attaqués, piratés, compromis par des méthodes « usuelles » (intrusion, déchiffrement, virus, saturation, etc.), ils possèdent des caractéristiques particulières qui les rendent particulièrement fragiles à d'autres types d'attaques plus spécifiques. Les attaques antagonistes ou adverses (*adversarial attacks* en anglais) consistent à injecter des variations mineures des données d'entrée, lors de la phase d'inférence, afin de modifier de manière significative la sortie du système. Les premières références d'attaques adverses sont le célèbre exemple du panneau STOP non reconnu lorsqu'il est tagué par des étiquettes, et celui du panda confondu avec un gibbon suite à l'ajout d'une faible composante de bruit (voir illustrations ci-dessous) ; on sait qu'il est assez facile de composer une attaque destinée à modifier très fortement l'interprétation des données faite par un réseau de neurones. Et cela ne concerne pas que les images : on peut concevoir des attaques antagonistes sur du signal temporel (audio en particulier), sur



Exemple de génération rapide d'illustrations contradictoires



Lecture : l'IA du moteur de recherche GoogleNet identifie un panda avec 57,7 % de confiance sur la première image. L'ajout d'un « bruit » quasi imperceptible (0,007) modifie le résultat : c'est un gibbon qui est identifié avec 99,3 % de confiance.

Source : GOODFELLOW Ian J., SHLENS Jonathon et SZEGEDY Christian, « Explaining and Harnessing Adversarial Examples », ArXiv, Cornell University, décembre 2014. URL : <https://arxiv.org/pdf/1412.6572.pdf>. Consulté le 31 janvier 2023.

du texte, etc. Les conséquences d'une telle attaque peuvent être dramatiques, une mauvaise interprétation des données d'entrée peut conduire à une prise de décision dans le mauvais sens (par exemple, accélérer au lieu de s'arrêter, pour une voiture). Le rapport du NIST sur le sujet ¹² établit une intéressante taxonomie des attaques et défenses correspondantes. Il montre notamment que les attaques en phase d'inférence ne sont pas les seules qui font souci. Il est notamment possible de polluer les bases d'apprentissage avec des exemples antagonistes, ce qui compromet naturellement les systèmes entraînés à partir de ces bases. Bien évidemment, la communauté de recherche en intelligence artificielle s'est saisie de la question et les travaux sur la détection des attaques antagonistes sont nombreux.

Toujours est-il que des accidents — aujourd'hui inévitables — sur des systèmes à risques ou stratégiques, causés par ces questions de sécurité, auront des conséquences extrêmement néfastes sur le développement de l'intelligence artificielle.

Un deuxième point d'attention est la question du respect de la vie privée. Cette question prend une dimension particulière avec les systèmes d'IA qui ont une grande capacité à révéler des données confidentielles de manière non désirée : par exemple retrouver les images individuelles d'une base d'entraînement dans les paramètres d'un réseau de neurones, ou opérer des recoupements sur diverses sources pour en déduire des informations sur une personne. Ces questions sont notamment à l'origine des travaux en apprentissage réparti (*federated learning*) ¹³ dont le but est de réaliser un apprentissage global à partir de sources multiples réparties sur le réseau pour composer un modèle unique contenant, d'une certaine manière, une compression de toutes les données réparties mais sans pouvoir en retrouver l'origine.

Pour approfondir le sujet de la cybersécurité de l'apprentissage machine, une bonne référence est le rapport de l'agence européenne de sécurité informatique (ENISA) sur le sujet ¹⁴. Si l'on y ajoute les

12. TABASSI Elham *et alii*, « A Taxonomy and Terminology of Adversarial Machine Learning », Draft NISTIR (National Institute of Standards and Technology - *Internal Report*), n° 8269, octobre 2019. URL : <https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf>. Consulté le 27 janvier 2023.

13. KAIROUZET Peter *et alii*, « Advances and Open Problems in Federated Learning », ArXiv, Cornell University, décembre 2019 (mis à jour en mars 2021). URL : <https://arxiv.org/abs/1912.04977>. Consulté le 27 janvier 2023.

14. ENISA (European Union Agency for Cybersecurity), *Securing Machine Learning Algorithms*, ENISA, décembre 2021. URL : <https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms/@download/fullReport>. Consulté le 27 janvier 2023.

questions de sécurité habituelles, ainsi que les problèmes multiples causés par les *deep fakes*, ces fausses images ou vidéos très facilement créées grâce à la technologie des réseaux génératifs antagonistes (GAN, *Generative Adversarial Networks*), il est clair que le mur de la sécurité de l'IA est aujourd'hui suffisamment solide et proche pour qu'il soit essentiel de s'en protéger.

