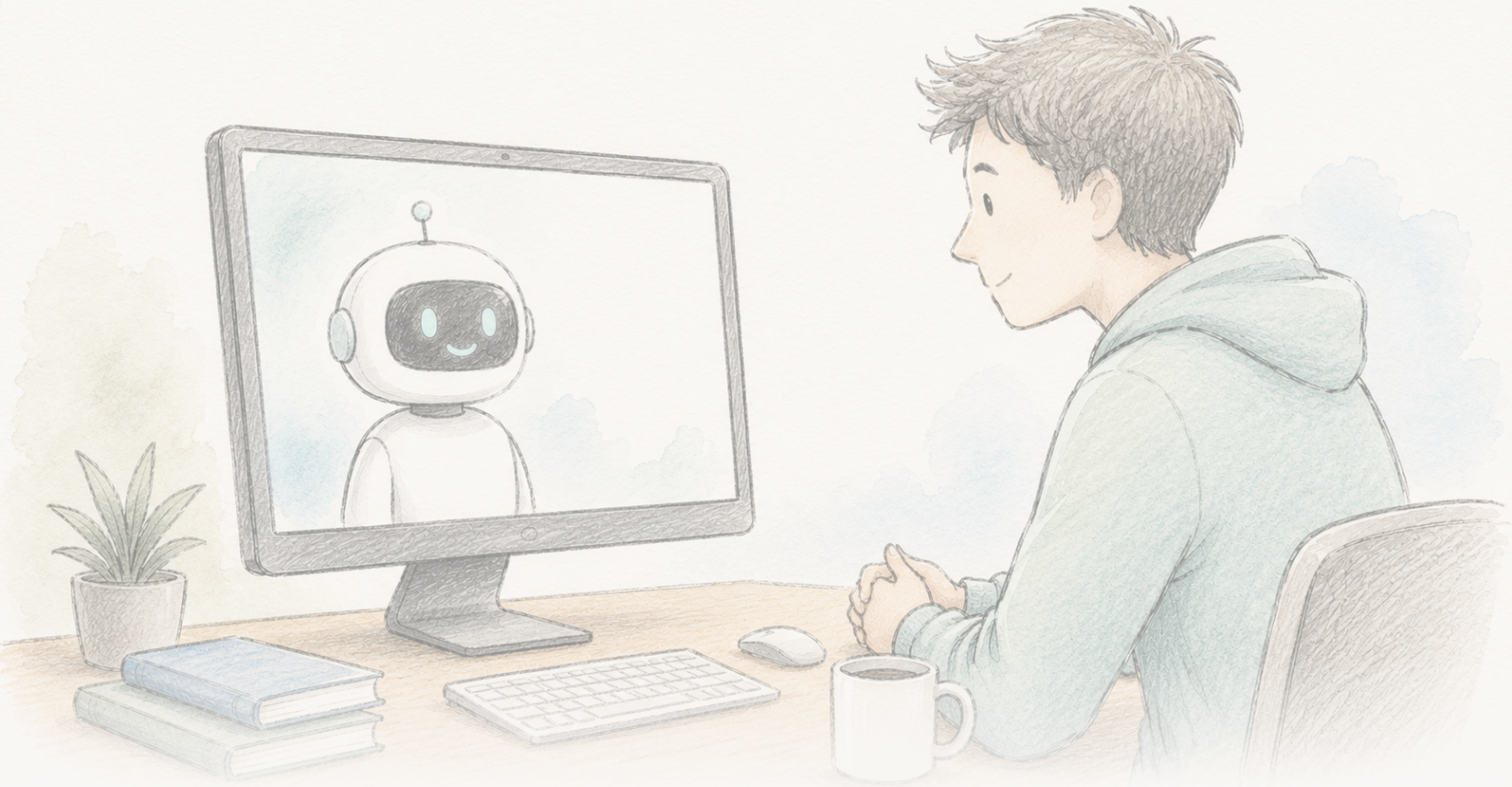


The Clean Paper

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

チャットボットとの対話で、陰謀論への信念が数か月弱まったように見えた

2024 年の研究では、*GPT-4 Turbo* との *3* 往復の会話により、本人が信じる陰謀論への信念が約 *20%* 低下し、効果は *2* か月続き、陰謀論の種類を越えて見られ、*AI* の主張も圧倒的に正確と評価された。*2026* 年 *6* 月、データ処理と再現性の問題を理由に論文には *Editorial Expression of Concern* が付けられた。著者らは修正解析でも効果の方向、有意性、大きさは保たれるとしており、*Science* は現在も評価中である。本当に興味深く、丁寧に設計された結果だが、いまは審査中。確定でも、反証済みでもない。

