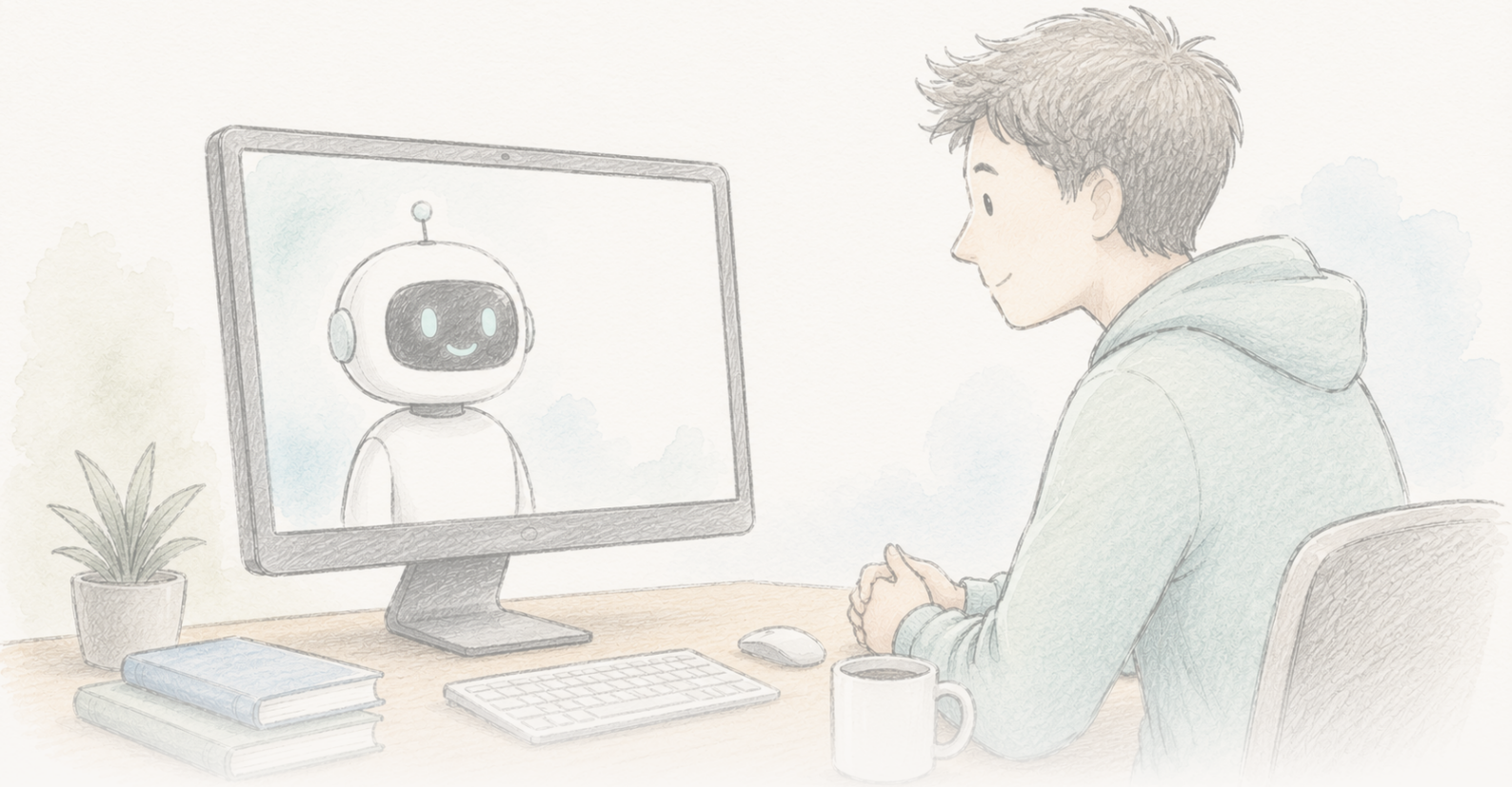


The Clean Paper

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

Ein Chatbot schien Verschwörungsüberzeugungen monatelang zu verringern

Eine Studie von 2024 fand, dass ein Gespräch über drei Runden mit GPT-4 Turbo den Glauben von Menschen an eine von ihnen vertretene Verschwörungstheorie um etwa 20 % verringerte — ein Effekt, der zwei Monate anhielt, über Verschwörungstypen hinweg generalisierte und auf KI-Behauptungen ruhte, die als weit überwiegend korrekt bewertet wurden. Im Juni 2026 wurde das Paper wegen Problemen mit Datenhandhabung und Reproduzierbarkeit unter eine Editorial Expression of Concern gestellt; die Autoren sagen, eine korrigierte Analyse erhalte den Befund in Richtung, Signifikanz und Größe, und Science prüft noch. Ein wirklich interessantes, sorgfältig gebautes Ergebnis — derzeit unter Begutachtung, weder erledigt noch widerlegt.

