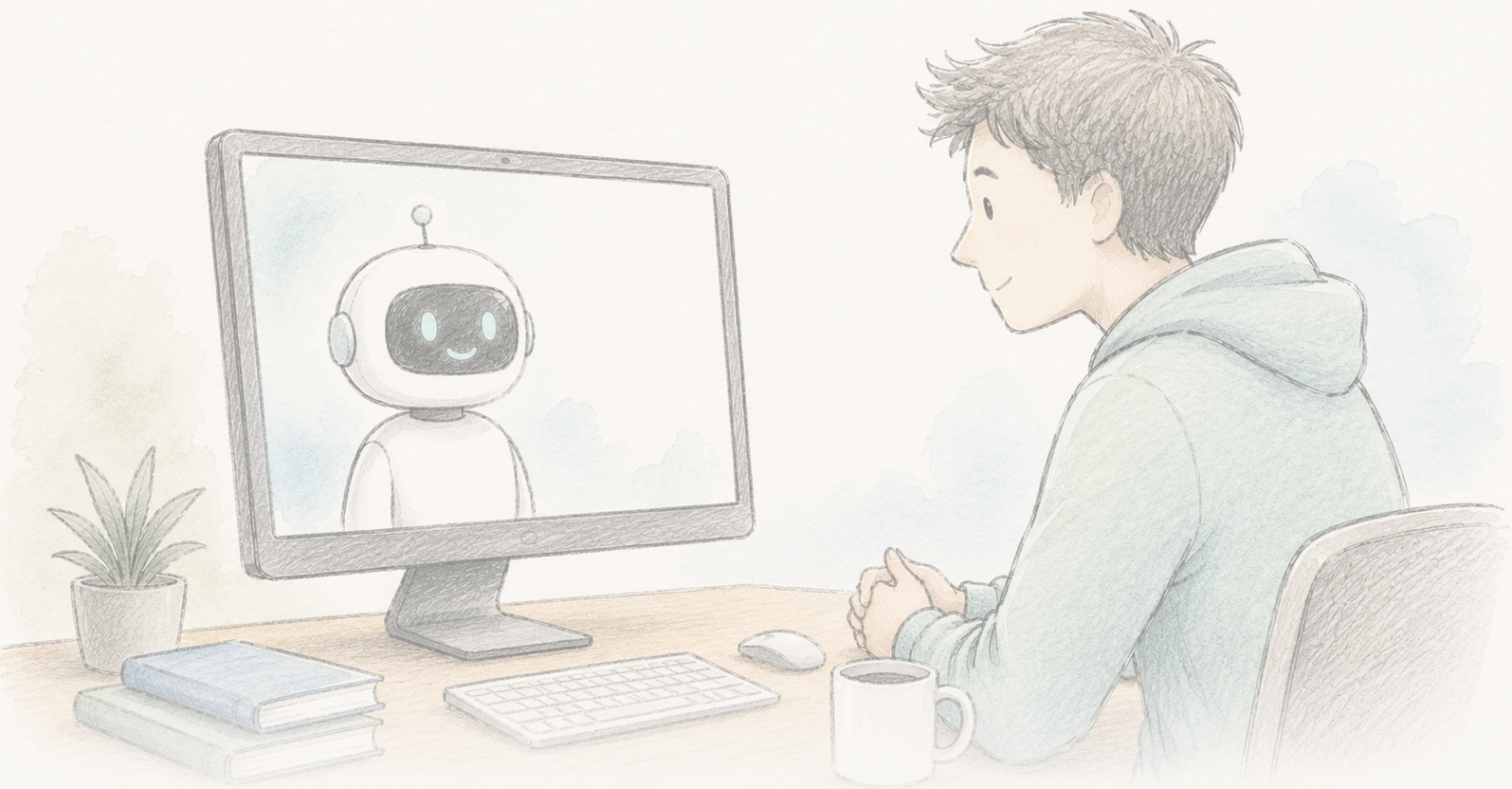




The CLEAN PAPER

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

'n Kletsbot het geblyk samesweringsgeloof maande lank te verminder

'n Studie van 2024 het gevind dat 'n gesprek van drie rondtes met GPT-4 Turbo mense se geloof in 'n samesweringsteorie wat hulle aanhang met sowat 20% verminder het — 'n effek wat twee maande gehou het, oor samesweringstipes veralgemeen het en gerus het op KI-bewerings wat as oorweldigend akkuraat beoordeel is. In Junie 2026 is die artikel onder 'n Editorial Expression of Concern geplaas oor probleme met datahantering en reproduseerbaarheid; die outeurs sê 'n gekorrigeerde ontleding behou die bevinding in rigting, beduidendheid en grootte, en Science beoordeel nog. 'n Werklik interessante, sorgvuldig geboude resultaat — tans onder beoordeling, nie afgehandel nie en nie weerlê nie.

