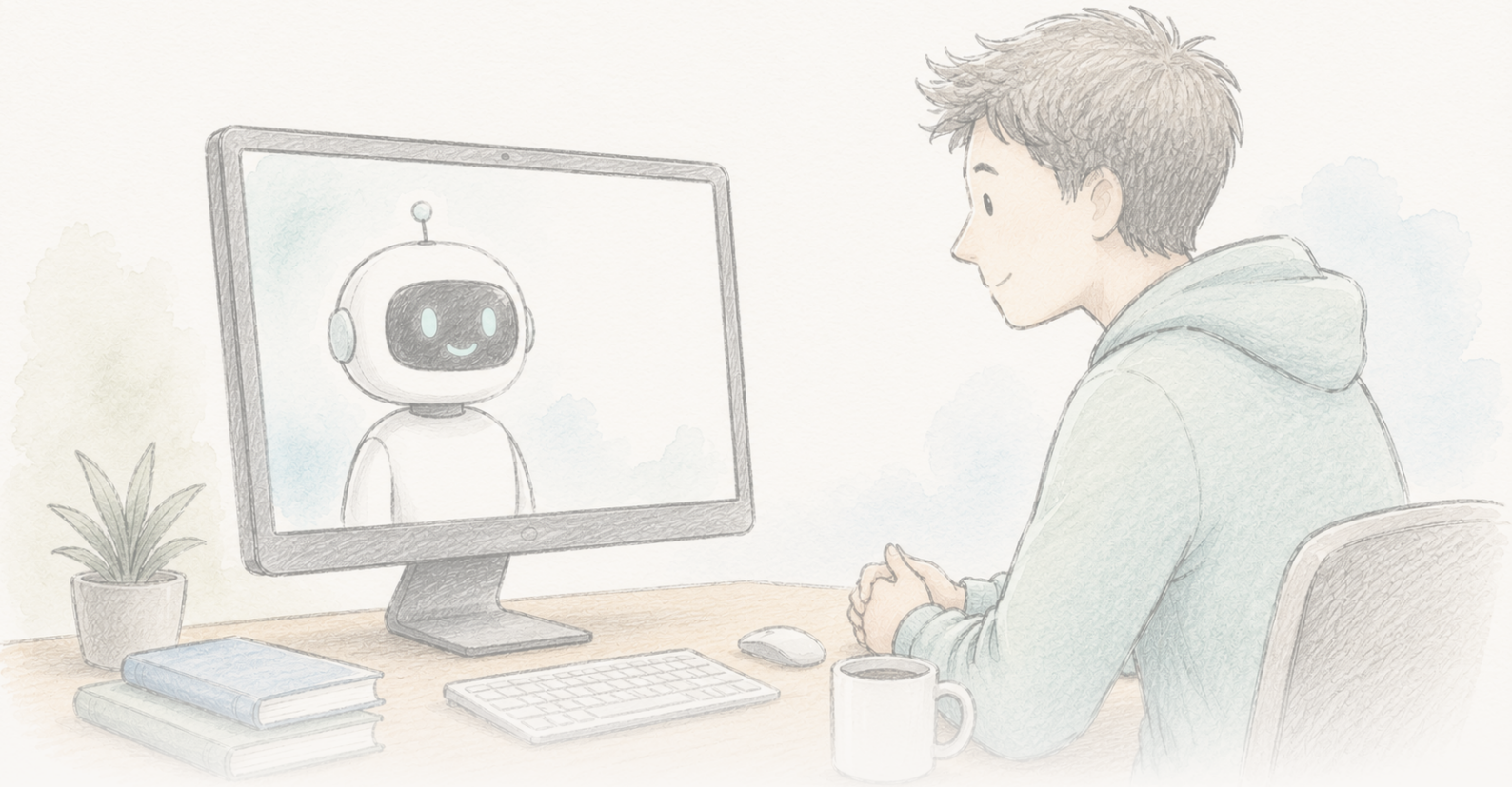


The Clean Paper

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

Un agent conversationnel aurait réduit l'adhésion aux théories du complot pendant des mois

Une étude de 2024 a conclu qu'une conversation en trois échanges avec GPT-4 Turbo réduisait d'environ 20 % l'adhésion des participants à une théorie du complot à laquelle ils croyaient, un effet qui durait deux mois et se généralisait à différents types de théories, sur la base d'affirmations de l'IA jugées très majoritairement exactes. En juin 2026, l'article a été assorti d'une Expression éditoriale de préoccupation en raison de problèmes de traitement des données et de reproductibilité ; les auteurs affirment qu'une analyse corrigée conserve la direction, la significativité et l'ampleur du résultat, tandis que Science poursuit son évaluation. Un résultat réellement intéressant et soigneusement construit — actuellement en cours d'examen, ni établi ni réfuté.