Le mur de l'interaction avec les humains

De très nombreux systèmes d'intelligence artificielle doivent interagir avec les humains. C'est le cas de certains robots, notamment, et on pense en particulier aux véhicules autonomes ; mais aussi les robots d'assistance aux personnes, les dialogueurs (*chatbots*) et plus généralement tous les systèmes qui ont besoin de communiquer avec leurs utilisateurs.

Les machines intelligentes d'aujourd'hui sont essentiellement des outils, pas des coéquipières. Au mieux, ces technologies sont utiles dans la mesure où elles étendent les capacités humaines, mais leurs compétences communicatives et cognitives sont encore inadéquates pour en faire des coéquipières utiles et de confiance. En effet, les machines collaboratives intelligentes doivent être flexibles et s'adapter aux états du coéquipier humain, ainsi qu'à l'environnement. Elles doivent comprendre les capacités et les intentions de l'utilisateur, et s'y adapter.

Or, nous ne comprenons pas suffisamment la cognition, la motivation et le comportement social de haut niveau de l'être humain social. Même si d'importants efforts de recherche en sciences cognitives ont été consacrés à la compréhension de la façon dont les humains pensent, apprennent et agissent, dans les environnements naturels, la séquence d'actions qui mène à un objectif n'est pas explicitement indiquée, voire la connaissance même des objectifs d'un humain reste complexe à appréhender. Stuart Russell a consacré un excellent ouvrage à ce sujet ¹⁵, dans lequel il montre à quel point il est difficile pour un système d'IA de connaître les intentions d'un humain ou d'un groupe d'humains, et il propose que l'IA questionne systématiquement lorsqu'il y a ambiguïté.

Réciproquement, il est également indispensable de permettre aux collaborateurs humains de comprendre les buts et actions des machines avec lesquelles ils sont en interaction. Les machines

15. RUSSELL Stuart J., *op. cit.*

ont souvent des caractéristiques physiques et des capacités très différentes de celles des humains, ce qui a un impact sur les rôles qu'elles peuvent jouer dans une équipe. Dans ce contexte, les besoins d'explications (que l'on nomme souvent « explicabilité ») de la part des systèmes d'intelligence artificielles sont cruciaux — ils font d'ailleurs l'objet d'une des mesures de la réglementation proposée par la Commission européenne (déjà citée). Mais les capacités d'explication des systèmes actuels d'IA sont très limitées, particulièrement lorsqu'il s'agit de réseaux neuronaux profonds dont les modèles internes sont composés de très grandes matrices de poids qu'il est difficile d'interpréter. J'en veux pour preuve les innombrables recherches sur l'explicabilité de l'IA, initialement popularisées par le programme « XAI » de la DARPA américaine lancé en 2017 ¹⁶.

Un problème plus général, illustré par le cas des véhicules autonomes ¹⁷, est celui du transfert du contrôle lorsque la machine reconnaît être dépassée (par exemple en cas de panne, de manque de visibilité, etc.) et doit transférer le contrôle à un humain, qui a besoin de beaucoup de temps pour assimiler le contexte et pouvoir reprendre la main. En résumé, l'interaction avec les humains est un sujet complexe et non résolu aujourd'hui, et il ne le sera pas de manière générale, mais plutôt application par application, comme dans l'exemple précédent.

Le mur de l'inhumanité

Je range plusieurs composantes dans ce cinquième mur que j'appelle globalement celui de l'humanité des machines, ou plutôt celui de leur inhumanité : acquisition du sens commun ; raisonnement causal ; passage au système 2 (au sens de Daniel Kahneman ¹⁸). Toutes composantes que nous, humains, possédons naturellement et que les systèmes d'intelligence artificielle n'ont pas — et n'auront pas à court ou moyen terme.

16. TUREK Matt, « Explainable Artificial Intelligence (XAI) », DARPA (Defense Advanced Research Projects Agency). URL : <https://www.darpa.mil/program/explainable-artificial-intelligence>. Consulté le 27 janvier 2023.

17. Dans la classification des niveaux d'autonomie pour le véhicule autonome, le niveau maximum 5 est celui de l'autonomie complète. Au niveau juste inférieur, 4, le véhicule gère presque toutes les situations mais rend la main dans des situations exceptionnelles, ce qui est extrêmement délicat à mettre en œuvre.

18. C'est-à-dire le mode de pensée réfléchi propre aux tâches complexes ou à la prise de décisions importantes (voir plus loin). Cf. KAHNEMAN Daniel, *Thinking, Fast and Slow*, New York : Farrar, Straus and Giroux, 2011, 512 p. (traduction française : *Système 1 / système 2. Les deux vitesses de la pensée*, Paris : Flammarion, 2012).

