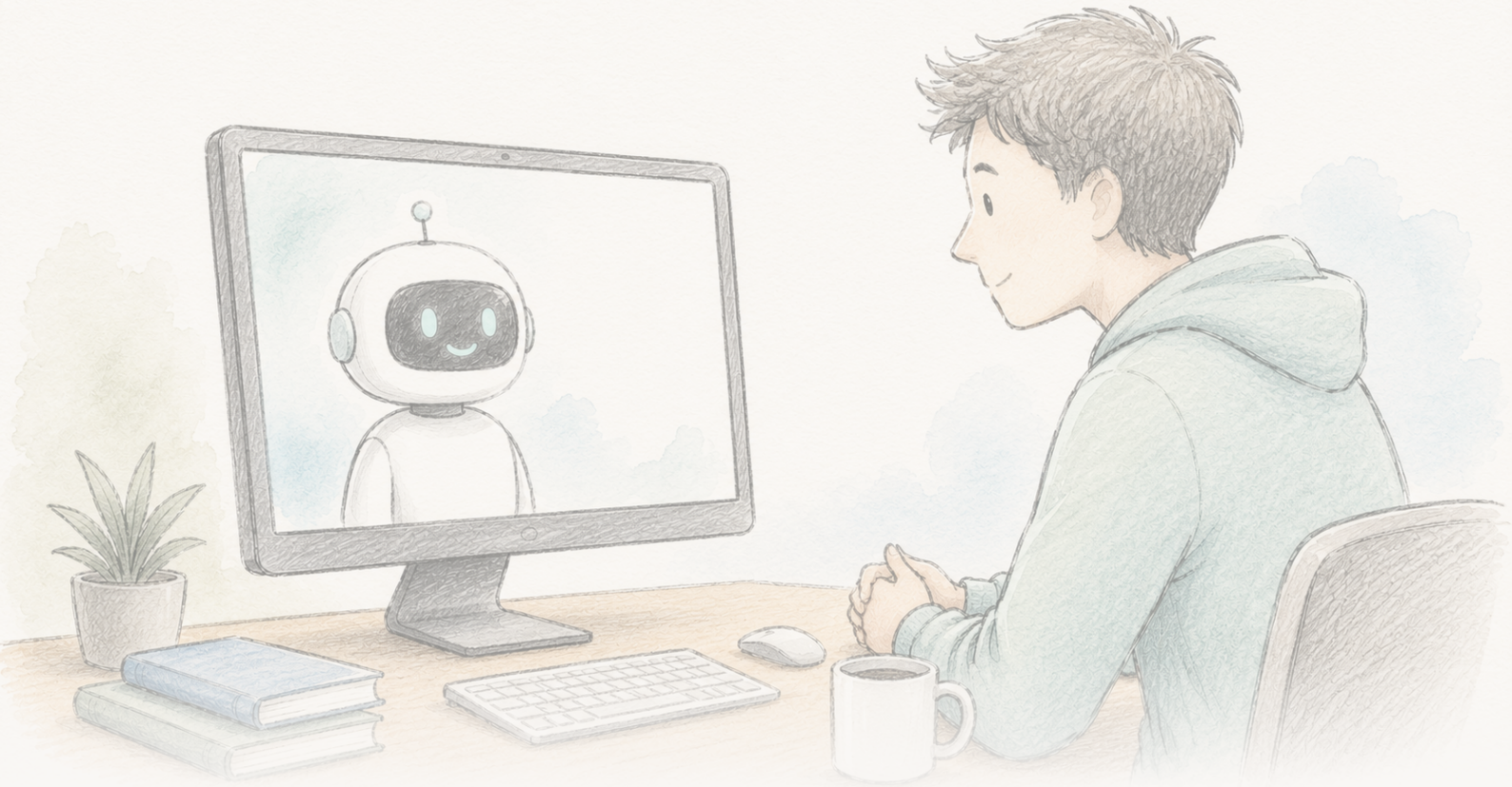


The Clean Paper

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

Een chatbot leek complotgeloof maandenlang te verminderen

Een studie uit 2024 vond dat een gesprek van drie rondes met GPT-4 Turbo het geloof van mensen in een door hen aangehangen complottheorie met ongeveer 20% verminderde — een effect dat twee maanden aanhield, generaliseerde over complottypes en rustte op AI-beweringen die als overweldigend accuraat werden beoordeeld. In juni 2026 werd de paper onder een Editorial Expression of Concern geplaatst wegens problemen met datahantering en reproduceerbaarheid; de auteurs zeggen dat een gecorrigeerde analyse de bevinding behoudt in richting, significantie en omvang, en Science beoordeelt nog. Een werkelijk interessant, zorgvuldig gebouwd resultaat — momenteel onder beoordeling, niet beklonken en niet ontkracht.

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 heeft aan deze paper een Editorial Expression of Concern gehecht terwijl het vragen over datahantering en reproduceerbaarheid beoordeelt. Dit is geen retraction en geen vaststelling van wangedrag; het betekent dat het gerapporteerde resultaat als voorlopig gelezen moet worden tot de beoordeling is afgerond.

Editorial Expression of Concern →

Feiten werken dus toch — en een vlag op de paper

