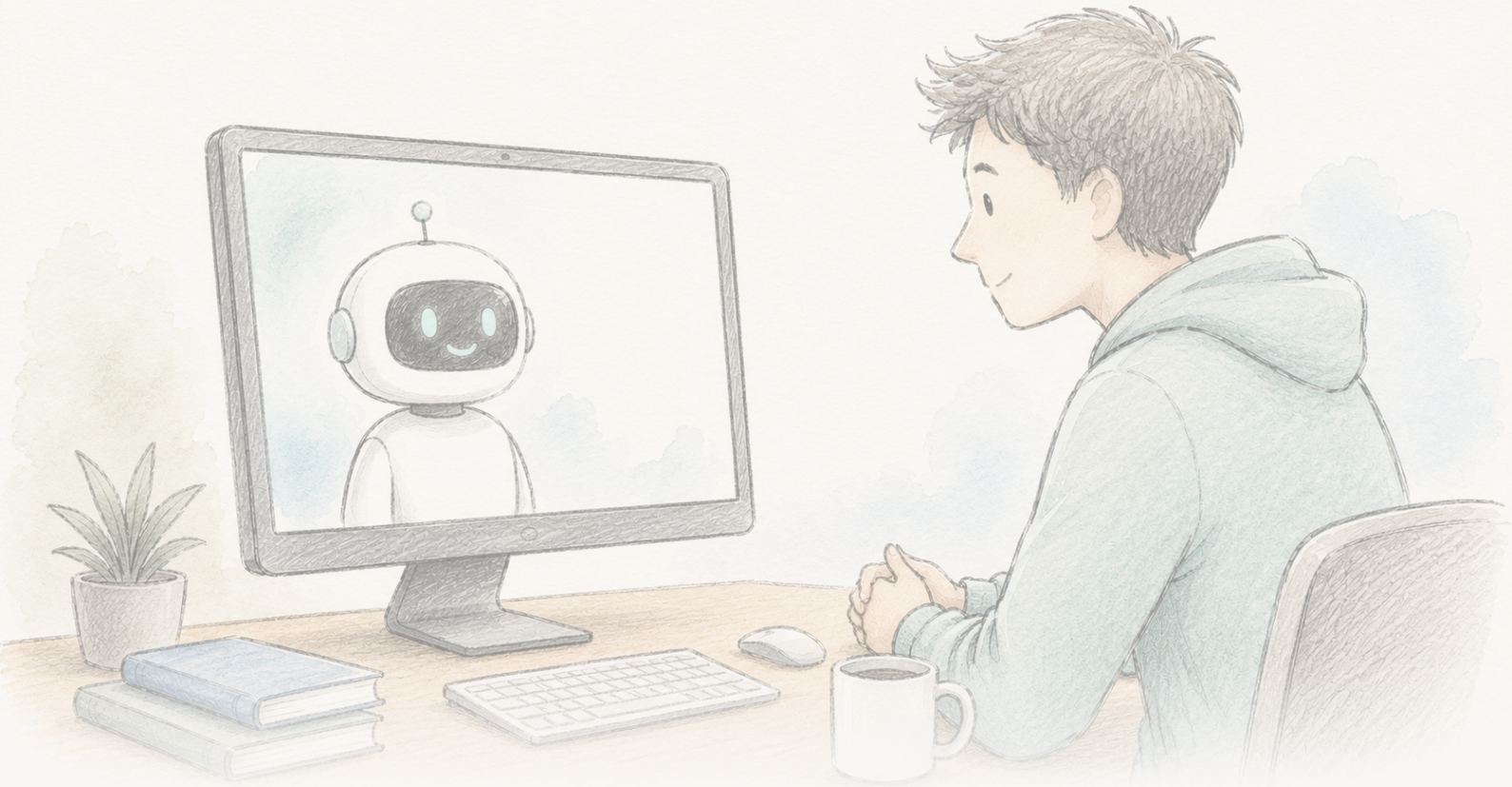


The Clean Paper

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

En chattbot tycktes minska tron på konspirationsteorier i flera månader

En studie från 2024 fann att ett samtal i tre omgångar med GPT-4 Turbo minskade människors tro på en konspirationsteori de omfattade med omkring 20 procent. Effekten bestod i två månader, återfanns över olika typer av konspirationsteorier och byggde på AI-påståenden som bedömdes som överväldigande korrekta. I juni 2026 försågs artikeln med en Editorial Expression of Concern på grund av problem med datahantering och reproducerbarhet. Författarna uppger att en korrigerad analys bevarar resultatets riktning, signifikans och storlek, och Science utvärderar fortfarande materialet. Ett genuint intressant och omsorgsfullt utformat resultat — nu under granskning, varken fastslaget eller motbevisat.

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 har försett artikeln med en Editorial Expression of Concern medan tidskriften utvärderar frågor om datahantering och reproducerbarhet. Det är inte en tillbakadragning och inte ett konstaterande av oredlighet; det betyder att det rapporterade resultatet bör betraktas som preliminärt tills granskningen är avslutad.

Editorial Expression of Concern →

Fakta trots allt — och en varningsflagga på studien

