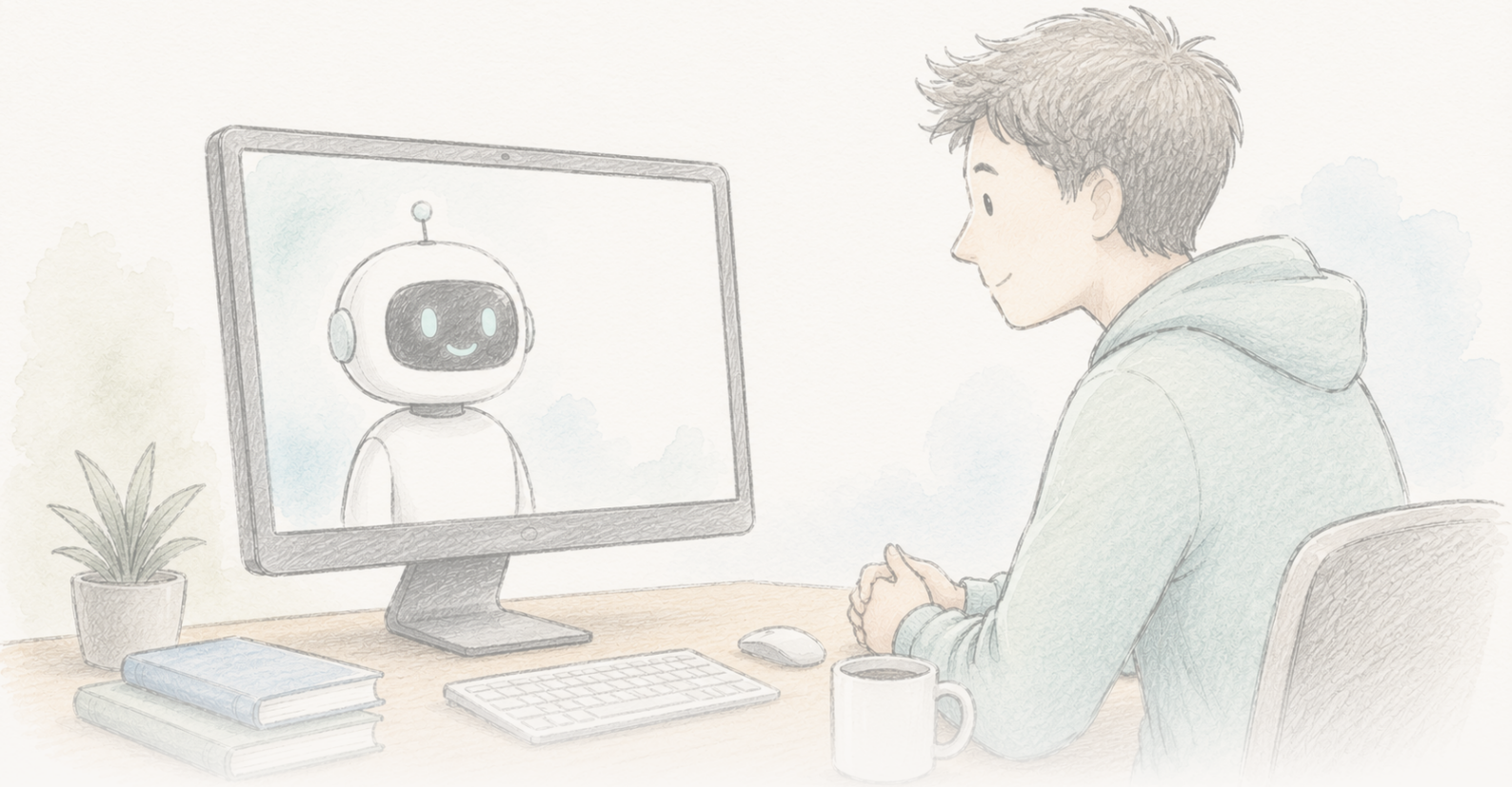


The Clean Paper

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

En chatbot så ud til at mindske troen på konspirationsteorier i flere måneder

Et studie fra 2024 fandt, at en samtale i tre runder med GPT-4 Turbo mindskede menneskers tro på en konspirationsteori, de selv troede på, med omkring 20 procent. Effekten varede i to måneder, gjaldt på tværs af forskellige typer konspirationsteorier og byggede på AI-påstande, der blev vurderet som overvejende korrekte. I juni 2026 blev artiklen forsynet med en Editorial Expression of Concern på grund af problemer med datahåndtering og reproducerbarhed. Forfatterne siger, at en korrigeret analyse bevarer resultatets retning, signifikans og størrelse, og Science vurderer stadig materialet. Et oprigtigt interessant og omhyggeligt udformet resultat — nu under vurdering, hverken afgjort eller afkræftet.

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 har forsynet artiklen med en Editorial Expression of Concern, mens tidsskriftet vurderer spørgsmål om datahåndtering og reproducerbarhed. Det er ikke en tilbagetrækning og ikke en konstatering af uredelighed; det betyder, at det rapporterede resultat bør betragtes som foreløbigt, indtil vurderingen er afsluttet.

Editorial Expression of Concern →

Fakta trods alt — og en advarsel på studiet