In 2024 rapporteerde een studie iets dat ingaat tegen een comfortabel stuk conventionele wijsheid. Geef iemand een kort gesprek van drie rondes met een AI — GPT-4 Turbo, geïnstrueerd om in te gaan op het specifieke bewijs dat die persoon aanvoert voor een complottheorie die hij zelf gelooft — en gemiddeld daalt zijn geloof in die theorie met ongeveer 20%. De daling was er twee maanden later nog. Het werkte bij klassieke complotten (de moord op John F. Kennedy, buitenaardse wezens, de Illuminati) en bij actuele (COVID-19, de Amerikaanse verkiezingen van 2020), en zelfs bij mensen wier overtuiging sterk was en met hun identiteit verweven. Toen een professionele factchecker 128 van de beweringen van de AI beoordeelde, waren er 127 waar, één misleidend en geen enkele onwaar.

De kop die rondging was dat je mensen *wel degelijk* uit het konijnenhol kunt praten — dat complotgelovigen niet buiten het bereik van bewijs staan, ze hebben alleen het juiste bewijs nodig, specifiek genoeg gebracht. Dat is een werkelijk interessante claim, en de studie was zorgvuldiger gebouwd dan het meeste overtuigingsonderzoek.

Toen, in juni 2026, plaatste *Science* een **Editorial Expression of Concern** bij de paper — een redactionele uiting van zorg. Dit is het deel dat de meeste navertellingen zullen overslaan, en het is precies het deel dat bepaalt hoeveel gewicht het resultaat op dit moment kan dragen.

