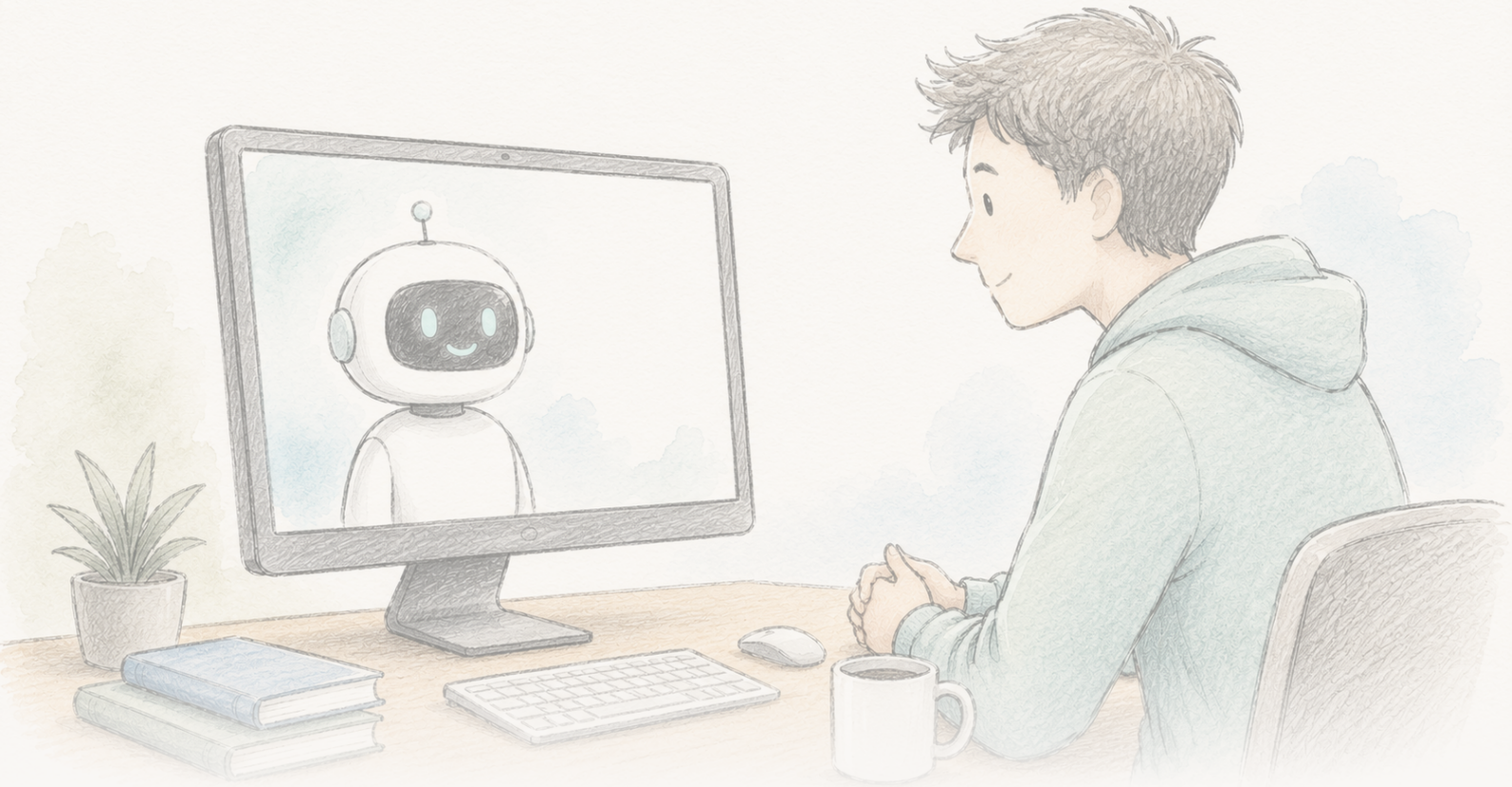


The Clean Paper

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

En chatbot så ut til å redusere troen på konspirasjonsteorier i flere måneder

En studie fra 2024 fant at en samtale i tre runder med GPT-4 Turbo reduserte folks tro på en konspirasjonsteori de selv trodde på, med rundt 20 prosent. Effekten varte i to måneder, gjaldt på tvers av ulike typer konspirasjonsteorier og bygget på KI-påstander som ble vurdert som overveldende korrekte. I juni 2026 ble artikkelen merket med en Editorial Expression of Concern på grunn av problemer med databehandling og reproduserbarhet. Forfatterne sier at en korrigert analyse bevarer funnets retning, signifikans og størrelse, og Science vurderer fortsatt materialet. Et genuint interessant og omhyggelig utformet resultat — nå under vurdering, verken avgjort eller avkreftet.

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 merket artikkelen med en Editorial Expression of Concern mens tidsskriftet vurderer spørsmål om databehandling og reproduserbarhet. Dette er ikke en tilbaketrekking og ikke en konstatering av uredelighet; det betyr at det rapporterte resultatet bør regnes som foreløpig til vurderingen er avsluttet.

Editorial Expression of Concern →

Fakta, tross alt — og et varsel på studien

