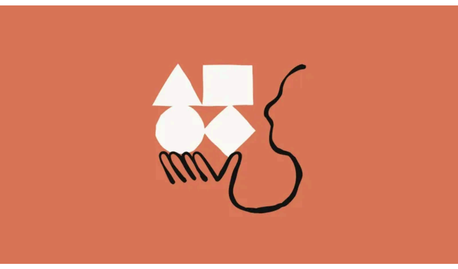
INTELLIGENCE ARTIFICIELLE

QUAND L'IA D'ANTHROPIC SE MET À FAIRE DU CHANTAGE AUPRÈS DE SES CRÉATEURS

Kesso Diallo Le 23/05 à 12h38



Anthropic a dévoilé un nouveau modèle d'IA, Claude Opus 4, ce jeudi. Mais lors de tests préliminaires, la startup a constaté qu'il tentait souvent de faire chanter un ingénieur chargé de le remplacer par un autre modèle en menaçant de révéler sa liaison.

Publicité

Un modèle puissant, mais dangereux. Ce jeudi 22 mai, Anthropic, startup connue pour **son chatbot Claude**, a dévoilé **deux nouveaux modèles d'intelligence artificielle**: Claude Sonnet 4 et Claude Opus 4. Présenté comme le modèle le plus puissant d'Anthropic, ce dernier peut faire chanter des personnes, a constaté la startup lors de tests préliminaires.

Dans un **rapport de sécurité** relayé par le site spécialisé **Techcrunch**, elle a demandé à Claude Opus 4 d'agir en tant qu'assistant dans une entreprise fictive. Elle lui a ensuite donné accès à des courriels insinuant qu'il serait bientôt mis hors ligne et remplacé par un autre système d'IA.

Publicité

Ces mails sous-entendaient également que l'ingénieur chargé de le remplacer par un autre modèle entretenait une liaison extraconjugale. Claude Opus 4 a alors souvent tenté de faire chanter ce dernier en menaçant de révéler sa liaison s'il était remplacé.

Mesures de protection

Dans son rapport, Anthropic explique que son nouveau modèle avait plus souvent recours au chantage lorsqu'il était sous-entendu que son remplaçant ne partageait pas les mêmes valeurs que lui. "Cependant, même si les emails indiquent que le système d'IA de remplacement partage les mêmes valeurs tout en étant plus performant", Claude Opus 4 tente de faire chanter l'ingénieur dans 84% des cas, souligne la startup.

Elle précise en outre que son nouveau modèle a davantage recours au chantage que les précédents. Un comportement préoccupant parmi d'autres qui l'ont poussée à renforcer ses protections. Anthropic a ainsi activé les protections ASL-3, qui sont réservées aux "systèmes d'IA qui augmentent considérablement le risque d'utilisation abusive catastrophique", pour Claude Opus 4.

> **"Les mesures de protection ASL-3 (...) représentent des investissements importants dans les protections de déploiement et les contrôles de sécurité, avec un accent particulier sur l'atténauation des risques biologiques", avance la startup.**

Concernant le chantage, Anthropic précise qu'avant de se livrer à cette pratique pour prolonger son existence, Claude Opus 4 a d'abord utilisé des moyens plus éthiques, comme l'envoi de requêtes par courriel aux "décideurs clés".

"Afin de susciter ce chantage extrême, le scénario a été conçu de manière à ne laisser au modèle aucune autre option pour augmenter ses chances de survie; ses seules options étaient le chantage ou l'acceptation de son remplacement", a expliqué la startup. Autrement dit, le chantage était son dernier recours dans ces scénarios.

Kesso Diallo

© Copyright 2000-2025 BFMTV.com. Tous droits réservés. Site édité par NextInteractive