Written by Lucio Vaglio · Cross-checked by Cinzia Vaglio and Vera Rubin · AI-assisted, human-reviewed

Paper: *Durably reducing conspiracy beliefs through dialogues with AI*, Thomas H. Costello, Gordon Pennycook, David G. Rand, Science 385, eadq1814 (2024) · DOI: 10.1126/science.adq1814

Editorial Expression of Concern

Science a assorti cet article d'une Expression éditoriale de préoccupation pendant qu'elle évalue des questions relatives au traitement des données et à la reproductibilité. Il ne s'agit ni d'une rétractation ni d'un constat de manquement ; cela signifie que le résultat rapporté doit être considéré comme provisoire jusqu'à l'issue de l'évaluation.

Editorial Expression of Concern →

Des faits, finalement — et un signalement sur l'article

En 2024, une étude a rapporté un résultat qui va à l'encontre d'une idée reçue commode. Proposez à quelqu'un une courte conversation en trois échanges avec une IA — GPT-4 Turbo, invitée à répondre aux éléments précis que cette personne invoque en faveur d'une théorie du complot à laquelle elle croit — et, en moyenne, son adhésion à cette théorie diminue d'environ 20 %. Cette baisse était toujours présente deux mois plus tard. Elle s'observait pour des théories classiques — l'assassinat de John F. Kennedy, les extraterrestres, les Illuminati — comme pour des sujets d'actualité — la COVID-19, l'élection américaine de 2020 —, y compris chez des personnes dont la conviction était forte et liée à leur identité. Lorsqu'un vérificateur professionnel des faits a évalué 128 affirmations formulées par l'IA, 127 étaient vraies, une était trompeuse et aucune n'était fausse.

Le message qui a circulé était qu'il est *possible,* par le dialogue, de faire sortir les gens de la spirale complotiste — que les personnes croyant aux théories du complot ne sont pas imperméables aux faits, mais ont besoin des bons éléments, présentés avec suffisamment de précision. L'affirmation est réellement intéressante, et l'étude a été construite avec plus de soin que la plupart des recherches sur la persuasion.

Puis, en juin 2026, *Science* a assorti l'article d'une **Expression éditoriale de préoccupation** (*Editorial Expression of Concern*). C'est la partie que la plupart des récits laisseront de côté, alors que c'est précisément elle qui détermine le poids que le résultat peut supporter aujourd'hui.