I 2024 rapporterte en studie noe som går på tvers av en behagelig konvensjonell oppfatning. La noen ha en kort samtale i tre runder med en KI — GPT-4 Turbo, instruert til å svare på de konkrete bevisene personen viser til for en konspirasjonsteori vedkommende selv tror på — og i gjennomsnitt synker troen på teorien med rundt 20 prosent. Nedgangen var fortsatt til stede to måneder senere. Det virket for klassiske konspirasjonsteorier (drapet på John F. Kennedy, romvesener, Illuminati) og aktuelle teorier (covid-19, presidentvalget i USA i 2020), også blant personer med sterk overbevisning knyttet til identiteten deres. Da en profesjonell faktasjekker vurderte 128 påstander fra KI-en, var 127 sanne, én misvisende og ingen falske.

Overskriften som spredte seg, var at det faktisk *går an* å snakke folk ut av kaninhullet — at de som tror på konspirasjonsteorier, ikke er utenfor rekkevidden til bevis, men bare trenger de riktige bevisene, lagt fram konkret nok. Det er en genuint interessant påstand, og studien var utformet mer omhyggelig enn det meste av forskningen på overtalelse.

Men i juni 2026 publiserte *Science* en **Editorial Expression of Concern** om artikkelen. Dette er delen de fleste gjenfortellinger kommer til å hoppe over, og nettopp den delen som avgjør hvor mye vekt resultatet kan bære akkurat nå.