I 2024 rapporterede et studie noget, der går imod en bekvem, konventionel antagelse. Lad en person føre en kort samtale i tre runder med en AI — GPT-4 Turbo, instrueret i at svare på de konkrete belæg, personen anfører for en konspirationsteori, vedkommende selv tror på — og i gennemsnit falder troen på teorien med omkring 20 procent. Faldet var stadig til stede to måneder senere. Det virkede for klassiske konspirationsteorier (mordet på John F. Kennedy, rumvæsener, Illuminati) og aktuelle teorier (covid-19, det amerikanske valg i 2020), selv blandt personer hvis overbevisning var stærk og knyttet til deres identitet. Da en professionel faktatjekker vurderede 128 påstande fra AI'en, var 127 sande, én vildledende og ingen falske.

Overskriften, der spredte sig, var, at man faktisk *kan* tale folk ud af kaninhullet — at mennesker, der tror på konspirationsteorier, ikke er uden for belæggenes rækkevidde, men blot har brug for de rigtige belæg, fremlagt tilstrækkeligt konkret. Det er en oprigtigt interessant påstand, og studiet var mere omhyggeligt udformet end det meste forskning i overtalelse.

Men i juni 2026 forsynede *Science* artiklen med en **Editorial Expression of Concern**. Det er den del, de fleste genfortællinger vil springe over, og netop den del, der afgør, hvor meget vægt resultatet kan bære lige nu.