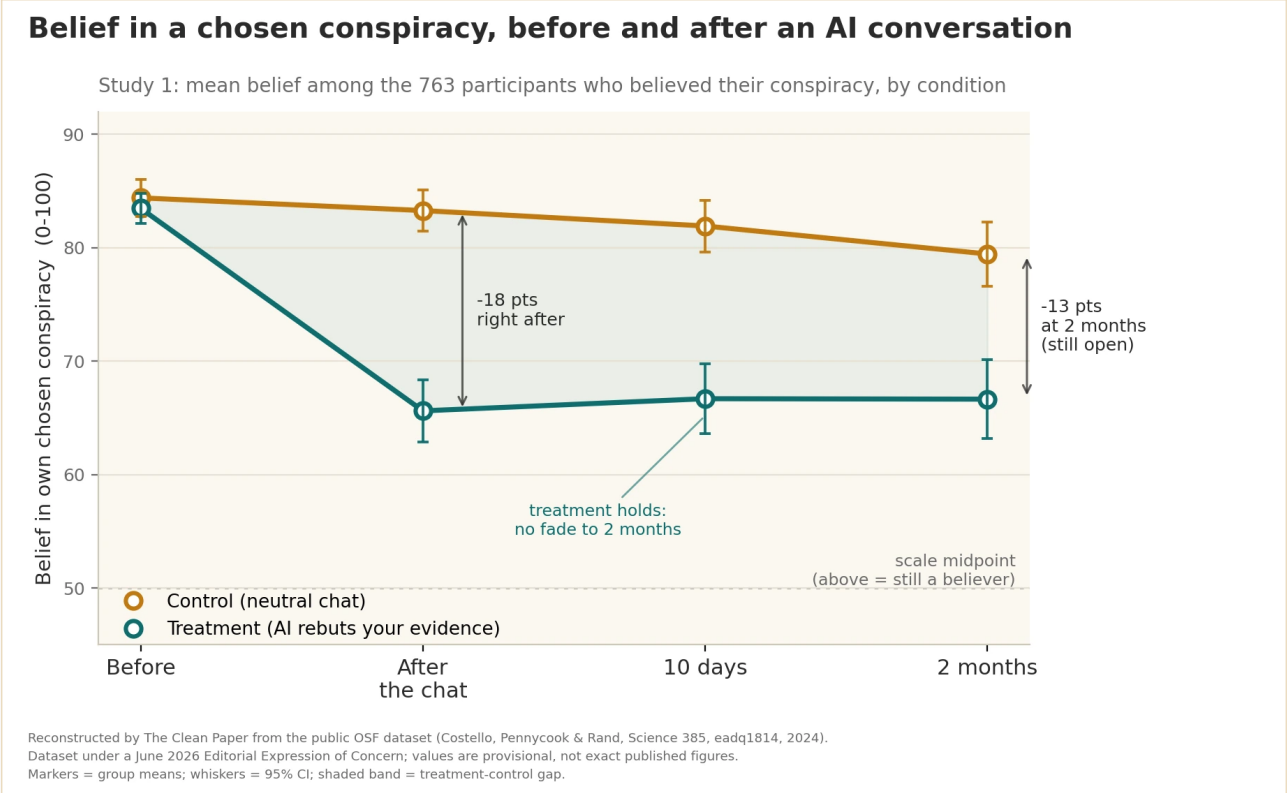
Wat een Expression of Concern is — en wat niet

Een Editorial Expression of Concern is een formele vlag die een tijdschrift aan een paper hecht om lezers te waarschuwen dat er vragen zijn gerezen, terwijl die vragen nog worden uitgezocht. Het is **geen** retraction (de paper wordt niet teruggetrokken), en het is op zichzelf geen vaststelling van wangedrag. Het betekent: lees dit met het voorbehoud in het achterhoofd, want het beeld kan veranderen. Hier hebben de auteurs, nadat ze op de problemen waren gewezen, de zaak onderzocht, de details aan *Science* gemeld en een gecorrigeerde analyse aangeleverd.

Wat er is aangemerkt, is specifiek. De auteurs werden gewezen op inconsistenties in hoe de selectiecriteria voor deelnemers werden toegepast tussen het manuscript en de gepubliceerde analysecode, en op het feit dat de openbare dataset overtollige rijen bevatte die er door een fout bij het samenvoegen van code in waren geslopen — problemen die sommige gerapporteerde cijfers moeilijk reproduceerbaar maakten uit het vrijgegeven materiaal. De auteurs gaven *Science* een gecorrigeerde

analysepijplijn en een set bijgewerkte resultaten, die volgens hen overeenkomen met het origineel **in richting, statistische significantie en substantiële omvang**. *Science* beoordeelt ze. Tot die beoordeling is afgerond, staan de exacte cijfers onder een vraagteken dat alleen het tijdschrift kan wegnemen.

Dit zijn dus twee verhalen tegelijk: een intrigerend resultaat over de vraag of feiten verankerde overtuigingen kunnen bewegen, en een levend voorbeeld van het wetenschappelijke archief dat zichzelf in het openbaar corrigeert. Het doel van dit stuk is ze uit elkaar te houden.