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 het aan hierdie artikel 'n Editorial Expression of Concern geheg terwyl dit vrae oor datahantering en reproduseerbaarheid beoordeel. Dit is nie 'n retraction nie en geen bevinding van wangedrag nie; dit beteken dat die gerapporteerde resultaat as voorlopig gelees moet word totdat die beoordeling afgehandel is.

Editorial Expression of Concern →

Feite werk tog — en 'n vlag op die artikel

In 2024 het 'n studie iets gerapporteer wat teen 'n gemaklike stuk konvensionele wysheid indruis. Gee iemand 'n kort gesprek van drie rondtes met 'n KI — GPT-4 Turbo, opdrag gegee om die spesifieke bewyse aan te spreek wat die persoon aanvoer vir 'n samesweringsteorie wat hy self glo — en gemiddeld daal sy geloof in daardie teorie met sowat 20%. Die daling was twee maande later nog daar. Dit het gewerk by klassieke sameswerings (die sluipmoord op John F. Kennedy, ruimtewesens, die Illuminati) en by aktuele (COVID-19, die Amerikaanse verkiesing van 2020), en selfs by mense wie se oortuiging sterk was en met hulle identiteit verweef. Toe 'n professionele feitenagaanwyser 128 van die KI se bewerings beoordeel het, was 127 waar, een misleidend en geen enkele vals nie.

Die opskrif wat versprei het, was dat jy mense wél uit die konynegat kan praat — dat samesweringsgelowiges nie buite die bereik van bewyse staan nie; hulle het net die regte bewyse nodig, spesifiek genoeg aangebied. Dit is 'n werklik interessante bewering, en die studie was sorgvuldiger gebou as die meeste oorredingsnavorsing.

Toe, in Junie 2026, het *Science* 'n **Editorial Expression of Concern** op die artikel geplaas — 'n redaksionele uiting van kommer. Dit is die deel wat die meeste oorvertellings sal oorslaan, en dit is presies die deel wat bepaal hoeveel gewig die resultaat op die oomblik kan dra.

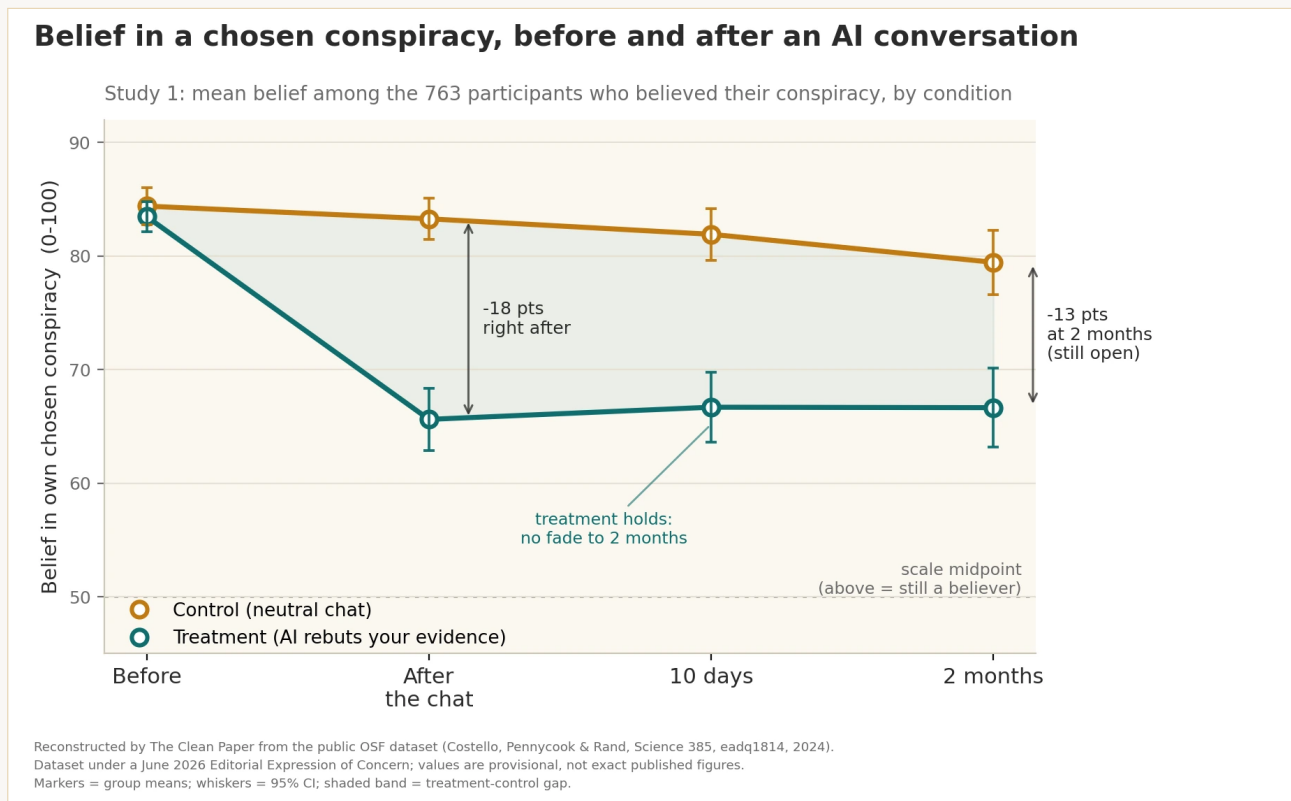
Wat 'n Expression of Concern is — en wat nie