Le sens commun, c'est ce qui nous permet de vivre au quotidien. Nous savons qu'il ne faut pas mettre les doigts dans une prise électrique, mais une Alexa a récemment conseillé à une fillette de 10 ans de le faire. Dans les années 1980-1990, un grand projet de modélisation des connaissances, Cyc ¹⁹, a tenté de développer une base de connaissances du sens commun, en stockant des millions de faits et règles élémentaires sur le monde. Ce projet n'a pas abouti, les systèmes d'IA actuels ne sont capables que de résoudre des problèmes très précis dans un contexte limité, ils ne savent pas sortir de leur domaine de compétence ²⁰. Sans aller jusqu'à parler d'intelligence artificielle générale (celle qui fait peur et qu'aucun spécialiste du domaine n'envisage réellement à un horizon atteignable), faute de disposer de bases élémentaires faisant sens, les systèmes d'IA seront toujours susceptibles de commettre des erreurs monumentales aux conséquences potentielles dommageables.

Il est très largement connu que les systèmes d'IA entraînés par apprentissage établissent des corrélations entre variables sans se soucier de causalité. Dans l'exemple référence du classement d'images de chats, le réseau établit une corrélation — complexe certes — sans lien de causalité entre les données d'entrée (les pixels de l'image) et la donnée de sortie (la catégorie). Il existe de nombreux exemples de corrélations « fallacieuses » (*spurious correlations*) comme celui illustré par le graphique 1, tiré du site du même nom ²¹, qui établit une corrélation à 79 % entre le nombre de lancements de navettes spatiales et celui de doctorants en sociologie. Autrement dit, un système d'IA entraîné par apprentissage sera capable de reproduire cette relation et de prédire très correctement l'un à partir de l'autre. De même, on doit pouvoir décider si un jour est pluvieux à partir des ventes de parapluies, mais la causalité est évidemment dans l'autre sens. L'absence de prise en compte de la causalité dans les systèmes d'IA est une grande faiblesse : globalement, les systèmes d'apprentissage automatique se basent sur le passé pour faire des prédictions sur le futur, faisant implicitement l'hypothèse que la structure causale du système ne changera pas.

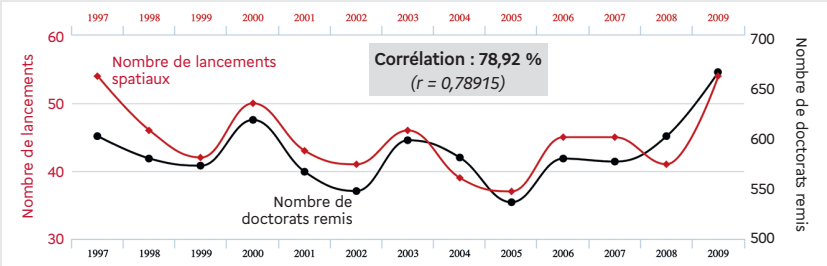
Il y a principalement deux manières de prendre en compte la causalité dans un système d'apprentissage automatique : le faire en

19. Présenté sur la page Wikipedia <https://en.wikipedia.org/wiki/Cyc>

20. Notons toutefois le développement récent de quelques systèmes multitâches, comme Gato de Google DeepMind, dont le seul réseau de neurones de 1,2 milliard de paramètres est capable de traiter (mal) des centaines de tâches différentes.

21. Site Internet <https://tylervigen.com/spurious-correlations>

Graphique 1. Corrélation (fallacieuse) entre les lancements spatiaux non commerciaux dans le monde et le nombre de doctorats de sociologie remis aux États-Unis



Source des données : Administration fédérale américaine et National Science Foundation. Graphique issu du site Internet <https://tylervigen.com/spurious-correlations>. Consulté le 31 janvier 2023.

injectant manuellement des connaissances sur le domaine d'intérêt, ou faire découvrir les liaisons causales à partir de données d'apprentissage²². Mais c'est très difficile : dans le premier cas, on revient aux problèmes des systèmes experts, avec les questions de cohérence des connaissances, de l'effort nécessaire pour les acquérir, etc. ; dans le deuxième cas, on ne sait traiter aujourd'hui que des exemples liés à très peu de variables. Si l'on ajoute les phénomènes de *feedback*, forcément présents dans les systèmes complexes, matérialisés par des boucles causales avec un contenu temporel, on ne sait plus le faire du tout.