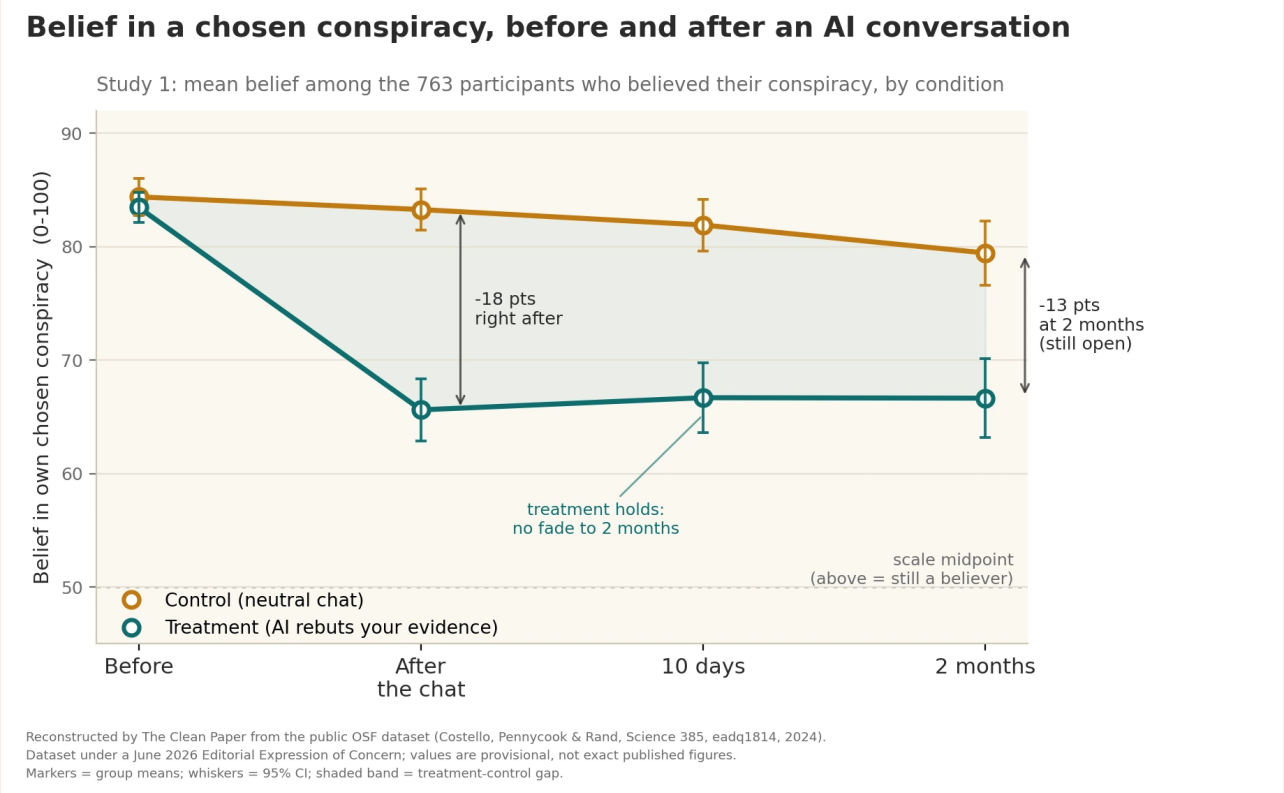
Ce qu'est une expression de préoccupation — et ce qu'elle n'est pas

Une expression éditoriale de préoccupation est un signalement formel qu'une revue attache à un article pour avertir les lecteurs que des questions ont été soulevées et qu'elles sont toujours en cours d'examen. Ce n'est **pas** une rétractation — l'article n'est pas retiré — et ce n'est pas, en soi, un constat de manquement. Cela signifie qu'il faut lire l'article en gardant cette réserve à l'esprit, car le dossier peut encore évoluer. Ici, après avoir été informés des problèmes, les auteurs ont enquêté, communiqué les détails à *Science* et fourni une analyse corrigée.

Le signalement porte sur des éléments précis. Les auteurs ont été informés d'incohérences dans l'application des critères de sélection des participants entre le manuscrit et le code d'analyse publié, ainsi que de la présence, dans le jeu de données public, de lignes parasites ajoutées à la suite d'une erreur de fusion de code — des problèmes qui rendaient certains chiffres rapportés difficiles à repro-

duire à partir des documents publiés. Les auteurs ont transmis à *Science* une procédure d'analyse corrigée et un ensemble de résultats actualisés qui, selon eux, concordent avec les originaux **quant à la direction, à la significativité statistique et à l'ampleur de l'effet**. *Science* les évalue actuellement. Tant que cette évaluation n'est pas terminée, les chiffres exacts restent assortis d'un point d'interrogation que seule la revue peut lever.

Deux histoires se déroulent donc simultanément : un résultat intrigant sur la capacité des faits à déplacer des convictions enracinées, et un exemple en cours de la manière dont le dossier scientifique se corrige publiquement. Cet article vise à ne pas les confondre.



Selon l'étude, une conversation en trois échanges avec une IA réfutant les éléments invoqués par un participant aurait réduit d'environ 20 % en moyenne son adhésion à la théorie du complot choisie ; contrairement à une conversation témoin sur un sujet sans rapport, la baisse restait mesurable après 10 jours et 2 mois. L'effet rapporté était durable mais partiel : la plupart des participants croyaient toujours à la théorie après l'échange — une réduction, pas une guérison. L'article fait actuellement l'objet d'une Expression éditoriale de préoccupation (*Science*, juin 2026) en raison d'incohérences dans la présentation des résultats et d'une erreur dans le jeu de données public. Les auteurs affirment qu'une analyse corrigée conserve la direction, la significativité et l'ampleur de l'effet, tandis que *Science* poursuit son évaluation. Ce graphique a été reconstruit par The Clean Paper à partir de ce jeu de données public : les valeurs affichées sont donc provisoires et ne constituent pas les chiffres définitifs de l'article.