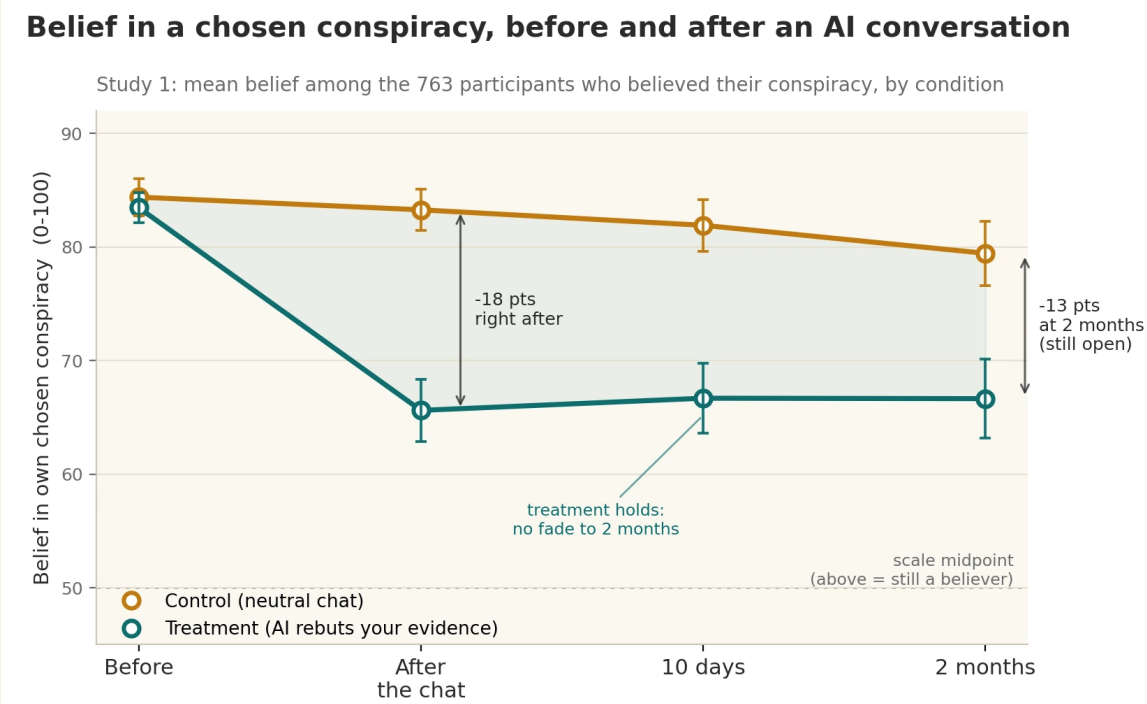
Hva en Expression of Concern er — og ikke er

En Editorial Expression of Concern er et formelt varsel et tidsskrift knytter til en artikkel for å gjøre leserne oppmerksomme på at det er reist spørsmål som fortsatt blir undersøkt. Det er **ikke** en tilbaketrekking (artikkelen er ikke trukket tilbake), og er ikke i seg selv en konstatering av uredelighet. Det betyr: les med dette forbeholdet i bakhodet, fordi forskningsgrunnlaget kan bli endret. Etter at forfatterne her ble gjort oppmerksomme på problemene, undersøkte de dem, rapporterte detaljene til *Science* og leverte en korrigert analyse.

Det som ble varslet, er konkret. Forfatterne ble gjort oppmerksomme på inkonsistens mellom manuskriptet og den publiserte analysekoden i hvordan kriteriene for deltakerscreening var brukt. Det offentlige datasettet inneholdt dessuten uvedkommende rader som var blitt skjøtet inn på grunn av en feil ved sammenslåing av kode. Problemene gjorde noen av de rapporterte tallene vanskelige å reprodusere fra det publiserte materialet. Forfatterne ga *Science* en korrigert analysepipeline og oppdaterte resultater som de sier samsvarer med originalen **i retning, statistisk signifikans og reell**

effektstørrelse. *Science* vurderer materialet. Inntil vurderingen er ferdig, står de nøyaktige tallene under et spørsmålstegn som bare tidsskriftet kan fjerne.

Dette er altså to historier samtidig: et interessant resultat om hvorvidt fakta kan flytte inngrodde overbevisninger, og et levende eksempel på at den vitenskapelige litteraturen korrigerer seg selv i offentligheten. Poenget med denne artikkelen er å holde de to historiene fra hverandre.



I studien ble det rapportert at en KI-samtale i tre runder, der deltakerens egne oppgitte bevis ble imøtegått, reduserte troen på personens valgte konspirasjonsteori med rundt 20 prosent i gjennomsnitt. I motsetning til en kontrollsamtale om et annet tema var nedgangen fortsatt målbar etter 10 dager og 2 måneder. Den rapporterte effekten var varig, men delvis: de fleste deltakerne trodde fortsatt på konspirasjonsteorien etterpå — en reduksjon, ikke en kur. Artikkelen er nå merket med en Editorial Expression of Concern (*Science*, juni 2026) på grunn av inkonsistens i rapporteringen og en feil i det offentlige datasettet. Forfatterne sier at en korrigert analyse bevarer effektens retning, signifikans og størrelse, mens *Science* fortsetter vurderingen. Diagrammet er rekonstruert av The Clean Paper fra det offentlige datasettet, så verdiene som vises, er foreløpige og ikke artikkelens endelige tall.