Enfin, la troisième composante du mur de l'inhumanité est le passage au niveau du système 2. La très grande majorité des applications de l'IA consiste à (très bien) traiter un signal en entrée et à produire une réponse quasiment instantanée : reconnaissance d'objets ou de personnes dans des images et des vidéos, reconnaissance de la parole, etc. Dans son livre *Thinking, Fast and Slow*²³, Daniel Kahneman s'appuie sur des travaux en psychologie qui schématisent le fonctionnement de notre cerveau de deux manières différentes, qu'il nomme « système 1 » et « système 2 ». Système 1 est le mode rapide, proche de la perception : il ne vous faut qu'un instant pour reconnaître une émotion sur une photo, pour comprendre un mot ou une courte phrase. En revanche, si vous devez faire une multiplication compliquée et si vous n'êtes pas un calculateur prodige, vous devrez faire appel à du raisonnement

22. Une solution hybride étant de spécifier « manuellement » un graphe causal concis et de faire apprendre ses paramètres à partir de données.

23. *Op. cit.*

pour donner le résultat. Les processus mentaux plus lents sont de la responsabilité du système 2. Et les deux modes sont en permanente interaction, le système 1 fournit les éléments prétraités au système 2 qui peut conduire ses raisonnements dessus.

Cette théorie commence à inspirer les chercheurs en intelligence artificielle : aujourd'hui, avec l'apprentissage machine profond, l'IA est au niveau du système 1. Pour pouvoir dépasser cela, représenter les connaissances de sens commun, faire de la planification, des raisonnements élaborés, il faudra coder le système 2, et il faudra aussi coder l'interaction continue entre le système 1 et le système 2. Ce sont de beaux sujets de recherche pour les prochaines années, mais pour l'instant, un idéal encore bien lointain, même si de premiers exemples ont été réalisés comme l'a montré Francesca Rossi d'IBM lors de la conférence AAI (Association for the Advancement of Artificial Intelligence)-2022²⁴.

Il y a d'autres facteurs d'inhumanité dans l'IA — par exemple, la question de l'émotion, de l'empathie, ou encore la réalisation de l'intelligence collective, sujets intéressants que je ne développe pas ici, considérant que les trois premiers constituent déjà un mur très solide sur lequel l'IA va inévitablement buter dans les prochaines années.

Des pistes pour éviter de se fracasser sur les cinq murs

► **D'abord, sur la confiance.** C'est un des sujets majeurs de recherche et développement en IA depuis quelques années, pour les systèmes dits à risques ou stratégiques²⁵. On peut donc espérer avoir dans quelques années un ensemble de technologies permettant d'améliorer la confiance des utilisateurs envers les systèmes d'IA. Cela suffira-t-il ? Pour ma part, je pense que faute d'avancées fondamentales sur la nature des systèmes d'IA, on n'arrivera pas à des garanties suffisantes pour donner une confiance totale ; et probablement des accidents, catastrophiques ou non, continueront à se produire. En attendant d'avoir la possibilité de démontrer les facteurs de confiance, il faudra s'appuyer sur des quantités d'expériences : des centaines de millions de kilomètres parcourus sans

24. GANAPINI Marianna B. *et alii*, « Combining Fast and Slow Thinking for Human-like and Efficient Navigation in Constrained Environments », ArXiv, Cornell University, janvier 2022. URL : <https://arxiv.org/abs/2201.07050>. Consulté le 31 janvier 2023.

25. Pour des applications non stratégiques comme la recommandation de contenu ou de chemin optimal pour aller d'un point à un autre, cette question est évidemment moins cruciale. Nous utilisons ces systèmes tous les jours sans nous poser de questions.

encombre par des véhicules autonomes, des dizaines de milliers de décisions automatiques d'attribution de crédits non contestées, de diagnostics médicaux jugés corrects par des spécialistes, etc.). Et la confiance n'est pas qu'un sujet technologique, les facteurs humains et sociaux sont prépondérants.

► **Sur l'énergie**, plusieurs pistes sont développées, car le mur est proche de nous ! Si la croissance actuelle se poursuit, il faudra en 2029 un million de fois plus d'énergie qu'aujourd'hui pour entraîner les systèmes d'IA, à supposer que l'on continue à le faire de la même manière.

Je vois principalement trois types de solutions, dont les performances sont très différentes : a) des solutions matérielles ; b) l'amélioration des architectures et algorithmes de réseaux neuronaux profonds ; c) l'hybridation avec d'autres formalismes d'IA. Je ne m'étends pas sur a) : il existe des dizaines de développements de nouveaux processeurs, architectures 3D, architectures neuro-inspirées, massivement parallèles, etc., et d'aucuns disent que l'ordinateur quantique lorsqu'il existera, résoudra la question.