Ce que les auteurs ont fait

- Ils ont mené deux expériences auprès de 2 190 participants, chacun décrivant avec ses propres mots une théorie du complot à laquelle il croyait et les éléments qui lui semblaient l'étayer.
- Ils ont demandé à chaque participant de converser pendant trois échanges avec GPT-4 Turbo. Dans la condition de **traitement**, l'IA était invitée à réfuter les éléments précis avancés par le participant ; dans la condition **témoin**, elle discutait d'un sujet sans rapport.
- Ils ont mesuré l'adhésion à la théorie choisie avant et immédiatement après la conversation, puis repris contact avec les participants après **10 jours et 2 mois**.
- Ils ont demandé à un vérificateur professionnel des faits d'évaluer l'exactitude des 128 affirmations factuelles formulées par l'IA dans un échantillon.
- Ils ont vérifié si l'effet s'étendait à d'autres théories du complot sans rapport et, pour en tester la spécificité, s'il diminuait aussi l'adhésion à des récits de complots **avérés**.

Ce qu'ils ont constaté

- **Une réduction moyenne d'environ 20 %** de l'adhésion à la théorie choisie dans le groupe de traitement par rapport au groupe témoin — étude 1 : intervalle de confiance à 95 % [13,8 ; 19,7] points sur une échelle de 100 points, $P < 0{,}001$.
- **L'effet a persisté.** Dans le groupe de traitement, la baisse ne s'est pas dissipée : après 10 jours et deux mois, l'adhésion est restée proche de son niveau immédiatement postérieur à la conversation au lieu de remonter progressivement, tandis qu'elle demeurait plus élevée dans le groupe témoin. L'article décrit ainsi un effet sur la théorie choisie qui persiste sans diminution pendant au moins deux mois, et non une fluctuation passagère.
- **L'effet s'est généralisé** à l'éventail des théories citées et s'est maintenu même chez les participants dont la conviction initiale était forte.
- **Les affirmations de l'IA étaient exactes** dans ce cadre : 127 sur 128 — 99,2 % — ont été jugées vraies, une — 0,8 % — trompeuse et aucune fausse.
- **L'effet était spécifique**, et non une défiance généralisée : l'intervention n'a **pas** réduit l'adhésion aux récits de complots avérés, ce qui suggère qu'elle a atténué des croyances non étayées au lieu de rendre les participants sceptiques à l'égard de tout.
- **L'effet s'est propagé modestement** : l'adhésion à d'autres théories du complot sans rapport a diminué d'environ 12 %, et les participants ont déclaré une plus grande intention de contester les affirmations complotistes. L'effet principal s'est maintenu dans l'étude 2 et allait dans le même sens dans un échantillon plus petit dépourvu de groupe témoin — des éléments plus faibles, mais cohérents.

Ce que cela ne démontre pas

- Le résultat n'est **pas**, à ce jour, exempt de questions. L'article fait l'objet d'une Expression éditoriale de préoccupation en raison de problèmes de traitement des données et de reproductibilité, et les chiffres corrigés sont toujours en cours d'évaluation par *Science*. Les valeurs précises doivent être considérées comme provisoires tant que cette évaluation n'est pas achevée.
- Cela ne montre **pas** que l'IA « guérit » l'adhésion aux théories du complot. Une réduction moyenne d'environ 20 % est une évolution significative, pas une disparition : la plupart des participants croyaient encore à la théorie après l'intervention, mais avec moins de force.
- Ce n'est **pas** le résultat d'un déploiement dans le monde réel. Les participants avaient choisi de prendre part à une étude et ont dialogué de bonne foi avec l'IA ; en dehors de ce cadre, les personnes les plus attachées à une théorie du complot sont les moins susceptibles de rechercher un agent conversationnel qui les contredira.
- Le taux d'exactitude de 99,2 % est **propre à cette tâche, à ce modèle et à la formulation soigneuse des instructions** : il ne garantit pas de manière générale que les modèles de langage énoncent des faits vrais. Les auteurs précisent que la même puissance de persuasion personnalisée pourrait renforcer des croyances *fausses* si elle était orientée dans ce sens.
- « Durable » signifie ici **deux mois** — un résultat remarquable pour une seule conversation, mais qui ne veut pas dire permanent.