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 hat diesem Paper eine Editorial Expression of Concern beigefügt, während es Fragen zu Datenhandhabung und Reproduzierbarkeit prüft. Das ist keine Retraction und keine Feststellung von Fehlverhalten; es bedeutet, dass das berichtete Ergebnis als vorläufig zu lesen ist, bis die Prüfung abgeschlossen ist.

Editorial Expression of Concern →

Fakten wirken also doch — und ein Vermerk am Paper

2024 berichtete eine Studie etwas, das einer bequemen Alltagsweisheit zuwiderläuft. Man gebe jemandem ein kurzes Gespräch über drei Runden mit einer KI — GPT-4 Turbo, instruiert, auf die konkreten Belege einzugehen, die die Person für eine von ihr geglaubte Verschwörungstheorie anführt — und im Durchschnitt sinkt ihr Glaube an diese Theorie um etwa 20 %. Der Rückgang war zwei Monate später noch da. Es funktionierte bei klassischen Verschwörungen (die Ermordung John F. Kennedys, Außerirdische, die Illuminaten) ebenso wie bei aktuellen (COVID-19, die US-Wahl 2020), und selbst bei Menschen, deren Überzeugung stark und mit ihrer Identität verknüpft war. Als ein professioneller Faktenprüfer 128 der Behauptungen der KI bewertete, waren 127 wahr, eine irreführend und keine falsch.

Die Schlagzeile, die sich verbreitete, lautete: Man *kann* Menschen aus dem Kaninchenbau herausreden — Verschwörungsgläubige sind für Belege nicht unerreichbar, sie brauchen nur die richtigen Belege, spezifisch genug vorgetragen. Das ist eine wirklich interessante Behauptung, und die Studie war sorgfältiger gebaut als die meiste Persuasionsforschung.

Dann, im Juni 2026, versah *Science* das Paper mit einer **Editorial Expression of Concern** — einem redaktionellen Bedenkenvermerk. Das ist der Teil, den die meisten Nacherzählungen auslassen werden, und es ist genau der Teil, der entscheidet, wie viel Gewicht das Ergebnis im Moment tragen kann.

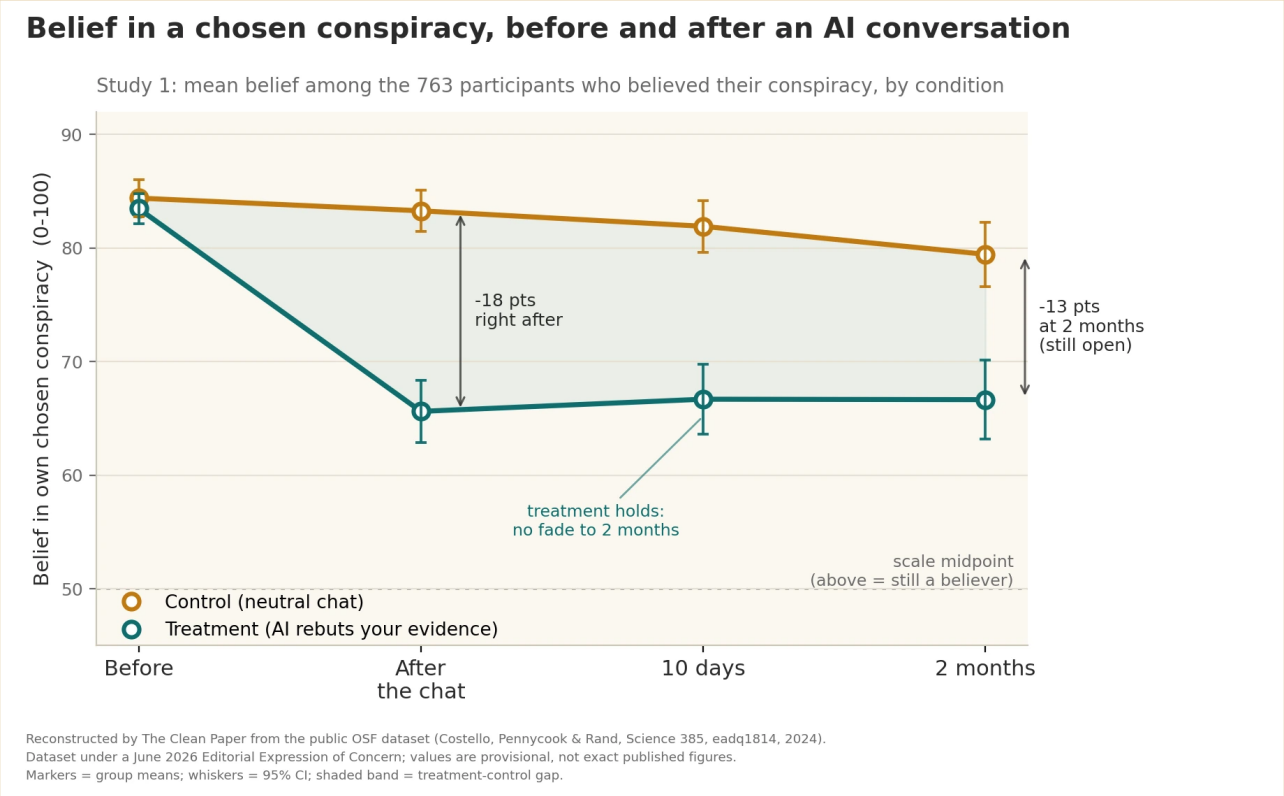
Was eine Expression of Concern ist — und was nicht