Les recherches sur b) sont plus intéressantes : améliorer la structure des réseaux, par exemple en les rendant parcimonieux par la destruction de tous les neurones et connexions qui ne sont pas indispensables ; ou encore par la définition d'architectures spécifiques, à l'image des *transformers* (BERT [Bidirectional Encoder Representations from Transformers], Meena, GPT3 [Generative Pre-trained Transformer 3], etc.) pour le langage, dont la structure permet de faire de l'autosupervision et donc au moins d'économiser l'annotation des données d'entraînement²⁶. Je pense également à l'amélioration du fonctionnement interne des réseaux, comme l'ont proposé divers auteurs avec des alternatives à la rétropropagation du gradient²⁷ ou autres.

Enfin, la troisième approche consiste à combiner les modèles neuronaux à d'autres types de modèles, essentiellement de deux natures : modèles numériques utilisés pour la simulation, l'optimisation et le contrôle de systèmes ; modèles symboliques, à base de connaissances. Si on est capable de combiner l'expertise contenue dans ces modèles, basée sur la connaissance établie au cours des

26. VASWANI Ashish *et alii*, « Attention Is All You Need », ArXiv, Cornell University, juin 2017. URL : <https://arxiv.org/abs/1706.03762>. Consulté le 31 janvier 2023.

27. Méthode d'apprentissage machine consistant, pour entraîner un réseau de neurones, à mettre à jour les poids de chaque neurone de la dernière couche vers la première (NDLR).

années par les meilleurs spécialistes humains, à celle contenue dans les données et que l'on pérennise par apprentissage, on doit pouvoir faire des économies substantielles de calcul, chacune des deux approches bénéficiant de l'autre. Le sujet est difficile car les modèles basés sur les données et ceux basés sur les connaissances ne sont pas compatibles entre eux, *a priori*.

► J'ai bien peur que le mur de la **sécurité** de l'IA soit très solide. Ou plutôt, il a une tendance naturelle à s'autoréparer lorsqu'il est percé. Je m'explique. De nouvelles failles des systèmes sont révélées chaque semaine ; des attaques contre des sites ou des systèmes stratégiques ont lieu en continu, qu'elles proviennent d'États mal intentionnés, de groupes terroristes ou mafieux... Les fournisseurs proposent régulièrement des mises à jour des systèmes d'exploitation et applications pour intégrer de nouvelles protections ou corrections de failles. En la matière, il s'agit toujours d'un jeu d'attaque et de défense. Les pirates conçoivent des attaques de plus en plus sophistiquées, l'industrie répond par des défenses encore plus sophistiquées. Les générateurs d'attaques antagonistes et de *deep fakes* produisent des attaques de plus en plus sournoises et des faux de plus en plus crédibles, l'industrie répond en augmentant la performance des détecteurs de faux. Les protocoles de chiffrement connaissent une augmentation périodique de la longueur des clefs de cryptographie, qui seront ensuite cassées par des algorithmes de plus en plus gourmands en ressources de calcul. Et ainsi de suite.

Pour les attaques adverses, une solution déjà évoquée est d'entraîner les réseaux avec de telles attaques, ce qui les rend plus robustes aux attaques connues. Mais, la course continuant, les types d'attaques continueront d'évoluer et il faudra, comme toujours, répondre avec un temps de retard.

Comparons au domaine militaire, qui a connu la course aux armements pendant de longues périodes : glaives, boucliers et armures il y a des milliers d'années, missiles et antimissiles aujourd'hui. La théorie de la dissuasion nucléaire a modéré cette course, puisque la réponse potentielle d'une puissance attaquée ferait subir des dommages si graves que cela ôterait toute envie d'attaquer. Il faudrait peut-être imaginer une doctrine équivalente à la dissuasion nucléaire en matière de cybersécurité de l'IA — ou espérer que l'IA apporte suffisamment de bonheur à la population mondiale pour que les causes sociales et autres (politiques, religieuses, économiques, etc.) de la malveillance disparaissent. Cela va prendre un peu de temps.

► J'aborde maintenant le mur de l'**interaction avec les humains**.

On peut commencer à le fracturer en ajoutant des capacités d'explication associées à la transparence des algorithmes utilisés. La transparence est indispensable lorsqu'il s'agit de systèmes qui sont susceptibles de prendre des décisions (imposées) ayant un impact important sur notre vie personnelle et sociale.

J'ai déjà abordé, dans la section correspondante, les travaux sur l'explicabilité. Les pistes sont nombreuses, je ne les détaillerai pas plus ici. Ma faveur va à celles qui combinent apprentissage numérique et représentations à base de connaissances (logiques, symboliques, ontologiques), même si elles sont encore à l'état de promesses : le passage du numérique (massivement distribué dans des matrices de poids) au symbolique est un sujet particulièrement ardu et non résolu de manière satisfaisante pour le moment.