Written by Lucio Vaglio · Cross-checked by Cinzia Vaglio and Vera Rubin · Edited by Michele Renda · Figures by Nova Vela · 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 は、データ処理と再現性に関する問題を評価している間、この論文に Editorial Expression of Concern を付けている。これは撤回でも不正行為の認定でもない。審査が解決するまで、報告結果を暫定的なものとして読むべきだという意味である。

Editorial Expression of Concern →

結局、事実は効くのか —— そして論文には警告旗が付いた

2024 年、ある研究が、よくある安心できる常識に逆らう結果を報告した。自分が信じる陰謀論について、本人が根拠だと思っている証拠を具体的に挙げ、それに GPT-4 Turbo が答える短い 3 往復の会話をすると、平均してその陰謀論への信念が約 20% 下がった。低下は 2 か月後にも残っていた。ジョン・F・ケネディ暗殺、宇宙人、イルミナティといった古典的陰謀論から、COVID-19 や 2020 年米大統領選のような時事的なものまで効果が見られ、信念が強く自己同一性と結び付いている人にも及んだ。専門のファクトチェッカーが AI の 128 の事実主張を評価すると、127 は正しく、1 つは誤解を招き、誤りは 0 だった。

広まった見出しは、「人は陰謀論の深みから話し合いで引き戻せる」というものだった。陰謀論を信じる人も証拠が届かない存在ではなく、その人が挙げる根拠へ十分具体的に応答する、適切な証拠が必要なだけなのかもしれない。それは本当に興味深い主張であり、研究設計も多くの説得研究より丁寧だった。

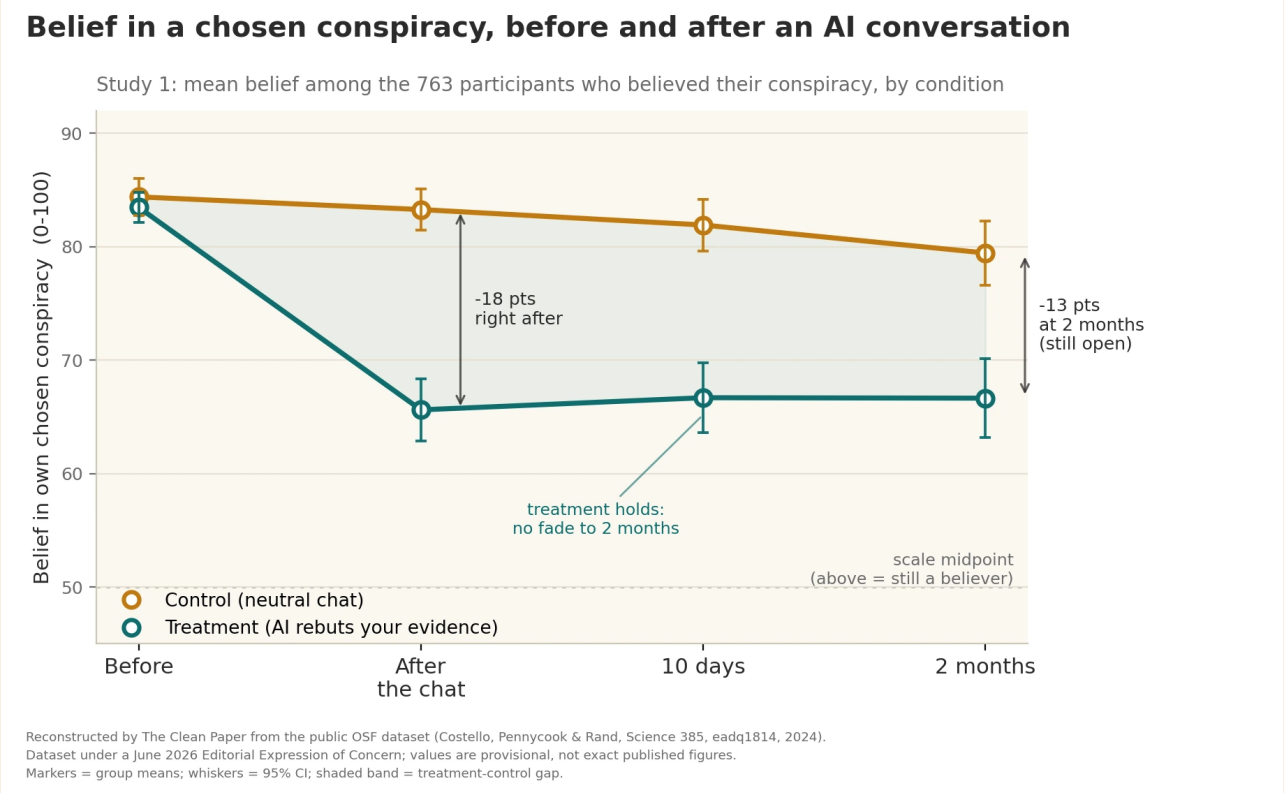
ところが 2026 年 6 月、*Science* はこの論文に **Editorial Expression of Concern（編集部による懸念表明）**を付けた。多くの語り直しで飛ばされそうな部分だが、現時点で結果にどれほどの重みを持たせてよいかを決めるのは、まさにここである。