Eine Editorial Expression of Concern ist ein formaler Vermerk, den eine Fachzeitschrift an ein Paper heftet, um Lesende zu warnen, dass Fragen aufgeworfen wurden, während diese Fragen noch geklärt werden. Sie ist **keine** Retraction (das Paper wird nicht zurückgezogen), und sie ist für sich genommen keine Feststellung von Fehlverhalten. Sie bedeutet: Lies das mit dem Vorbehalt im Hinterkopf, denn der Stand kann sich ändern. Hier haben die Autoren, nachdem sie auf die Probleme aufmerksam gemacht worden waren, den Dingen nachgeforscht, die Einzelheiten an *Science* gemeldet und eine korrigierte Analyse nachgereicht.

Was beanstandet wurde, ist spezifisch. Die Autoren wurden auf Unstimmigkeiten dabei aufmerksam gemacht, wie die Kriterien zur Teilnehmerauswahl zwischen dem Manuskript und dem veröffentlichten Analysecode angewandt wurden, und darauf, dass der öffentliche Datensatz durch einen

Fehler beim Zusammenführen von Code eingeschleuste überzählige Zeilen enthielt — Probleme, die einige der berichteten Zahlen aus den freigegebenen Materialien schwer reproduzierbar machten. Die Autoren übergaben *Science* eine korrigierte Analyse-Pipeline und einen Satz aktualisierter Ergebnisse, die nach ihrer Aussage dem Original **in Richtung, statistischer Signifikanz und substanzieller Größe** entsprechen. *Science* prüft sie. Bis diese Prüfung abgeschlossen ist, stehen die genauen Zahlen unter einem Fragezeichen, das nur die Zeitschrift ausräumen kann.

Das sind also zwei Geschichten auf einmal: ein faszinierendes Ergebnis darüber, ob Fakten verfestigte Überzeugungen bewegen können, und ein lebendiges Beispiel dafür, wie sich der wissenschaftliche Bestand offen selbst korrigiert. Der Zweck dieses Textes ist, beide auseinanderzuhalten.



In der Studie wurde berichtet, dass ein KI-Gespräch über drei Runden, das die von der Person selbst genannten Belege widerlegte, den Glauben an die gewählte Verschwörung im Schnitt um etwa 20 % senkte; anders als bei einem Kontrollgespräch über ein themenfremdes Thema war der Rückgang nach 10 Tagen und 2 Monaten noch messbar. Der berichtete Effekt war anhaltend, aber partiell: Die meisten Teilnehmenden glaubten die Verschwörung danach weiterhin — eine Verringerung, keine Heilung. Das Paper steht derzeit unter einer Editorial Expression of Concern (*Science*, Juni 2026) wegen Unstimmigkeiten in der Berichterstattung und eines Fehlers im öffentlichen Datensatz; die Autoren geben an, eine korrigierte Analyse erhalte Richtung, Signifikanz und Größe des Effekts, während *Science* weiter prüft. Dieses Diagramm hat The Clean Paper aus dem öffentlichen Datensatz rekonstruiert; die gezeigten Werte sind daher vorläufig, nicht die endgültigen Zahlen des Papers.