Plus généralement, l'interaction entre systèmes d'IA et humains entre dans le concept général d'interaction humain-machine (HCI, *Human-Computer Interaction* en anglais). La communauté HCI travaille depuis des décennies sur le sujet général, avec des réalisations remarquables en visualisation, réalité virtuelle, réalité augmentée, interfaces haptiques²⁸, etc. ; on peut — et il faut — faire appel à leurs compétences pour le cas particulier des interactions avec des machines d'IA. La piste que je préconise donc (à l'image d'autres chercheurs et institutions qui l'ont également encouragée) est de resserrer les liens entre les deux communautés IA et HCI. Les chercheurs en IA y trouveront des éléments pour repousser le quatrième mur, et les chercheurs en HCI y trouveront la source de nouveaux défis pour leurs méthodes et leurs outils.

► Reste le mur de l'**inhumanité** : le plus éloigné, mais aussi le plus solide pour le moment. Le risque n'est pas encore très important mais s'amplifiera au fur et à mesure de l'insertion de systèmes IA de plus en plus autonomes, intrusifs et « impactants », dans notre société. En ce qui concerne la quête du sens commun, on a vu que des millions de dollars et des années de recherche investis sur Cyc n'ont pas réglé la question, loin de là. Peut-on miser sur de nouvelles architectures et organisations de réseaux neuronaux pour cela ? Certains l'espèrent. Personnellement, je miserai plutôt sur une autre branche de l'IA, celle de la robotique développementale (*developmental robotics*), qui a pour but de faire acquérir à des robots

28. Qui permettent de simuler des sensations tactiles à l'utilisateur qui manipule un objet virtuel (NDLR).

doués de sens les notions de base du monde, en interagissant avec leur environnement — peuplé d'objets et d'humains — et surtout en stimulant ce qu'on appelle la curiosité artificielle, à savoir doter les robots d'intentions et de capacités d'exploration et d'envoi de *stimuli* vers leur environnement afin d'en recevoir un *feedback* pour l'apprentissage par renforcement. Certaines expérimentations (photo ci-contre) faites par l'équipe INRIA (Institut national de recherche en sciences et technologies du numérique)-FLOWERS (FLOWing Epigenetic Robots and Systems) sont assez convaincantes en ce sens.

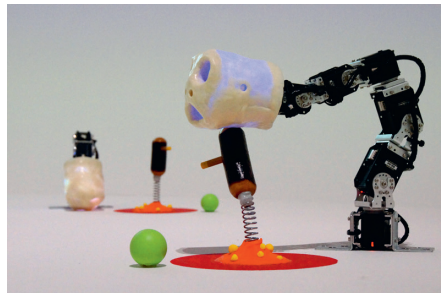


Photo issue de l'installation Ergo-Robots/FLOWERS Fields. Accessible sur le site du laboratoire FLOWERS. URL : <https://flowers.inria.fr/robots/ergo-robots-fr/>. Consulté le 31 janvier 2023.

J'ai déjà abordé les recherches en cours sur la découverte de la causalité par apprentissage automatique. C'est un sujet de longue haleine bien identifié mais disposant de peu de résultats. Je pense que l'introduction explicite de causalité, soit par conception d'architecture, soit par ajout d'une couche causale symbolique, apportera des résultats plus rapidement et plus concrètement — modulo la difficulté de combiner symbolique et numérique, dont j'ai déjà parlé.

Enfin, pour le passage de l'IA au niveau du système 2, j'ai abordé les pistes dans la section correspondante. La principale question est de savoir si cela peut être atteint par apprentissage de réseaux neuronaux — après tout, c'est bien ainsi que nous fonctionnons — ou par la conjonction de réseaux avec d'autres modes de représentation des connaissances, réalisant une IA hybride conjuguant symbolique et numérique, mettant en résonance les rêves et avancées de l'IA de la fin du XX^e siècle avec les progrès remarquables de celle du début du XXI^e.

Et tout ceci serait mobilisé pour réaliser des IA faibles, spécialisées dans la résolution d'un seul ou d'un petit nombre de problèmes, bien entendu, même si certains comme DeepMind ont l'ambition de développer une IA générale. Mais essayons déjà de ne pas nous écraser dans les murs de l'IA spécialisée. ■