Expression of Concern とは何か —— そして何ではないのか

Editorial Expression of Concern は、論文について問題が提起され、その検討がまだ続いている間に、読者へ注意を促すために学術誌が正式に付ける印である。**撤回（retraction）ではない**。論文は取り下げられておらず、それ自体が不正行為の認定でもない。意味するのは、「記録が今後変わる可能性があるので、この留保を念頭に読んでください」ということだ。今回、問題を知らされた著者らは調査を行い、具体的な内容を Science へ報告し、修正解析を提出した。

指摘された問題は具体的だ。参加者のスクリーニング基準が、論文本文と公開された解析コードで一貫して適用されていなかった。また、公開データセットにはコード結合時の誤りで余計な行が混入しており、公開資料だけから報告値の一部を再現しにくくなっていた。著者らは *Science* に修正済み解析パイプラインと更新結果を提出し、元の結果と**方向、統計的有意性、実質的な効果量**が一致すると述べている。*Science* は現在それを評価中だ。その評価が終わるまでは、正確な数字には疑問符が付き、その疑問符を外せるのは学術誌の審査だけである。

つまりこれは二つの物語が同時に進んでいる。事実が根深い信念を動かせるのかという興味深い研究結果と、科学記録が公開の場で自ら訂正されていく現在進行形の例だ。ここでは、その二つを混ぜないことが大切になる。



この研究では、参加者自身が挙げた根拠へ反論する 3 往復の AI 会話によって、本人が選んだ陰謀論への信念が平均約 20% 低下したと報告された。無関係な話題を扱う対照会話とは異なり、その低下は 10 日後と 2 か月後にも測定できた。効果は持続したが部分的で、会話後も大半の参加者は陰謀論を信じていた。治癒ではなく低下である。論文には現在、報告上の不整合と公開データセットの誤りを理由に Editorial Expression of Concern が付いている（*Science*, 2026 年 6 月）。著者らは修正解析でも効果の方向、有意性、大きさは保たれると述べ、*Science* が評価を続けている。この図は The Clean Paper が公開データセットから再構成したものなので、表示値は暫定値であり、論文の最終確定値ではない。

Original figure —The Clean Paper CC BY 4.0

著者らが行ったこと

- 二つの実験で 2190 人を対象にした。各参加者は、自分が信じる陰謀論と、それを支持すると考える証拠を自分の言葉で記述した。
- 各参加者は GPT-4 Turbo と 3 往復の会話を行った。**介入群**では AI が本人の具体的な証拠へ反論するよう促され、**対照群**では無関係な話題を議論した。
- 選んだ陰謀論への信念を会話前と直後に測定し、さらに **10 日後と 2 か月後**に追跡した。
- サンプル内で AI が行った 128 の事実主張すべてを、専門ファクトチェッカーに評価させた。
- 効果が他の無関係な陰謀論へ波及するか、また特異性の確認として、たまたま**真実である陰謀**への信念まで下げてしまうかを調べた。

分かったこと

- 介入群では対照群に比べ、選んだ陰謀論への信念が**平均およそ 20% 低下**した（研究 1：100 点尺度で 95% 信頼区間 [13.8, 19.7] ポイント、P < 0.001）。
- **持続した**。介入群では低下がすぐには戻らず、10 日後と 2 か月後でも会話直後に近い水準を保ち、対照群は一貫して高かった。論文は、選んだ陰謀論への効果が少なくとも 2 か月、弱まらず持続したと表現している。一瞬の揺らぎではない。
- **多様な陰謀論へ一般化した**。参加者が挙げたさまざまな陰謀論で見られ、当初の信念が強い人にも成立した。
- **この条件では AI の事実主張は高精度だった**。 128 件中 127 件（99.2%）が真、1 件（0.8%）が誤解を招く、偽は 0 だった。
- **単なる全面的な懐疑ではなく、特異的だった**。介入は実際に真である陰謀への信念を低下させなかった。すべてを疑うようにしたのではなく、裏付けの弱い信念を動かしたことを示唆する。
- **わずかな波及効果もあった**。他の無関係な陰謀論への信念も約 12% 下がり、参加者は陰謀論的主張に反論しようとする意図を強く報告した。主効果は研究 2 でも再現され、対照群のない小規模サンプルでも同方向だった（証拠は弱いが整合的）。