År 2024 rapporterade en studie något som går emot en bekväm konventionell föreställning. Låt någon föra ett kort samtal i tre omgångar med en AI — GPT-4 Turbo, instruerad att bemöta de specifika belägg personen anför för en konspirationsteori som hen själv tror på — och i genomsnitt minskar tron på teorin med omkring 20 procent. Minskningen fanns kvar två månader senare. Det fungerade för klassiska konspirationsteorier (mordet på John F. Kennedy, utomjordingar, Illuminati) liksom för aktuella sådana (covid-19, det amerikanska valet 2020), och även bland personer vars övertygelse var stark och knuten till identiteten. När en professionell faktagranskare bedömde 128 påståenden från AI:n var 127 sanna, ett vilseledande och inget falskt.

Rubriken som spreds var att det faktiskt *går* att prata människor ur kaninhålet — att personer som tror på konspirationsteorier inte är utom räckhåll för belägg, utan bara behöver rätt belägg, tillräckligt specifikt presenterade. Det är ett genuint intressant påstående, och studien var mer omsorgsfullt utformad än merparten av forskningen om övertalning.

Men i juni 2026 försåg *Science* artikeln med en **Editorial Expression of Concern**. Det är den del som de flesta återberättelser lär hoppa över, och samtidigt just den del som avgör hur stor tyngd resultatet kan bära just nu.