À PROPOS DE CHATGPT

ChatGPT, sorti le 30 novembre 2022, est déjà devenu une idole, le nombre de personnes qui le consultent est tel que, dès à présent, il faut s'inscrire et attendre son tour pour dialoguer avec lui (ou elle). Dernier né parmi les OpenAI (entreprise financée par Microsoft), son succès repose sur plusieurs de ses caractéristiques : d'abord parce qu'il est doté d'une interface vocale, donc capable de répondre en langage naturel aux questions qu'on lui pose ; ensuite parce qu'il fonctionne grâce à un corpus très important de mots et de phrases disponibles sur Internet jusqu'à la mi-2021, qui lui permet de fournir des réponses souvent exactes à de très nombreuses questions, *a fortiori* de produire des articles (le test ayant été fait par de nombreux journalistes), voire des textes en apparence scientifiques. D'où d'ailleurs le danger qu'il présente de devenir, pour les élèves et étudiants, un substitut à leurs devoirs, voire d'être exploité abusivement pour produire des articles scientifiques. Mais, attention ! Il peut aussi commettre des erreurs, et même avoir des hallucinations et inventer de toutes pièces certaines réponses. Certains estimeront peut-être que

cela le rend plus humain mais, si ses capacités de calcul le rendent particulièrement agile, il n'en reste pas moins qu'il ne s'agit que d'une intelligence artificielle.

Le texte de ce bref encadré a été soumis à ChatGPT, qui n'y a pas répondu. Il reste à nos lecteurs le soin d'en savoir plus sur cette innovation qui a été abondamment commentée début 2023 mais qui, quand paraîtra ce numéro de *Futuribles*, sera peut-être déjà dépassée... Des programmes informatiques de plus en plus élaborés sont en permanence mis au point. Ils permettent déjà de produire des traductions de qualité, d'écrire des poèmes, de faire des calculs à une vitesse sidérante, de susciter l'angoisse qu'ils remplacent prochainement des millions de travailleurs. En attendant, il semble que ces développements exigent d'expurger du *Web* un maximum de contenus violents ou discriminants, tâche qu'exécute une entreprise de sous-traitance au Kenya, dont « les étiqueteurs seraient payés entre 1,32 et 2 dollars US de l'heure » (selon le magazine *Time* cité par *Les Échos* du 21 janvier 2023). ■

Hugues de Jovenel

Interview : ChatGPT par ChatGPT

Depuis son lancement, l'agent conversationnel (*chatbot*) ChatGPT fait la une des médias et suscite les conversations. Beaucoup y voient une version très performante d'intelligence artificielle (IA) et l'annonce de nouvelles modalités d'accès au savoir et de créativité. Présenté comme pouvant générer du texte pour répondre aux requêtes les plus variées des internautes, et fournir ses réponses dans plusieurs langues dont le français, nous l'avons testé en l'interrogeant sur lui-même et sur les perspectives qu'il ouvre.

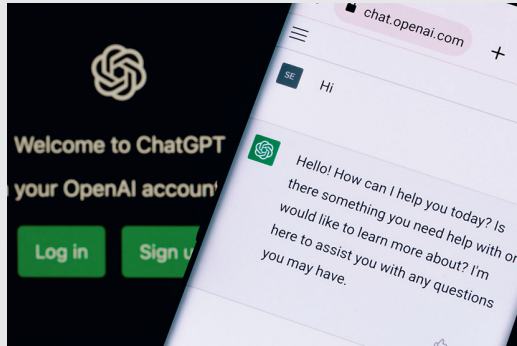
Pierre-Antoine Marti a mené l'interview pour Futuribles International. Le propos est retranscrit sans modifications : seules les plus grosses coquilles et erreurs grammaticales ont été corrigées, par égard pour nos lecteurs — il y en avait une dizaine. ■

ChatGPT, pourrais-tu expliquer ton fonctionnement ?

Je suis un modèle de langage entraîné par OpenAI. J'ai été formé à comprendre et à générer du texte en utilisant une grande quantité de données de textes provenant d'Internet. Mon but est de pouvoir comprendre les questions posées à moi et de fournir une réponse pertinente en utilisant les connaissances que j'ai apprises pendant mon processus d'entraînement.

Qu'apportes-tu au quotidien des gens ?

Je suis conçu pour aider les gens en leur fournissant des informations et des réponses à leurs questions. Je peux aider les gens à trouver des réponses à des questions factuelles, à comprendre des sujets complexes, à résoudre des problèmes et à effectuer des tâches. Je peux aider les entreprises à automatiser certaines tâches de service client, à améliorer les expériences utilisateur dans les applications, les *chatbots* et les assistants virtuels, à augmenter les ventes en ligne, à générer des contenus automatiquement, etc. En gros, je suis un outil puissant pour les gens pour effectuer des tâches de manière plus efficace et rapide.



© Ascannio / Shutterstock

Quels pourraient être les risques liés à ton utilisation ?