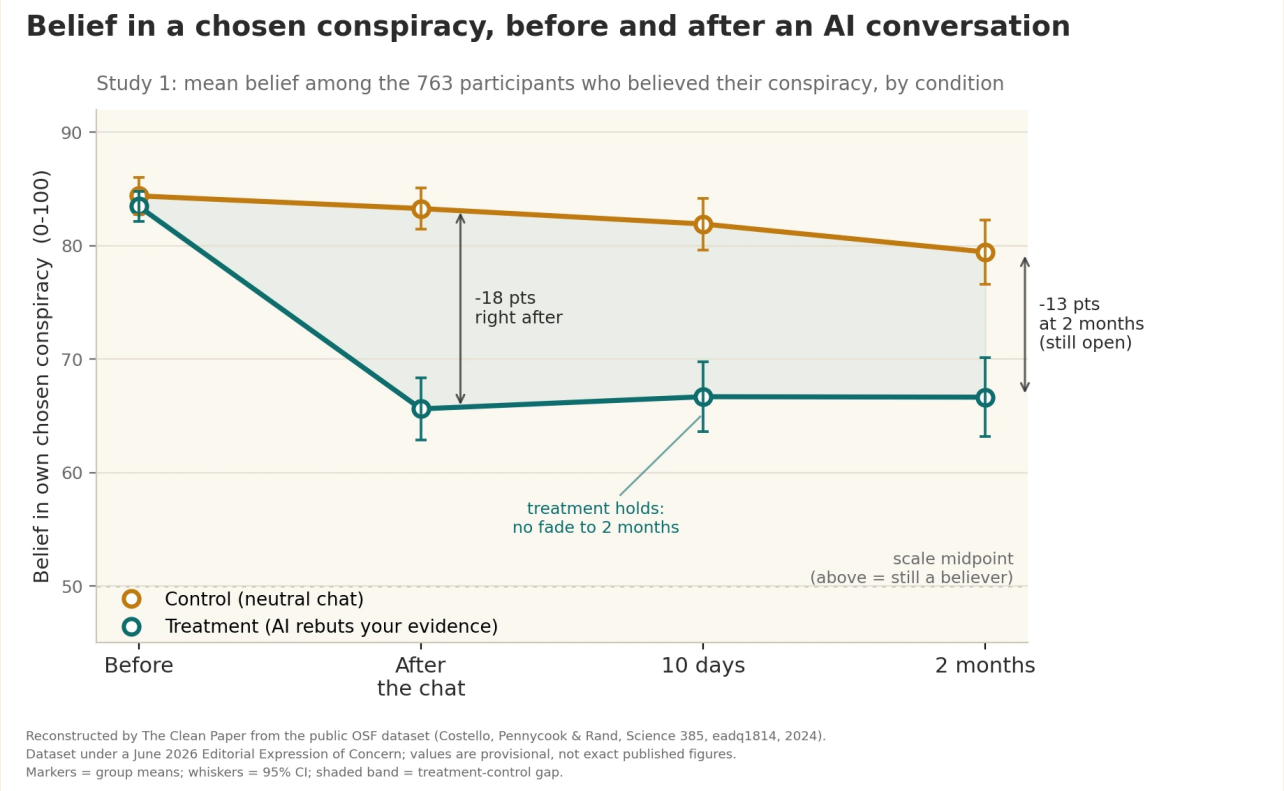
Hvad en Expression of Concern er — og ikke er

En Editorial Expression of Concern er en formel advarsel, som et tidsskrift knytter til en artikel for at gøre læserne opmærksomme på, at der er rejst spørgsmål, mens de stadig undersøges. Det er **ikke** en tilbagetrækning (artiklen er ikke trukket tilbage), og det er ikke i sig selv en konstatering af videnskabelig uredelighed. Det betyder: læs med dette forbehold i tankerne, fordi forskningsgrundlaget kan ændre sig. Efter at forfatterne her blev gjort opmærksomme på problemerne, undersøgte de dem, indberettede detaljerne til *Science* og indsendte en korrigeret analyse.

Det, der blev markeret, er konkret. Forfatterne blev gjort opmærksomme på uoverensstemmelser mellem manuskriptet og den offentliggjorte analysekode i anvendelsen af kriterierne for screening af deltagere. Det offentlige datasæt indeholdt desuden uvedkommende rækker, som var blevet indsat ved en fejl under sammenfletning af kode. Problemerne gjorde nogle af de rapporterede tal vanskelige at reproducere ud fra det offentliggjorte materiale. Forfatterne gav *Science* en korrigeret

analysepipeline og opdaterede resultater, som de siger stemmer overens med originalen **i retning, statistisk signifikans og reel effektstørrelse**. *Science* vurderer materialet. Indtil vurderingen er afsluttet, står de nøjagtige tal under et spørgsmålstegn, som kun tidsskriftet kan fjerne.

Der er altså tale om to historier på én gang: et spændende resultat om, hvorvidt fakta kan flytte fastgroede overbevisninger, og et levende eksempel på, at den videnskabelige litteratur korrigerer sig selv offentligt. Formålet med denne artikel er at holde de to historier adskilt.