In de studie werd gerapporteerd dat een AI-gesprek van drie rondes, dat het door de deelnemer zelf genoemde bewijs weerlegde, het geloof in diens gekozen complot met gemiddeld zo'n 20% verlaagde; anders dan bij een controlegesprek over een los onderwerp was de daling na 10 dagen en 2 maanden nog meetbaar. Het gerapporteerde effect was duurzaam maar gedeeltelijk: de meeste deelnemers geloofden het complot daarna nog steeds — een vermindering, geen genezing. De paper staat momenteel onder een Editorial Expression of Concern (*Science*, juni 2026) wegens rapportage-inconsistenties en een fout in de openbare dataset; de auteurs zeggen dat een gecorrigeerde analyse richting, significantie en omvang van het effect behoudt, terwijl *Science* blijft beoordelen. Deze grafiek is door The Clean Paper gereconstrueerd uit die openbare dataset, dus de getoonde waarden zijn voorlopig, niet de definitieve cijfers van de paper.

Wat de auteurs deden

- Ze voerden twee experimenten uit met 2190 deelnemers die elk — in eigen woorden — een complottheorie noemden die ze geloofden, en het bewijs dat die volgens hen ondersteunde.

- Elke deelnemer voerde een gesprek van drie rondes met GPT-4 Turbo. In de **behandel**-conditie was de AI geïnstrueerd het specifieke bewijs van de deelnemer te weerleggen; in de **controle**-conditie besprak hij een los onderwerp.
- Het geloof in het gekozen complot werd gemeten vóór en direct na het gesprek, en de deelnemers werden opnieuw benaderd na **10 dagen en 2 maanden**.
- Een professionele factchecker beoordeelde de juistheid van alle 128 feitelijke beweringen die de AI in een steekproef deed.
- Er werd gecontroleerd of het effect oversloeg naar andere, niet-verwante complotten — en, als specificiteitstest, of het ook het geloof drukte in complotten die toevallig **waar** zijn.

Wat ze vonden

- **Een gemiddelde vermindering van ruwweg 20%** van het geloof in het gekozen complot in de behandelgroep ten opzichte van de controle (studie 1: 95%-betrouwbaarheidsinterval [13,8; 19,7] punten op een 100-puntsschaal, $P < 0{,}001$).
- **Het hield stand.** In de behandelgroep vervaagde de daling niet: het geloof bleef na 10 dagen en twee maanden dicht bij het niveau van vlak na het gesprek in plaats van terug te kruipen, terwijl de controle de hele tijd hoger bleef — de paper beschrijft het effect op het gekozen complot als minstens twee maanden onverminderd aanhoudend, geen momentane wiebel.
- **Het generaliseerde** over de reeks complotten die mensen noemden, en hield ook stand bij deelnemers wier overtuiging aanvankelijk sterk was.
- **De beweringen van de AI waren accuraat** in deze setting: 127 van de 128 (99,2%) beoordeeld als waar, 1 (0,8%) als misleidend, geen als onwaar.
- **Het was specifiek**, geen blinde twijfel: de interventie verminderde het geloof in ware complotten **niet** — wat erop wijst dat ze onhoudbare overtuigingen bewoog in plaats van mensen overal cynisch over te maken.
- **Het sloeg bescheiden over** — het geloof in andere, niet-verwante complotten daalde met zo'n 12%, en deelnemers rapporteerden een sterkere intentie om tegen complotclaims in te gaan. Het hoofdeffect hield stand in studie 2 en wees dezelfde kant op in een kleinere steekproef zonder controlegroep (zwakker bewijs, maar consistent).

Wat dit niet bewijst

- Het staat op dit moment **niet** onbetwist overeind. De paper staat onder een Editorial Expression of Concern wegens problemen met datahantering en reproduceerbaarheid, en de gecorrigeerde cijfers worden nog door *Science* beoordeeld. Behandel de specifieke waarden als voorlopig tot die beoordeling er ligt.
- Het toont **niet** aan dat AI complotgeloof "geneest". Een gemiddelde vermindering van ~20% is een betekenisvolle verschuiving, geen uitwissing — de meeste deelnemers geloofden de theorie daarna nog steeds, alleen minder.
- Het is **geen** resultaat uit de echte wereld. Deelnemers meldden zich aan voor een studie en gingen

te goeder trouw met de AI in gesprek; in het wild zijn de mensen die het sterkst aan een complot hechten het minst geneigd een chatbot op te zoeken die met ze in discussie gaat.
- De 99,2% nauwkeurigheid is **specifiek voor deze taak, dit model en deze zorgvuldige prompting** — geen algemene garantie dat taalmodellen ware dingen zeggen. De auteurs zijn er expliciet over dat dezelfde gepersonaliseerde overtuigingskracht *onware* overtuigingen zou kunnen aanjagen als ze daarop gericht werd.
- "Duurzaam" betekent hier **twee maanden** — indrukwekkend voor één gesprek, maar niet hetzelfde als blijvend.

