



IDÉE 12

L'IA POURRAIT DEVENIR PLUS FIABLE ET TRANSPARENTE

À l'avenir, l'IA pourrait avoir amélioré ses capacités de raisonnement, lui permettant de produire de meilleures analyses, de faire moins d'erreurs factuelles et d'être plus transparente. Cependant, ces améliorations peuvent s'avérer insuffisantes pour surmonter les problèmes liés aux mauvaises données.

AUJOURD'HUI

L'IA peut générer du texte et des images de haute qualité, mais peut commettre des erreurs logiques ou factuelles. Les réseaux de neurones¹ sont en mesure de reconnaître les modèles sans nécessairement comprendre le contenu ou le contexte. Par exemple, un grand modèle de langage (GML) recourt à la probabilité pour générer une sortie mot par mot, en fonction de la fréquence à laquelle les mots apparaissent à côté d'autres mots². Ne comprenant pas la signification des mots, il peut sortir des **hallucinations**³. Certains développeurs et développeuses cherchent à améliorer la factualité et l'exactitude en donnant aux GML l'accès à des bases de connaissances externes grâce à un processus appelé **génération augmentée de récupération (RAG)**⁴. Cependant, la RAG se fie à la qualité des données sources. Par exemple, l'Aperçu IA de Google Search utilise la RAG pour générer des résumés de requêtes de recherche. Toutefois, son incapacité à distinguer les sources faisant autorité des blagues sur les médias sociaux l'a amené à formuler des recommandations telles que mettre de la colle sur la pizza et regarder directement le soleil⁵.



Réseau de neurones

Un type d'IA modélisée sur le cerveau humain, dans lequel des nœuds interconnectés (neurones) traitent l'information en couches pour reconnaître les modèles, apprendre des données et prendre des décisions. Les GML et les générateurs d'images constituent un type de réseau neuronal.

Hallucination

Lorsqu'un système d'IA générative présente des informations fausses ou trompeuses comme vraies, telles que de fausses allégations, des sources inventées, des réponses à du contenu qui n'était pas dans la requête ou des images représentant des choses impossibles dans la réalité.

Récupération augmentée de génération

Un GML faisant référence à une base de connaissances faisant autorité, en dehors de ses données d'entraînement, avant de générer une réponse.

IA explicable vs interprétable

Une « IA explicable » peut fournir une explication raisonnable de la façon dont elle génère des sorties. Une IA est « interprétable » si un humain peut comprendre et expliquer comment elle fonctionne en interne.

Les réseaux de neurones manquent de transparence et peuvent être difficiles à comprendre et à contrôler. Les calculs qu'ils effectuent régulièrement s'avèrent si complexes que même les spécialistes humains éprouvent de la difficulté à comprendre comment ils transforment les entrées en sorties⁶. En d'autres termes, les réseaux de neurones manquent d'**interprétabilité**. Il est donc difficile de s'assurer qu'un système d'IA générative ne peut pas créer de contenu nuisible, ou qu'un système d'IA décisionnelle ne prend pas en compte des facteurs interdits tels que la race ou le sexe (voir l'idées L'utilisation de l'IA pour prédire le comportement humain pourrait s'avérer inefficace). Certains développeurs et développeuses essaient de surmonter cette limite en demandant au système de fournir une **explication** motivant la prise de décision. Cependant, on ignore actuellement si l'explication d'un modèle d'IA reflète avec précision sa pondération réelle des variables dans une décision⁷.



AVENIRS

Les futurs systèmes d'IA pourraient combiner différentes approches pour devenir plus fonctionnels et polyvalents. Les systèmes d'IA hybrides utilisent plus d'un type d'IA : par exemple, l'IA neurosymbolique combine la reconnaissance des formes des réseaux de neurones avec les règles interprétables par l'homme de l'**IA symbolique**⁸. À l'avenir, davantage de systèmes d'IA pourraient être composés de plusieurs systèmes avec des forces qui compenseraient les faiblesses des uns et des autres. Ces systèmes hybrides pourraient s'avérer plus performants, précis et très performants. Ils pourraient conduire au développement de tout nouveaux types d'IA.

Les systèmes d'IA pourraient s'avérer plus transparents et interprétables, les rendant ainsi moins biaisés. Une meilleure interprétabilité pourrait permettre aux développeuses et développeurs d'empêcher plus facilement un système d'IA décisionnel de prendre en compte à tort des facteurs, tels que la race. Par exemple, une banque pourrait montrer à sa clientèle pourquoi son système d'IA lui a refusé du crédit, lui permettant de demander un recours si – par exemple – elle estime que l'IA accorde trop d'importance à son code postal, lequel reflète qu'elle vit dans une zone multiculturelle de la ville. Cependant, il est peu probable que cette explication élimine complètement les biais (voir les idées Les biais des systèmes d'IA pourraient perdurer et L'utilisation de l'IA pour prédire le comportement humain pourrait s'avérer inefficace), d'autant plus que les biais peuvent être inhérents aux données.