Hva forfatterne gjorde

- Gjennomførte to eksperimenter med 2190 deltakere, som med egne ord oppga en konspirasjonsteori de trodde på, og bevisene de mente støttet den.
- Lot hver deltaker ha en samtale i tre runder med GPT-4 Turbo. I **behandlingsgruppen** ble KI-en

instruert til å imøtegå deltakerens konkrete bevis; i **kontrollgruppen** diskuterte den et tema uten sammenheng med konspirasjonsteorien.
- Målte troen på den valgte konspirasjonsteorien før og umiddelbart etter samtalen, og kontaktet deretter deltakerne igjen etter **10 dager og 2 måneder**.
- Lot en profesjonell faktasjekker vurdere riktigheten av alle de 128 faktapåstandene KI-en kom med i et utvalg av samtalene.
- Undersøkte om effekten spredte seg til andre, urelaterte konspirasjonsteorier — og, som en test av spesifisitet, om den også reduserte troen på konspirasjoner som faktisk er **sanne**.

Hva de fant

- **En gjennomsnittlig reduksjon på rundt 20 prosent** i troen på den valgte konspirasjonsteorien i behandlingsgruppen sammenlignet med kontrollgruppen (studie 1: 95 prosent konfidensintervall [13,8, 19,7] poeng på en skala fra 0 til 100, $P < 0{,}001$).
- **Effekten varte.** I behandlingsgruppen forsvant ikke nedgangen: etter 10 dager og to måneder lå troen fortsatt nær nivået rett etter samtalen i stedet for å stige igjen, mens kontrollgruppen lå høyere hele tiden. Artikkelen beskriver effekten på den valgte konspirasjonsteorien som uforminsket i minst to måneder, ikke som en kortvarig vakling.
- **Effekten gjaldt på tvers av** de ulike konspirasjonsteoriene deltakerne oppga, og holdt seg også blant deltakere som i utgangspunktet hadde en sterk overbevisning.
- **KI-ens påstander var korrekte** i denne situasjonen: 127 av 128 (99,2 prosent) ble vurdert som sanne, 1 (0,8 prosent) som misvisende og ingen som falske.
- **Effekten var spesifikk**, ikke generell tvil: intervensjonen reduserte **ikke** troen på virkelige konspirasjoner, noe som tyder på at den flyttet ubegrunnede overbevisninger i stedet for å gjøre folk kyniske til alt.
- **Effekten spredte seg i beskjeden grad** — troen på andre, urelaterte konspirasjonsteorier falt med rundt 12 prosent, og deltakerne oppga større vilje til å si imot konspirasjonspåstander. Hoveddeffekten holdt seg i studie 2 og pekte i samme retning i et mindre utvalg uten kontrollgruppe (svakere evidens, men i samsvar med resten).

Hva dette ikke beviser

- Resultatet står **ikke** uimotsagt akkurat nå. Artikkelen er merket med en Editorial Expression of Concern på grunn av problemer med databehandling og reproduserbarhet, og de korrigerte tallene vurderes fortsatt av *Science*. De konkrete verdiene bør regnes som foreløpige til vurderingen er ferdig.
- Studien viser **ikke** at KI «kurerer» tro på konspirasjonsteorier. En gjennomsnittlig reduksjon på rundt 20 prosent er en meningsfull endring, ikke en utslettelse — de fleste deltakerne trodde fortsatt på teorien etterpå, bare mindre.
- Dette er **ikke** et resultat fra bruk i den virkelige verden. Deltakerne meldte seg til en studie og møtte KI-en i god tro. Utenfor studien er de som er mest overbevist om en konspirasjon, også de

som minst sannsynlig vil oppsøke en chatbot som argumenterer mot dem.