I studiet blev en AI-samtale i tre runder, der imødegik deltagerens egne angivne belæg, rapporteret at mindske troen på personens valgte konspirationsteori med omkring 20 procent i gennemsnit. I modsætning til en kontrolsamtale om et uvedkommende emne kunne faldet stadig måles efter 10 dage og 2 måneder. Den rapporterede effekt var vedvarende, men delvis: de fleste deltagere troede stadig på konspirationsteorien bagefter — en reduktion, ikke en kur. Artiklen er i øjeblikket forsynet med en Editorial Expression of Concern (*Science*, juni 2026) på grund af uoverensstemmelser i rapporteringen og en fejl i det offentlige datasæt. Forfatterne siger, at en korrigeret analyse bevarer effektens retning, signifikans og størrelse, mens *Science* fortsætter sin vurdering. Diagrammet er rekonstrueret af The Clean Paper ud fra det offentlige datasæt, så de viste værdier er foreløbige og ikke artiklens endelige tal.

Hvad forfatterne gjorde

- Gennemførte to eksperimenter med 2.190 deltagere, som med egne ord angav en konspirationsteori, de troede på, og de belæg, de mente støttede den.

- Lod hver deltager føre en samtale i tre runder med GPT-4 Turbo. I **behandlingsgruppen** blev AI'en instrueret i at imødegå deltagerens konkrete belæg; i **kontrolgruppen** diskuterede den et emne uden forbindelse til konspirationsteorien.
- Målte troen på den valgte konspirationsteori før og umiddelbart efter samtalen og kontaktede derefter deltagerne igen efter **10 dage og 2 måneder**.
- Lod en professionel faktatjekker vurdere nøjagtigheden af alle 128 faktuelle påstande, som AI'en fremsatte i et udvalg af samtalerne.
- Undersøgte, om effekten bredte sig til andre, uvedkommende konspirationsteorier — og, som en test af specificitet, om den også mindskede troen på konspirationer, der faktisk er **sande**.

Hvad de fandt

- **En gennemsnitlig reduktion på omkring 20 procent** i troen på den valgte konspirationsteori i behandlingsgruppen sammenlignet med kontrolgruppen (studie 1: 95-procents konfidensinterval [13,8; 19,7] point på en 100-punktsskala, $P < 0{,}001$).
- **Effekten holdt ved.** I behandlingsgruppen forsvandt faldet ikke: efter 10 dage og to måneder lå troen fortsat tæt på niveauet umiddelbart efter samtalen i stedet for at stige igen, mens kontrolgruppen lå højere hele vejen. Artiklen beskriver effekten på den valgte konspirationsteori som uformindsket i mindst to måneder, ikke som et kortvarigt udsving.
- **Effekten gjaldt på tværs af** de konspirationsteorier, deltagerne angav, og holdt også blandt deltagere, hvis overbevisning oprindeligt var stærk.
- **AI'ens påstande var korrekte** i denne sammenhæng: 127 af 128 (99,2 procent) blev vurderet som sande, 1 (0,8 procent) som vildledende og ingen som falske.
- **Effekten var specifik**, ikke udtryk for generel tvivl: interventionen mindskede **ikke** troen på virkelige konspirationer, hvilket tyder på, at den flyttede uunderbyggede overbevisninger frem for at gøre folk kyniske over for alt.
- **Effekten bredte sig i beskedent omfang** — troen på andre, uvedkommende konspirationsteorier faldt med omkring 12 procent, og deltagerne angav en større hensigt om at sige imod konspiratoriske påstande. Hovedeffekten holdt i studie 2 og pegede i samme retning i en mindre stikprøve uden kontrolgruppe (svagere evidens, men samstemmende).

Hvad dette ikke beviser

- Resultatet står **ikke** uimodsagt lige nu. Artiklen er forsynet med en Editorial Expression of Concern på grund af problemer med datahåndtering og reproducerbarhed, og de korrigerede tal vurderes stadig af *Science*. Betragt de konkrete værdier som foreløbige, indtil vurderingen er afsluttet.
- Studiet viser **ikke**, at AI »kurerer« tro på konspirationsteorier. En gennemsnitlig reduktion på omkring 20 procent er en meningsfuld forskydning, ikke en udslettelse — de fleste deltagere troede stadig på teorien bagefter, blot mindre.
- Det er **ikke** et resultat fra brug i den virkelige verden. Deltagerne meldte sig til et studie og gik ind i samtalen med AI'en i god tro. Uden for studiet er de mest overbeviste tilhængere af en