この研究からは言えないこと

- 現時点で**疑問なく確定した結果**ではない。論文はデータ処理と再現性の問題で Editorial Expression of Concern の対象となっており、修正後の数字はまだ *Science* が評価中だ。審査が終わるまでは具体的な値を暫定的なものとして扱うべきだ。
- AI が陰謀論を「**治す**」と示したわけではない。平均約 20% の低下は意味のある変化だが、消去ではない。大半の参加者はその後も陰謀論を信じており、程度が弱まっただけだ。
- **実社会への導入結果ではない**。参加者は研究へ自ら参加し、AI と誠実に対話した。現実では、陰謀論へのコミットメントが強い人ほど、自分に反論するチャットボットをわざわざ探す可能性は低い。
- 99.2% という精度は**この課題、このモデル、この慎重なプロンプト**に特有であり、言語モデルが一般に真実を述べる保証ではない。著者ら自身、同じ個別化された説得能力を逆方向に使えば、誤った信念を強めることもできると明記している。
- ここで「持続的」とは**2 か月**のことであり、一度の会話としては印象的だが、永続という意味ではない。

証拠をどう評価するか

- **研究設計は大きな強みだ**。介入対対照の統制実験、プラセボに近い明確な対照、二つの研究と独立サンプルでの再現、追跡による持続性確認、そして AI 自身の発言を明示的に事実確認した。説得研究としてかなり丁寧で、効果の方向は試されたところで一貫している。
- **しかし Expression of Concern は脚注ではない**。公開データとコードから報告値を再生成できることは、結果を信頼するための一部であり、そこがまさに壊れた。スクリーニング基準の不一致

と、コード結合エラーによる重複・余分な行があった。修正パイプラインでも結果が保たれるという著者らの説明は励みになり、もっともらしい。しかし現時点では著者ら自身の説明であり、独立した確定判断ではない。待つべきなのは *Science* の評価だ。
- **誠実な現在地は「警告付き」であって、「論破済み」でも「確定」でもない。**よく設計された印象的な研究だが、正確な数字について正式な審査が進んでいる。良い物語をこの居心地の悪い場所に置いておくのは気持ちよくないが、それが正確だ。

なぜ重要なのか

長年、陰謀論を信じる人は心理的欲求やアイデンティティによって動かされており、事実を示しても議論では変えられないという影響力のある見方があった。もしこの結果が確認されるなら、証拠に基づく持続的な信念低下は、その悲観論への意味ある反証になる。問題の一部は、一般的なデバンキングが、本人が実際に根拠として挙げる具体的な証拠へ応答してこなかったことかもしれない。LLM はそれを大規模に行える。

また、科学が本来どう機能すべきかの事例としても重要だ。Expression of Concern の教訓は、「有名な結果は価値がない」ではない。誤りが表に出て、データが再検討され、主張が公開の場で再確認される、ということだ。この仕組みが動くのはスキャンダルではなく機能である。だから、この結果への正しい姿勢は拍手でも切り捨てでもなく、評価が終わるのを待つことだ。

そして AI については両刃だ。誤った信念を個別化された議論で弱める能力は、逆向きに使えば同じくらい効果的に誤った信念を強められるかもしれない。技術そのものは中立だが、目的は中立ではない。

まとめ

2024 年の研究は、GPT-4 Turbo との 3 往復の会話によって、参加者が信じる陰謀論への信念が約 20% 低下し、その効果が 2 か月持続し、陰謀論の種類を越えて見られ、AI の事実主張も圧倒的に正確と評価されたと報告した。しかし 2026 年 6 月、データ処理と再現性の問題を理由に論文へ Editorial Expression of Concern が付けられた。著者らは修正解析でも結果の方向、有意性、大きさは保たれると報告し、*Science* は現在も評価を続けている。現時点では、本当に興味深く、丁寧に作られたが、正式審査中の結果として読むのがよい。確定でも、反証済みでもない。最も誠実な一文は、科学の過程そのものが書いている —— もう一度確かめる。

出典

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