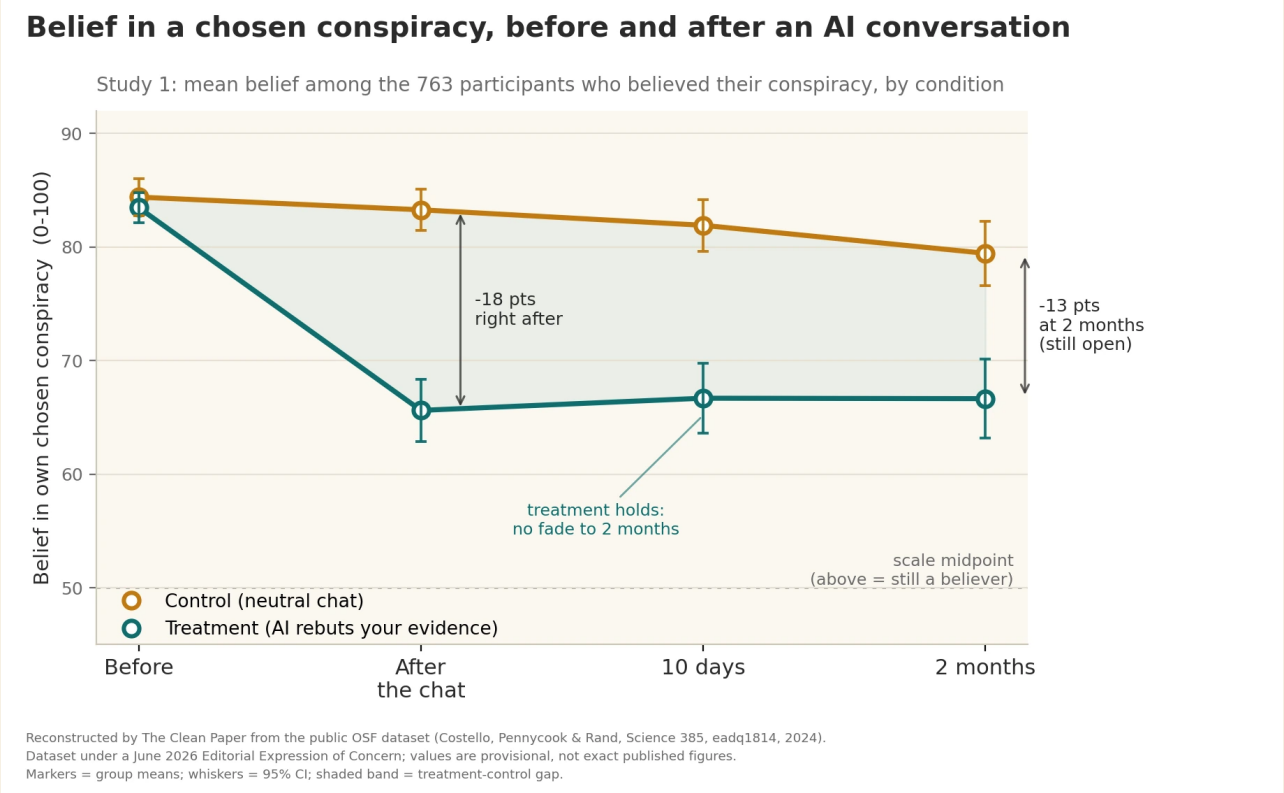
Vad en Expression of Concern är — och inte är

En Editorial Expression of Concern är en formell varningsflagga som en tidskrift fäster vid en artikel för att uppmärksamma läsarna på att frågor har väckts medan de fortfarande utreds. Det är **inte** en tillbakadragning (artikeln har inte dragits tillbaka), och det är inte i sig ett konstaterande av oredlighet. Budskapet är: läs med detta förbehåll i åtanke, eftersom forskningsunderlaget kan komma att ändras. När författarna här fick kännedom om problemen utredde de dem, redovisade detaljerna för *Science* och lämnade in en korrigerad analys.

Det som flaggades är specifikt. Författarna uppmärksammades på inkonsekvenser mellan manuskriptet och den publicerade analyskoden i hur kriterierna för urval av deltagare hade tillämpats. Den offentliga datamängden innehöll dessutom ovidkommande rader som hade fogats in genom ett fel när kod slogs samman. Problemen gjorde vissa av de rapporterade siffrorna svåra att återskapa från det publicerade materialet. Författarna gav *Science* en korrigerad analyspipeline och uppdaterade resultat som de uppger överensstämmer med originalet **i riktning, statistisk signifikans och faktisk**

effektstorlek. *Science* utvärderar materialet. Tills den granskningen är klar står de exakta siffrorna under ett frågetecken som bara tidskriften kan undanröja.

Det här är alltså två berättelser på samma gång: ett tankeväckande resultat om huruvida fakta kan rubba djupt rotade föreställningar och ett pågående exempel på hur det vetenskapliga underlaget korrigerar sig självt offentligt. Syftet med den här artikeln är att hålla isär dem.