Hoe sterk is het bewijs

- **Het ontwerp is een echte kracht.** Gecontroleerde experimenten met behandel- en controlegroep, een schone placebo-achtige controle, replicatie over twee studies en een onafhankelijke steekproef, follow-ups op duurzaamheid en een expliciete nauwkeurigheidscheck op de eigen beweringen van de AI — dit is zorgvuldiger dan het meeste overtuigingsonderzoek, en de richting van het effect is overal consistent waar ze is getest.
- **Maar de Expression of Concern is geen voetnoot.** De gerapporteerde cijfers kunnen regenereren uit de vrijgegeven data en code is deel van wat een resultaat betrouwbaar maakt, en dat is precies wat brak — inconsistenties in selectiecriteria en gedupliceerde rijen uit een fout bij het samenvoegen van code. De verklaring van de auteurs dat een gecorrigeerde pijplijn het resultaat behoudt is bemoedigend en plausibel, maar het is het eigen relaas van de auteurs, nog geen onafhankelijk oordeel. De beoordeling van *Science* is waar op te wachten valt.
- **De eerlijke status is "gevlagd", niet "ontkracht" en niet "bevestigd".** Een goed gebouwde, opvallende studie waarvan de exacte cijfers onder formele beoordeling staan. Dat is een ongemakkelijke plek om een goed verhaal achter te laten, en het is de juiste.

Waarom het ertoe doet

Jarenlang gold de invloedrijke opvatting dat complotgelovigen worden gedreven door psychologische behoeften en identiteit, en dus niet met feiten uit hun overtuigingen te praten zijn. Een duurzame, op bewijs gebaseerde vermindering — als die standhoudt — is een betekenisvol tegenwicht tegen dat pessimisme: het suggereert dat het probleem er deels in lag dat generiek debunken nooit inging op het *specifieke* bewijs dat een gelovige daadwerkelijk aanvoert — iets wat een taalmodel op schaal kan doen.

Het doet er ook toe als casestudy in hoe wetenschap hoort te werken. De les van de Expression of Concern is niet dat gevierde resultaten waardeloos zijn; het is dat fouten boven water komen, data opnieuw wordt bekeken en claims in het openbaar worden nagelopen. Dat die machinerie draait is een kenmerk, geen schandaal — en het is de reden dat de juiste houding tegenover dit resultaat geduld is, geen applaus en geen wegwuiven.

En het snijdt aan twee kanten bij AI. Hetzelfde vermogen om een onware overtuiging met een op maat

gesneden argument leeg te laten lopen, zou er een net zo effectief kunnen opblazen. De techniek is neutraal; alleen het doel is dat niet.

Heldere samenvatting

Een studie uit 2024 vond dat een gesprek van drie rondes met GPT-4 Turbo het geloof van mensen in een door hen aangehangen complottheorie met ongeveer 20% verminderde — een effect dat twee maanden aanhield, generaliseerde over complottypes en rustte op AI-beweringen die als overweldigend accuraat werden beoordeeld. In juni 2026 werd de paper onder een Editorial Expression of Concern geplaatst wegens problemen met datahantering en reproduceerbaarheid; de auteurs rapporteren dat een gecorrigeerde analyse de bevinding behoudt in richting, significantie en omvang, en *Science* beoordeelt nog. Lees het als een werkelijk interessant, zorgvuldig gebouwd resultaat dat momenteel onder beoordeling staat — niet beklonken, niet ontkracht. De eerlijkste zin erover schrijft het proces zelf op dit moment: controleer het opnieuw.

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