L'IA pourrait être moins susceptible d'halluciner, bien qu'elle ne soit aussi bonne que les données, la logique et la formation auxquelles elle a accès. Les modèles hybrides, avec la capacité de raisonner de différentes manières, pourraient halluciner moins et fournir une analyse de meilleure qualité. Par exemple, ils peuvent utiliser une RAG pour recueillir des informations en ligne, puis utiliser une IA symbolique pour évaluer si l'information est crédible ou une blague. Cependant, l'amélioration des capacités de raisonnement ne résoudra pas le problème sous-jacent d'un écosystème d'informations pollué (voir l'idée L'IA pourrait briser Internet tel que nous le connaissons actuellement). Les futurs systèmes d'IA pourraient aborder ce problème en collectant plus de données directement via des capteurs. Ils pourraient également être assez intelligents pour reconnaître le défaut d'informations adéquates pour fournir une réponse précise.

IA symbolique

Une approche qui tente d'imiter le raisonnement humain, dans lequel la connaissance est représentée comme des symboles et manipulée conformément aux règles d'un système logique formel, comme la déduction et l'induction.



IMPLICATIONS

- Une transparence accrue et une réduction des préjugés pourraient **permettre à l'IA d'être utilisée dans des domaines où elle serait jugée inappropriée aujourd'hui**, comme l'application de la loi ou les procédures judiciaires
- Les outils d'IA pourraient permettre de **recueillir plus rapidement et plus facilement des informations de haute qualité et d'effectuer des analyses**, améliorant ainsi les processus décisionnels et la recherche
 - › **Les aides à la recherche en IA pourraient déstabiliser les emplois** de premier échelon comme les assistantes et assistants de recherche, les analystes juniors ou encore les avocates et avocats juniors
 - › **L'analyse de l'IA serait probablement préférable lorsqu'elle était limitée à des sources fiables et de haute qualité**, comme la documentation interne d'un organisme ou une base de données de revues académiques
- **Une analyse et un raisonnement de meilleure qualité pourraient accélérer l'adoption de l'IA** parmi les organismes actuellement peu enclins à prendre des risques. Cela pourrait réduire le besoin de surveillance et de contrôle de la qualité, mais aussi diminuer la probabilité que des erreurs soient remarquées
 - › **Les hallucinations peuvent devenir moins évidentes**, ce qui les rend plus difficiles à détecter
- Même avec un meilleur raisonnement, **des entrées fausses ou biaisées pourraient nuire aux performances et à la fiabilité de l'IA**
 - › **L'IA pourrait encore répandre de la désinformation**, lorsque l'argumentation et le raisonnement sont valides, mais que les hypothèses sont fausses

Notes de fin de texte

- 1 « [Que sont les hallucinations d'IA? | IBM](#) », 1^{er} septembre 2023.
- 2 Naveed, Humza, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes et Ajmal Mian. « [A Comprehensive Overview of Large Language Models](#) ». arXiv, 27 décembre 2023, 4.
- 3 « [Que sont les hallucinations d'IA? | IBM](#) », 1^{er} septembre 2023.
- 4 Martineau, Kim. « [Qu'est-ce que la génération augmentée de récupération?](#) » IBM Research Blog, 22 août 2023.
- 5 Friedman, Alan. « [L'Aperçu IA de Google peut dire aux utilisatrices et utilisateurs de manger des roches et de la colle, ou encore que la recherche Google est un monopole.](#) » PhoneArena, 24 mai 2024.
- 6 Fan, Feng-Lei, Jinjun Xiong, Mengzhou Li et Ge Wang. « [On Interpretability of Artificial Neural Networks : A Survey](#) ». *IEEE Transactions on Radiation and Plasma Medical Sciences* 5, no. 6 (novembre 2021): 741–60.
- 7 Fan, Feng-Lei, Jinjun Xiong, Mengzhou Li et Ge Wang. « [On Interpretability of Artificial Neural Networks : A Survey](#) ». *IEEE Transactions on Radiation and Plasma Medical Sciences* 5, no. 6 (novembre 2021): 741–60.
- 8 Hitzler, Pascal, Aaron Eberhart, Monireh Ebrahimi, Md Kamruzzaman Sarker et Lu Zhou. « [Neuro-Symbolic Approaches in Artificial Intelligence](#) ». *National Science Review* 9, no. 6 (4 juin 2022): nwac035.

© Sa Majesté le Roi du chef du Canada, 2025

Pour obtenir des informations sur les droits de reproduction : <https://horizons.gc.ca/fr/contactez-nous/>

PDF : PH4-222/2025F-PDF

ISBN : 978-0-660-76906-6

Also available in English under the title: AI could become more reliable and transparent.

AVERTISSEMENT

Horizons de politiques Canada (Horizons de politiques) est le centre d'excellence en prospective du gouvernement du Canada. Notre mandat est de doter le gouvernement du Canada d'une perspective et d'un état d'esprit tournés vers l'avenir afin de renforcer la prise de décisions. Le contenu de ce document ne représente pas nécessairement le point de vue du gouvernement du Canada ou des agences et ministères participants.