I studien rapporterades ett AI-samtal i tre omgångar, där deltagarens egna angivna belägg bemöttes, minska tron på personens valda konspirationsteori med omkring 20 procent i genomsnitt. Till skillnad från ett kontrollsamtal om ett orelaterat ämne var minskningen fortfarande mätbar efter 10 dagar och 2 månader. Den rapporterade effekten var varaktig men partiell: de flesta deltagare trodde fortfarande på konspirationsteorin efteråt — en minskning, inte ett botemedel. Artikeln är för närvarande försedd med en Editorial Expression of Concern (*Science*, juni 2026) på grund av inkonsekvenser i rapporteringen och ett fel i den offentliga datamängden. Författarna uppger att en korrigerad analys bevarar effektens riktning, signifikans och storlek, medan *Science* fortsätter sin utvärdering. Diagrammet har rekonstruerats av The Clean Paper från den offentliga datamängden, så de visade värdena är preliminära och inte artikelns slutliga siffror.

Vad författarna gjorde

- Genomförde två experiment med 2 190 deltagare, som med egna ord angav en konspirationsteori de trodde på och de belägg som de ansåg stödde den.
- Lät varje deltagare föra ett samtal i tre omgångar med GPT-4 Turbo. I **behandlingsgruppen** in-

struerades AI:n att bemöta deltagarens specifika belägg; i **kontrollgruppen** diskuterade den ett orelaterat ämne.
- Mätte tron på den valda konspirationsteorin före och omedelbart efter samtalet och kontaktade sedan deltagarna igen efter **10 dagar och 2 månader**.
- Lät en professionell faktagranskare bedöma riktigheten i samtliga 128 faktapåståenden som AI:n gjorde i ett urval av samtalen.
- Undersökte om effekten spred sig till andra, orelaterade konspirationsteorier — och, som ett test av specificitet, om den även minskade tron på konspirationer som faktiskt är **sanna**.

Vad de fann

- **En genomsnittlig minskning på omkring 20 procent** i tron på den valda konspirationsteorin i behandlingsgruppen jämfört med kontrollgruppen (studie 1: 95-procentigt konfidensintervall [13,8; 19,7] poäng på en 100-gradig skala, $P < 0{,}001$).
- **Effekten bestod.** I behandlingsgruppen avtog inte minskningen: efter 10 dagar och två månader låg tron kvar nära nivån direkt efter samtalet i stället för att stiga igen, medan kontrollgruppen låg högre hela tiden. Artikeln beskriver effekten på den valda konspirationsteorin som oförminskad i minst två månader, inte som en tillfällig svacka.
- **Effekten generaliserades** över de olika konspirationsteorier som deltagarna angav och fanns kvar även bland deltagare vars ursprungliga övertygelse var stark.
- **AI:ns påståenden var korrekta** i den här situationen: 127 av 128 (99,2 procent) bedömdes som sanna, 1 (0,8 procent) som vilseledande och inget som falskt.
- **Effekten var specifik**, inte ett uttryck för allmänt tvivel: interventionen minskade **inte** tron på verkliga konspirationer, vilket tyder på att den påverkade ogrundade föreställningar snarare än gjorde människor cyniska till allt.
- **Effekten spred sig i måttlig grad** — tron på andra, orelaterade konspirationsteorier minskade med omkring 12 procent, och deltagarna uppgav en större avsikt att säga emot konspiratoriska påståenden. Huvudeffekten kvarstod i studie 2 och pekade i samma riktning i ett mindre urval utan kontrollgrupp (svagare evidens, men samstämmig).

Vad detta inte bevisar

- Resultatet står **inte** oemotsagt just nu. Artikeln har fått en Editorial Expression of Concern på grund av problem med datahantering och reproducerbarhet, och de korrigerade siffrorna utvärderas fortfarande av *Science*. Betrakta de specifika värdena som preliminära tills granskningen är klar.
- Studien visar **inte** att AI ”botar” tro på konspirationsteorier. En genomsnittlig minskning på omkring 20 procent är en betydande förskjutning, inte ett utplånande — de flesta deltagare trodde fortfarande på teorin efteråt, men mindre starkt.
- Det är **inte** ett resultat från användning i verkligheten. Deltagarna valde att delta i en studie och samtalade med AI:n i god tro. Utanför studien är de som är mest övertygade om en konspiration