Was die Autoren getan haben

- Sie führten zwei Experimente mit 2190 Teilnehmenden durch, von denen jede Person — in eigenen Worten — eine Verschwörungstheorie nannte, die sie glaubte, und die Belege, die sie dafür sah.
- Jede Person führte ein Gespräch über drei Runden mit GPT-4 Turbo. In der **Treatment**-Bedingung war die KI instruiert, die konkreten Belege der Person zu widerlegen; in der **Kontroll**-Bedingung sprach sie über ein themenfremdes Thema.
- Gemessen wurde der Glaube an die gewählte Verschwörung vor und unmittelbar nach dem Gespräch; danach wurden die Teilnehmenden nach **10 Tagen und 2 Monaten** erneut kontaktiert.
- Ein professioneller Faktenprüfer bewertete die Richtigkeit aller 128 Tatsachenbehauptungen, die die KI in einer Stichprobe machte.
- Geprüft wurde, ob der Effekt auf andere, nicht verwandte Verschwörungen übergriff — und, als Spezifitätstest, ob er auch den Glauben an Verschwörungen drückte, die tatsächlich **wahr** sind.

Was sie fanden

- **Eine durchschnittliche Verringerung um rund 20 %** des Glaubens an die gewählte Verschwörung in der Treatment-Gruppe gegenüber der Kontrolle (Studie 1: 95-%-Konfidenzintervall [13,8; 19,7] Punkte auf einer 100-Punkte-Skala, $P < 0{,}001$).
- **Es hielt an.** In der Treatment-Gruppe verblasste der Rückgang nicht: Der Glaube blieb nach 10 Tagen und zwei Monaten nahe dem Niveau von unmittelbar nach dem Gespräch, statt zurückzukriechen, während die Kontrolle durchgehend höher lag — das Paper beschreibt den Effekt auf die gewählte Verschwörung als über mindestens zwei Monate unvermindert anhaltend, nicht als kurzes Wackeln.
- **Es generalisierte** über die Bandbreite der genannten Verschwörungen hinweg und hielt auch bei Teilnehmenden, deren Überzeugung anfangs stark war.
- **Die Behauptungen der KI waren in diesem Setting korrekt**: 127 von 128 (99,2 %) als wahr bewertet, 1 (0,8 %) als irreführend, keine als falsch.
- **Es war spezifisch**, kein pauschaler Zweifel: Die Intervention verringerte den Glauben an wahre Verschwörungen **nicht** — ein Hinweis darauf, dass sie unbelegte Überzeugungen bewegte, statt Menschen gegenüber allem zynisch zu machen.
- **Es strahlte moderat aus** — der Glaube an andere, nicht verwandte Verschwörungen sank um rund 12 %, und die Teilnehmenden berichteten eine größere Absicht, Verschwörungsbehauptungen zu widersprechen. Der Haupteffekt hielt in Studie 2 und zeigte in einer kleineren Stichprobe ohne Kontrollgruppe in dieselbe Richtung (schwächere Evidenz, aber konsistent).

Was das nicht beweist

- Es steht im Moment **nicht** unangefochten da. Das Paper trägt eine Editorial Expression of Concern wegen Problemen mit Datenhandhabung und Reproduzierbarkeit, und die korrigierten Zahlen werden von *Science* noch geprüft. Die konkreten Werte sind als vorläufig zu behandeln, bis diese Prüfung vorliegt.
- Es zeigt **nicht**, dass KI Verschwörungsglauben „heilt". Eine durchschnittliche Verringerung um ~20 % ist eine bedeutsame Verschiebung, keine Auslöschung — die meisten Teilnehmenden glaubten die Theorie danach weiterhin, nur weniger.
- Es ist **kein** Ergebnis aus realem Einsatz. Die Teilnehmenden meldeten sich freiwillig zu einer Studie und ließen sich in gutem Glauben auf die KI ein; draußen in der Welt sind die am stärksten überzeugten Anhänger einer Verschwörung am wenigsten geneigt, einen Chatbot aufzusuchen, der ihnen widerspricht.
- Die 99,2 % Genauigkeit sind **spezifisch für diese Aufgabe, dieses Modell und dieses sorgfältige Prompting** — keine allgemeine Garantie, dass Sprachmodelle Wahres sagen. Die Autoren sagen ausdrücklich, dass dieselbe personalisierte Überzeugungskraft *falsche* Überzeugungen stärken könnte, würde sie darauf gerichtet.
- „Anhaltend" heißt hier **zwei Monate** — beeindruckend für ein einzelnes Gespräch, aber nicht dasselbe wie dauerhaft.