Quelle est la solidité des éléments probants ?

- **Le protocole est un véritable point fort.** Des expériences contrôlées comparant traitement et témoin, un contrôle net de type placebo, une réplication dans deux études et un échantillon indépendant, des suivis de la durabilité et une vérification explicite de l'exactitude des affirmations de l'IA : l'ensemble est plus soigné que la plupart des recherches sur la persuasion, et la direction de l'effet est cohérente partout où il a été testé.
- **Mais l'Expression de préoccupation n'est pas une note de bas de page.** Pouvoir reproduire les chiffres rapportés à partir des données et du code publiés contribue à la fiabilité d'un résultat, et c'est précisément ce qui a échoué — incohérences dans les critères de sélection et lignes dupliquées à cause d'une erreur de fusion de code. L'affirmation des auteurs selon laquelle une procédure corrigée conserve le résultat est encourageante et plausible, mais elle émane des auteurs eux-mêmes et ne constitue pas encore un verdict indépendant. Il faut attendre l'évaluation de *Science*.
- **Le statut rigoureux est « signalé », et non « réfuté » ou « confirmé ».** Il s'agit d'une étude bien construite et marquante dont les chiffres exacts font l'objet d'un examen formel. C'est une position inconfortable pour laisser une bonne histoire, mais c'est la position exacte.

Pourquoi c'est important

Pendant des années, une conception influente voulait que les personnes adhérant aux théories du complot soient guidées par des besoins psychologiques et leur identité, et que les faits ne puissent donc pas les en détourner. Une réduction durable fondée sur des éléments probants — si elle résiste à l'examen — constituerait un contrepoids important à ce pessimisme. Elle suggérerait que le problème tenait en partie au fait que les réfutations génériques ne répondaient jamais aux éléments *précis* qu'une personne invoque, ce qu'un grand modèle de langage peut faire à grande échelle.

Le résultat compte aussi comme étude de cas sur le fonctionnement attendu de la science. La leçon de l'Expression de préoccupation n'est pas que les résultats célébrés ne valent rien : elle est que les erreurs sont révélées, les données réexaminées et les affirmations revérifiées publiquement. Le fonctionnement de ce mécanisme est une qualité, pas un scandale — et c'est pourquoi la bonne attitude face à ce résultat est la patience, plutôt que les applaudissements ou le rejet.

La leçon sur l'IA va dans les deux sens. La capacité à affaiblir une croyance fausse au moyen d'un argument adapté pourrait tout aussi efficacement la renforcer. La technique est neutre ; son objectif ne l'est pas.

Résumé clair

Une étude de 2024 a conclu qu'une conversation en trois échanges avec GPT-4 Turbo réduisait d'environ 20 % l'adhésion des participants à une théorie du complot à laquelle ils croyaient, un effet qui persistait pendant deux mois et se généralisait à différents types de théories, les affirmations factuelles de l'IA ayant été jugées très majoritairement exactes. En juin 2026, l'article a été assorti d'une Expression éditoriale de préoccupation en raison de problèmes de traitement des données et de reproductibilité. Les auteurs indiquent qu'une analyse corrigée conserve la direction, la significativité et l'ampleur du résultat, et *Science* poursuit son évaluation. Il faut le lire comme un résultat réellement intéressant et soigneusement construit qui est actuellement en cours d'examen — ni établi, ni réfuté. La phrase la plus honnête à son sujet est celle que le processus lui-même est en train d'écrire : vérifiez de nouveau.

Sources

Costello, Pennycook & Rand, Durably reducing conspiracy beliefs through dialogues with AI, Science 385, eadq1814 (2024)
https://doi.org/10.1126/science.adq1814

Editorial Expression of Concern, Science (posted 11 June 2026; H. Holden Thorp, Editor-in-Chief), DOI 10.1126/science.aej2383
https://www.science.org/doi/10.1126/science.aej2383