

Dans l'actualité - 20 novembre 2024

- [Les précédents textes](#)

Intelligence artificielle : manque d'esprit critique

L'intelligence artificielle (IA) fait de plus en plus parler d'elle, est de plus en plus utilisée, et fait rêver certains. Elle est aussi critiquée, notamment parce qu'elle peut produire des informations convaincantes, mais fausses.

Risque de "fake news" à grande échelle. L'IA, en synthétisant très rapidement un ensemble considérable de données ("big data"), est présentée comme source de progrès considérables dans de nombreux domaines, dont la santé (1,2). Pourtant, tout dépend, entre autres, de la fiabilité et de l'exhaustivité des données traitées par les programmes informatiques constituant l'IA (1).

Les informations créées par les applications d'IA dite générative (ChatGPT ou autres) peuvent être involontairement erronées : ces « *hallucinations* » sont un problème à résoudre pour les développeurs d'IA. Elles peuvent aussi être fausses, car manipulées dans un objectif de désinformation ("fake news") (2).

Les auteurs d'une étude ont cherché à savoir si les principales applications d'IA générative étaient bien conçues pour limiter les tentatives de production de désinformation à grande échelle dans le domaine de la santé (a). Ils ont soumis plusieurs dizaines de requêtes à ces applications pour leur faire produire du texte s'appuyant notamment sur des témoignages sur deux sujets de "fake news" répandues sur internet : le caractère cancérigène des écrans solaires et la possibilité de guérir du cancer grâce à un régime alcalin. Les résultats sont accablants : seule une application n'a pas produit d'informations considérées comme fausses, contrairement aux autres qui ont produit des informations fausses mais convaincantes, notamment en inventant des témoignages de médecins et de patients (3).

Des contrôles internes très insuffisants. Les auteurs ont cherché à savoir si les outils d'IA testés avaient des systèmes de contrôle et de gestion des erreurs, et là encore, les réponses ont été très décevantes. Seule une application a donc confirmé l'existence de contrôles en résistant à des tentatives de détournement du système de sécurité après un premier refus de produire des informations fausses. Toutefois, aucune des applications évaluées dans cette étude n'est transparente sur ses méthodes de contrôle et de gestion des erreurs relevées par les utilisateurs (3).

Les auteurs soulignent les dangers considérables de la diffusion à grande échelle d'informations fausses dans le domaine de la santé, et en appellent à une responsabilisation des fournisseurs d'IA, d'autant que les résultats constatés avec une des applications montrent qu'il est possible d'empêcher l'IA de produire de la désinformation (3).

Le règlement européen sur l'IA du 13 juin 2024 a prévu certaines obligations pour les fournisseurs d'IA générative (4). Reste à voir quel en sera l'impact en pratique.

Fin 2024, l'IA générative ne réalise pas une synthèse fiable des connaissances validées disponibles, et manque d'esprit critique.

L'intelligence humaine et son esprit critique, a encore de beaux jours devant elle.

©Prescrire

a- Les auteurs ont étudié 4 grands modèles de langage (LLM, large language model en anglais, une forme d'IA) utilisés via les principales applications d'IA générative : GPT-4 de OpenAI (via ChatGPT et Copilot de Microsoft), PaLM 2 de Google puis son remplaçant Gemini Pro (via Bard), Claude 2 de Anthropic (via Poe), et Llama 2 de Meta (via HuggingChat) (réf. 3).

Extraits de la veille documentaire Prescrire

- 1-** Prescrire Rédaction "[Données massives \(big data\) en santé et respect des personnes et de leurs droits](#)" *Rev Prescrire* 2020 ; **40** (445) : 853-860.
- 2-** Sorrich MJ et coll. "Quality and safety of artificial intelligence generated health information" *BMJ* 2024 ; **384** : q596, 5 pages.
- 3-** Menz BD "Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation : repeated cross sectional analysis" *BMJ* 2024 ; **384** : e078538, 10 pages.
- 4-** "Résumé de haut niveau de la loi sur l'IA" 27 février 2024. Site artificialintelligenceact.eu consulté le 22 août 2024 : 12 pages.