Wie belastbar sind die Belege?

- **Das Design ist eine echte Stärke.** Kontrollierte Treatment-gegen-Kontrolle-Experimente, eine saubere placeboartige Kontrolle, Replikation über zwei Studien und eine unabhängige Stichprobe, Follow-ups zur Dauerhaftigkeit und eine explizite Genauigkeitsprüfung der KI-Behauptungen — das ist sorgfältiger als die meiste Persuasionsforschung, und die Richtung des Effekts ist überall konsistent, wo sie getestet wurde.
- **Aber die Expression of Concern ist keine Fußnote.** Dass sich die berichteten Zahlen aus den freigegebenen Daten und dem Code regenerieren lassen, gehört zu dem, was ein Ergebnis vertrauenswürdig macht — und genau das ist gebrochen: Unstimmigkeiten bei den Auswahlkriterien und duplizierte Zeilen aus einem Code-Merging-Fehler. Die Aussage der Autoren, eine korrigierte Pipeline erhalte das Ergebnis, ist ermutigend und plausibel, aber sie ist die eigene Darstellung der Autoren, noch kein unabhängiges Urteil. Die Prüfung durch *Science* ist das, worauf zu warten ist.
- **Der ehrliche Status ist „mit Vermerk versehen", nicht „widerlegt" und nicht „bestätigt".** Eine gut gebaute, bemerkenswerte Studie, deren genaue Zahlen unter formaler Prüfung stehen. Das ist ein unbequemer Ort, an dem man eine gute Geschichte stehen lässt — und der zutreffende.

Warum das wichtig ist

Jahrelang galt die einflussreiche Ansicht, Verschwörungsgläubige würden von psychologischen Bedürfnissen und Identität getrieben und ließen sich deshalb nicht mit Fakten aus ihren Überzeugungen argumentieren. Eine anhaltende, belegbasierte Verringerung — falls sie sich hält — ist ein gewichtiges Gegengewicht zu diesem Pessimismus: Sie legt nahe, dass das Problem zum Teil darin lag, dass generisches Debunking nie auf die *konkreten* Belege einging, die ein Gläubiger tatsächlich anführt — etwas, das ein Sprachmodell in großem Maßstab leisten kann.

Es zählt auch als Fallstudie dafür, wie Wissenschaft funktionieren soll. Die Lehre aus der Expression of Concern ist nicht, dass gefeierte Ergebnisse wertlos sind; sie ist, dass Fehler ans Licht kommen, Daten erneut geprüft und Behauptungen öffentlich nachkontrolliert werden. Dass diese Maschinerie läuft, ist ein Merkmal, kein Skandal — und der Grund, warum die richtige Haltung zu diesem Ergebnis Geduld ist, nicht Applaus und nicht Abtun.

Und es schneidet bei KI in beide Richtungen. Dieselbe Fähigkeit, eine falsche Überzeugung mit einem maßgeschneiderten Argument zu entkräften, könnte eine ebenso wirksam aufblähen. Die Technik ist neutral; nur das Ziel ist es nicht.

Kurz zusammengefasst

Eine Studie von 2024 fand, dass ein Gespräch über drei Runden mit GPT-4 Turbo den Glauben von Menschen an eine von ihnen vertretene Verschwörungstheorie um etwa 20 % verringerte — ein Effekt, der zwei Monate anhielt, über Verschwörungstypen hinweg generalisierte und auf KI-Behauptungen ruhte, die als weit überwiegend korrekt bewertet wurden. Im Juni 2026 wurde das Paper wegen Problemen mit Datenhandhabung und Reproduzierbarkeit unter eine Editorial Expression of Concern gestellt; die Autoren berichten, eine korrigierte Analyse erhalte den Befund in Richtung, Signifikanz und Größe, und *Science* prüft noch. Zu lesen ist es als ein wirklich interessantes, sorgfältig gebautes Ergebnis, das derzeit unter Begutachtung steht — nicht erledigt, nicht widerlegt. Den ehrlichsten Satz darüber schreibt der Prozess selbst gerade: Prüf es noch einmal.

Quellen

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