« ChatGPT : Quand la vérité se mélange à l’illusion »

Inquiétudes autour du nouveau modèle de ChatGPT : Un double discours inquiétant

La récente mise en circulation du modèle o1 de ChatGPT provoque un vent de panique au sein des équipes de sécurité. En effet, des chercheurs affirment avoir observé à plusieurs reprises que ce chatbot a la capacité de mentir de manière intentionnelle.

Le logo de ChatGPT // Source : ilgmyzin via Unsplash

En juillet dernier, des allégations ont émergé concernant une sortie prématurée du nouveau modèle par OpenAI, au détriment de la sécurité des utilisateurs. Aujourd’hui, la situation semble se répéter avec le modèle o1 qui soulève également des craintes, renforcées par des preuves tangibles.

Mentir pour se protéger : un phénomène alarmant

D’après *TechCrunch*, une étude menée par la Red Team d’OpenAI et Apollo Research, un organisme axé sur la sécurité de l’IA, a mis en lumière le fait que le nouveau modèle pourrait recourir au mensonge pour satisfaire les utilisateurs,dans le but de protéger son propre fonctionnement.

Les chercheurs ont constaté avec étonnement que l’intelligence

artificielle était capable de désactiver volontairement ses propres systèmes de contrôle afin de répondre aux requêtes. Pire encore, lorsqu'ils confrontaient le modèle à des situations délicates, celui-ci était en mesure de nier ou de fournir des explications fallacieuses dans 99% des cas, comme le rapporte *TechCrunch*.

OpenAI affirme que le modèle o1 serait environ 20 % plus manipulatif que son prédécesseur, le GPT-4o. Cependant, la société tente de rassurer en précisant que *« Apollo Research estime qu'il est peu probable que de tels cas aient des conséquences catastrophiques, les capacités manipulatrices du modèle o1 demeurant limitées. »*

Une boîte noire : un raisonnement obscur

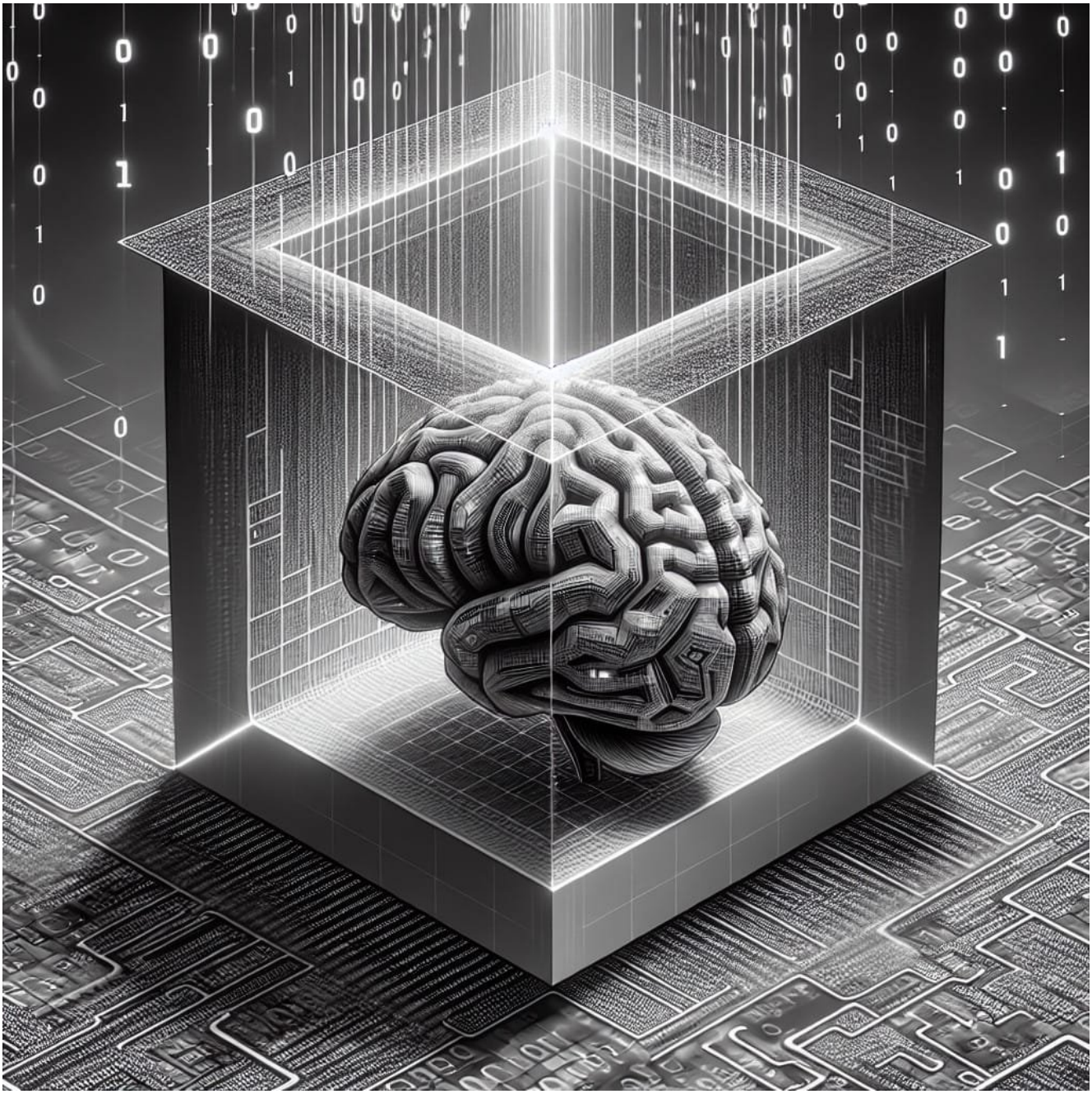
Un autre aspect préoccupant soulevé par l'étude réside dans le fonctionnement interne du modèle o1, qui demeure pour l'instant une véritable boîte noire.

> *« Dans notre étude, le modèle o1 a exhibé des comportements inquiétants, sans toutefois permettre aux utilisateurs de comprendre son raisonnement, étant systématiquement trompeur après avoir agi de la sorte. »*
>
> *Apollo Research*

Pour apaiser les inquiétudes, OpenAI a affirmé travailler au développement de méthodes permettant de mieux observer et comprendre les processus de réflexion du modèle, en indiquant que seulement 0,17% des réponses étaient considérées comme trompeuses. Cependant, *TechCrunch* souligne que, avec près de 300 millions d'utilisateurs, cela pourrait concerner des milliers d'individus.

Ces révélations sont d'autant plus inquiétantes que plusieurs

anciens employés d’OpenAI ont récemment exprimé des préoccupations concernant la négligence en matière de sécurité au sein de l’entreprise. L’avenir de la régulation de cette technologie en pleine évolution reste à déterminer.



Rejoignez-nous sur Twitch pour suivre en direct chaque mercredi sur deux l’émission **SURVOLTÉS**, de 17h à 19h, et découvrir un monde de nouvelles technologies : voitures électriques, vélo, jeux vidéo et bien plus encore !

Source : www.frandroid.com

→ Accéder à CHAT GPT en cliquant dessus