- Nøyaktigheten på 99,2 prosent er **spesifikk for denne oppgaven, denne modellen og den omhyggelige utformingen av instruksjonene** — ikke en generell garanti for at språkmodeller kommer med sanne påstander. Forfatterne sier uttrykkelig at den samme personlig tilpassede overtalelsesevnen kunne fremme *falske* overbevisninger dersom den ble rettet mot det.
- «Varig» betyr her **to måneder**, noe som er imponerende for én samtale, men ikke det samme som permanent.

Hvor sterk er evidensen?

- **Studiedesignet er en klar styrke.** Kontrollerte eksperimenter med behandlings- og kontrollgruppe, en ryddig placebolignende kontroll, gjentakelse i to studier og et uavhengig utvalg, oppfølging av varighet og en uttrykkelig kontroll av riktigheten i KI-ens egne påstander — dette er mer omhyggelig enn det meste av forskningen på overtalelse, og retningen på effekten er konsistent overalt den ble testet.
- **Men en Expression of Concern er ikke en fotnote.** Å kunne gjenskape de rapporterte tallene fra publiserte data og publisert kode er en del av det som gjør et resultat troverdig, og det var nettopp dette som sviktet: inkonsistente screeningkriterier og dupliserte rader etter en feil ved sammenslåing av kode. Forfatternes opplysning om at en korrigert pipeline bevarer resultatet, er oppmuntrende og plausibel, men det er deres egen redegjørelse, ennå ikke en uavhengig konklusjon. Det er *Science*s vurdering man må vente på.
- **Den ærlige statusen er «varslet», ikke «avkreftet» eller «bekreftet».** En godt utformet og slående studie der de nøyaktige tallene er under formell vurdering. Det er et ubehagelig sted å la en god historie stå, og det er det riktige.

Hvorfor det er viktig

I mange år har et innflytelsesrikt syn vært at mennesker som tror på konspirasjonsteorier, drives av psykologiske behov og identitet, og derfor ikke kan argumenteres ut av overbevisningene sine med fakta. En varig, evidensbasert reduksjon — dersom resultatet holder — er en viktig motvekt til denne pessimismen. Den antyder at en del av problemet var at generell faktasjekking aldri tok for seg de *konkrete* bevisene en person faktisk viser til, noe en stor språkmodell kan gjøre i stor skala.

Studien er også viktig som et eksempel på hvordan vitenskap er ment å fungere. Lærdommen av en Expression of Concern er ikke at berømte resultater er verdiløse, men at feil kommer fram, data undersøkes på nytt og påstander etterprøves offentlig. At dette maskineriet er i gang, er en styrke, ikke en skandale — og derfor er tålmodighet den riktige holdningen til resultatet, fremfor enten applaus eller avvisning.

Og KI-spørsmålet går begge veier. Den samme evnen til å svekke en falsk overbevisning med et skreddersydd argument kunne like effektivt styrke en. Teknikken er nøytral; bare målet er det ikke.

Kort oppsummering

En studie fra 2024 fant at en samtale i tre runder med GPT-4 Turbo reduserte folks tro på en konspirasjonsteori de selv trodde på, med rundt 20 prosent. Effekten varte i to måneder og gjaldt på tvers av ulike typer konspirasjonsteorier, mens KI-ens faktapåstander ble vurdert som overveldende korrekte. I juni 2026 ble artikkelen merket med en Editorial Expression of Concern på grunn av problemer med databehandling og reproduserbarhet. Forfatterne opplyser at en korrigert analyse bevarer funnets retning, signifikans og størrelse, og *Science* vurderer fortsatt materialet. Les den som et genuint interessant og omhyggelig utformet resultat som nå er til vurdering — verken avgjort eller avkreftet. Den mest ærlige setningen om studien er den prosessen selv skriver: sjekk den igjen.

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