också minst benägna att söka upp en chattbot som argumenterar emot dem.
- Träffsäkerheten på 99,2 procent är **specifik för denna uppgift, denna modell och den omsorgsfulla utformningen av instruktionerna** — inte en allmän garanti för att språkmodeller framför sanna påståenden. Författarna är tydliga med att samma personligt anpassade övertalningsförmåga skulle kunna främja *falska* föreställningar om den riktades åt det hållet.
- ”Varaktig” betyder här **två månader**, vilket är imponerande för ett enda samtal men inte detsamma som permanent.

Hur starka är beläggen

- **Studiens upplägg är en verklig styrka.** Kontrollerade experiment med behandlings- och kontrollgrupp, en tydlig placeboliknande kontroll, upprepning i två studier och ett oberoende urval, uppföljningar av varaktigheten samt en uttrycklig kontroll av riktigheten i AI:ns egna påståenden — detta är mer omsorgsfullt än merparten av forskningen om övertalning, och effektens riktning är samstämmig överallt där den testades.
- **Men en Expression of Concern är ingen fotnot.** Att kunna återskapa de rapporterade siffrorna från publicerade data och publicerad kod är en del av det som gör ett resultat trovärdigt, och det var just där det brast: inkonsekventa urvalskriterier och dubblerade rader efter ett fel vid sammanslagning av kod. Författarnas uppgift att en korrigerad pipeline bevarar resultatet är uppmuntrande och rimlig, men det är deras egen redogörelse, ännu inte ett oberoende utlåtande. Det är *Science*s utvärdering som måste inväntas.
- **Den uppriktiga statusen är ”flaggad”, inte ”motbevisad” eller ”bekräftad”.** En väl utformad och slående studie vars exakta siffror granskas formellt. Det är ett obekvämt ställe att lämna en bra berättelse på, och det är det korrekta.

Varför det spelar roll

I åratal har en inflytelserik uppfattning varit att människor som tror på konspirationsteorier drivs av psykologiska behov och identitet och därför inte kan övertygas med fakta. En varaktig, evidensbaserad minskning — om resultatet håller — är en viktig motvikt till denna pessimism. Den antyder att problemet delvis var att generell faktagranskning aldrig tog sig an de *specifika* belägg som en person faktiskt åberopar, något en stor språkmodell kan göra i stor skala.

Studien är också viktig som fallstudie i hur vetenskap är tänkt att fungera. Lärdomen av en Expression of Concern är inte att uppmärksammade resultat är värdelösa, utan att fel kommer fram, data granskas på nytt och påståenden kontrolleras igen offentligt. Att detta maskineri arbetar är en styrka, inte en skandal — och därför är tålamod rätt hållning till resultatet, snarare än vare sig applåder eller avfärdande.

Och AI-frågan har två sidor. Samma förmåga att försvaga en falsk föreställning med ett skräddarsytt argument skulle lika effektivt kunna förstärka en. Tekniken är neutral; bara syftet är det inte.

Kort sammanfattning

En studie från 2024 fann att ett samtal i tre omgångar med GPT-4 Turbo minskade människors tro på en konspirationsteori de omfattade med omkring 20 procent. Effekten bestod i två månader och återfanns över olika typer av konspirationsteorier, samtidigt som AI:ns faktapåståenden bedömdes som överväldigande korrekta. I juni 2026 försågs artikeln med en Editorial Expression of Concern på grund av problem med datahantering och reproducerbarhet. Författarna uppger att en korrigerad analys bevarar resultatets riktning, signifikans och storlek, och *Science* utvärderar fortfarande materialet. Läs den som ett genuint intressant och omsorgsfullt utformat resultat som för närvarande granskas — varken fastslaget eller motbevisat. Den mest uppriktiga meningen om studien är den som processen själv skriver: kontrollera den igen.

Källor

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