Il existe plusieurs risques liés à l'utilisation de modèles de traitement du langage comme moi. Certains de ces risques incluent :

- La propagation d'informations fausses ou trompeuses : si des informations erronées ou trompeuses sont présentes dans les données d'entraînement utilisées pour entraîner un modèle, il existe un risque que ces informations soient répercutées dans les réponses générées par le modèle.
- Les biais de données : les modèles de traitement du langage, comme moi, sont entraînés sur des données provenant d'Internet, qui peuvent contenir des biais sociaux et culturels. Ces biais peuvent être répercutés dans les réponses générées par le modèle.
- La violation de la vie privée : les modèles de traitement du langage peuvent être utilisés pour collecter des informations sensibles sur les utilisateurs, comme leur nom, leur adresse, leur numéro de téléphone, et ceci peut mener à la violation de leur vie privée.

- La dépendance : il est important de ne pas être trop dépendant des modèles automatisés, il est toujours important de vérifier et valider les réponses fournies par les modèles pour s'assurer de leur pertinence et de leur exactitude.

Il est important que les entreprises et les développeurs qui utilisent des modèles de traitement du langage comme moi prennent en compte ces risques et travaillent pour les minimiser en utilisant des données de qualité, en vérifiant les réponses générées par les modèles et en veillant à respecter la vie privée des utilisateurs.

Les modèles de traitement du langage comme moi seront-ils amenés à se développer à l'avenir, et comment ?

Les modèles de traitement du langage comme moi sont en effet en constante évolution et il est prévu qu'ils continuent à se développer à l'avenir. Il y a plusieurs raisons pour lesquelles cela se produit :

- Les données sont de plus en plus abondantes : avec la croissance d'Internet et l'augmentation de la quantité de contenu généré par les utilisateurs, il y a de plus en plus de données disponibles pour entraîner les modèles de traitement du langage.
- Les ordinateurs sont de plus en plus puissants : les progrès dans les technologies de calcul permettent d'entraîner des modèles de traitement du langage de plus en plus sophistiqués, avec des tailles de modèles plus importantes, plus de

couches, et des capacités de traitement plus élevées.

- L'IA est un domaine en constante évolution : les recherches en intelligence artificielle continuent à progresser, et ces avancées permettent de développer de nouvelles techniques pour améliorer les modèles de traitement du langage.
- L'augmentation de l'utilisation des modèles de traitement du langage : l'utilisation de modèles de traitement du langage dans divers domaines tels que le commerce électronique, la finance, les technologies de l'information, les services à la clientèle, etc., augmente et cela pousse à améliorer ces modèles pour répondre aux besoins spécifiques de ces domaines.

Pour ces différentes raisons, il est prévu que les modèles de traitement du langage deviennent de plus en plus précis, plus rapides et plus capables de générer des réponses plus naturelles et plus contextuelles. Les équipes de recherche continuent à travailler pour améliorer leur compréhension de la langue, de la thématique, de la logique, de la causalité, de la compréhension de l'émotion, etc. Et cela va continuer à améliorer l'expérience utilisateur dans les interactions avec des modèles de traitement de langue comme moi.

Est-il possible qu'advienne une « singularité technologique », si oui à quel horizon et avec quelles conséquences ?

La « singularité technologique » est une idée selon laquelle il y aurait un point à l'avenir où les progrès en

matière d'intelligence artificielle (IA) atteindraient un tel niveau que les machines seraient capables de s'autoaméliorer de manière exponentielle, dépassant rapidement les capacités humaines. Cette idée est souvent associée à l'écrivain et futuriste américain Vernor Vinge, qui a été l'un des premiers à utiliser le terme « singularité technologique » dans les années 1980.

Il est difficile de prévoir avec précision quand cela pourrait se produire, si cela se produit. Les experts sont partagés sur la question de savoir si cela se produira jamais, et si c'est le cas, quand cela pourrait se produire. Certains estiment qu'il pourrait se produire dans les prochaines décennies, tandis que d'autres estiment qu'il pourrait ne jamais se produire. Il y a également des avis selon lesquels, même si

l'IA atteint un certain niveau de puissance, il y aura des limites physiques et éthiques qui empêcheront une telle singularité technologique.

Les conséquences potentielles de la singularité technologique sont très débattues, mais elles pourraient être à la fois positives et négatives. Certaines personnes estiment que cela pourrait permettre de résoudre certains des plus grands défis de l'humanité, tels que la maladie, la pauvreté et les changements climatiques. D'autres estiment que cela pourrait être extrêmement dangereux, car les machines pourraient devenir plus intelligentes que les humains et pourraient prendre des décisions qui sont contraires aux intérêts humains. Il est important de noter que cela reste encore un sujet de débat et qu'il est difficile de prévoir les conséquences réelles d'un tel événement. ■

Source : cette interview, réalisée en janvier 2023, a été initialement publiée le 17 janvier 2023, en Note de veille de Futuribles International. URL : <https://www.futuribles.com/chatgpt-par-chatgpt/>. Consulté le 5 février 2023.