konspiration også de mindst tilbøjelige til at opsøge en chatbot, der vil argumentere imod dem.
- Nøjagtigheden på 99,2 procent er **specifik for denne opgave, denne model og den omhyggelige udformning af instruktionerne** — ikke en generel garanti for, at sprogmodeller fremsætter sande påstande. Forfatterne understreger, at den samme personligt tilpassede overtalelsesevne kunne fremme *falske* overbevisninger, hvis den blev rettet mod det.
- »Vedvarende« betyder her **to måneder**, hvilket er imponerende for en enkelt samtale, men ikke det samme som permanent.

Hvor stærk er evidensen?

- **Studiets design er en reel styrke.** Kontrollerede eksperimenter med behandlings- og kontrolgruppe, en klar placebolignende kontrol, gentagelse i to studier og en uafhængig stikprøve, opfølgning på varigheden samt en udtrykkelig kontrol af rigtigheden af AI'ens egne påstande — dette er mere omhyggeligt end det meste forskning i overtalelse, og effektens retning er konsistent overalt, hvor den blev testet.
- **Men en Expression of Concern er ikke en fodnote.** Muligheden for at genskabe de rapporterede tal ud fra offentliggjorte data og kode er en del af det, der gør et resultat troværdigt, og det var netop her, det brød sammen: inkonsistente screeningskriterier og duplikerede rækker efter en fejl ved sammenfletning af kode. Forfatternes oplysning om, at en korrigeret pipeline bevarer resultatet, er opmuntrende og plausibel, men det er deres egen redegørelse, endnu ikke en uafhængig konklusion. Det er *Science*s vurdering, man må vente på.
- **Den ærlige status er »markeret«, ikke »afkræftet« eller »bekræftet«.** Et veludformet og opsigtsvækkende studie, hvis nøjagtige tal er under formel vurdering. Det er et ubekvemt sted at efterlade en god historie, og det er det korrekte.

Hvorfor det er vigtigt

I årevis har en indflydelsesrig opfattelse været, at mennesker, der tror på konspirationsteorier, drives af psykologiske behov og identitet og derfor ikke kan argumenteres ud af deres overbevisninger med fakta. En vedvarende, evidensbaseret reduktion — hvis resultatet holder — er en væsentlig modvægt til denne pessimisme. Den antyder, at en del af problemet var, at generel faktatjek aldrig tog fat på de *konkrete* belæg, som en person faktisk henviser til, noget en stor sprogmodel kan gøre i stor skala.

Studiet er også vigtigt som et eksempel på, hvordan videnskab er beregnet til at fungere. Læren af en Expression of Concern er ikke, at berømte resultater er værdiløse, men at fejl kommer frem, data undersøges på ny, og påstande efterprøves offentligt. At dette maskineri arbejder, er en styrke, ikke en skandale — og derfor er tålmodighed den rette holdning til resultatet frem for enten bifald eller afvisning.

Og AI-spørgsmålet går begge veje. Den samme evne til at svække en falsk overbevisning med et skræddersyet argument kunne lige så effektivt styrke en. Teknikken er neutral; kun formålet er det ikke.

Kort sammenfatning

Et studie fra 2024 fandt, at en samtale i tre runder med GPT-4 Turbo mindskede menneskers tro på en konspirationsteori, de selv troede på, med omkring 20 procent. Effekten varede i to måneder og gjaldt på tværs af forskellige typer konspirationsteorier, mens AI'ens faktuelle påstande blev vurderet som overvejende korrekte. I juni 2026 blev artiklen forsynet med en Editorial Expression of Concern på grund af problemer med datahåndtering og reproducerbarhed. Forfatterne oplyser, at en korrigeret analyse bevarer resultatets retning, signifikans og størrelse, og *Science* vurderer stadig materialet. Læs den som et oprigtigt interessant og omhyggeligt udformet resultat, der i øjeblikket er under vurdering — hverken afgjort eller afkræftet. Den mest ærlige sætning om studiet er den, processen selv skriver: kontrollér det igen.

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