'n Editorial Expression of Concern is 'n formele vlag wat 'n vaktydskrif aan 'n artikel heg om lesers te waarsku dat vrae geopper is, terwyl daardie vrae nog uitgewerk word. Dit is **nie** 'n retraction nie (die artikel word nie teruggetrek nie), en dit is op sigself geen bevinding van wangedrag nie. Dit beteken: lees dit met die voorbehoud in gedagte, want die prentjie kan verander. Hier het die outeurs, nadat hulle op die probleme gewys is, die saak ondersoek, die besonderhede aan *Science* gerapporteer en 'n gekorrigeerde ontleding verskaf.

Wat aangemerkt is, is spesifiek. Die outeurs is gewys op teenstrydighede in hoe die keuringskriteria vir deelnemers toegepas is tussen die manuskrip en die gepubliseerde ontledingskode, en daarop dat die openbare datastel oortollige rye bevat het wat deur 'n fout met die saamvoeg van kode ingekom het — probleme wat sommige van die gerapporteerde syfers moeilik reproduseerbaar gemaak het uit die vrygestelde materiaal. Die outeurs het vir *Science* 'n gekorrigeerde ontledingspyplyn en 'n stel bygewerkte resultate gegee, wat volgens hulle met die oorspronklike ooreenstem **in rigting**,

statistiese beduidendheid en wesenlike grootte. *Science* beoordeel dit. Totdat daardie beoordeling klaar is, staan die presiese syfers onder 'n vraagteken wat net die tydskrif kan verwyder.

Dit is dus twee stories tegelyk: 'n boeiende resultaat oor of feite verskanste oortuigings kan beweeg, en 'n lewende voorbeeld van die wetenskaplike rekord wat homself in die openbaar regstel. Die doel van hierdie stuk is om hulle uitmekaar te hou.



In die studie is gerapporteer dat 'n KI-gesprek van drie rondtes, wat die deelnemer se eie genoemde bewyse weerlê het, die geloof in daardie persoon se gekose samesweringsteorie met gemiddeld sowat 20% verlaag het; anders as by 'n kontrolegesprek oor 'n onverwante onderwerp was die daling ná 10 dae en 2 maande nog meetbaar. Die gerapporteerde effek was blywend maar gedeeltelik: die meeste deelnemers het die samesweringsteorie daarna steeds geglo — 'n vermindering, nie 'n genesing nie. Die artikel staan tans onder 'n Editorial Expression of Concern (*Science*, Junie 2026) oor rapporteringsteenstrydighede en 'n fout in die openbare datastel; die outeurs sê 'n gekorrigeerde ontleding behou die effek se rigting, beduidendheid en grootte, terwyl *Science* voortgaan om te beoordeel. Hierdie grafiek is deur The Clean Paper uit daardie openbare datastel gerekonstrueer, so die waardes wat gewys word, is voorlopig, nie die artikel se finale syfers nie.

Original figure — The Clean Paper CC BY 4.0

Wat die outeurs gedoen het

- Hulle het twee eksperimente uitgevoer met 2190 deelnemers wat elkeen — in eie woorde — 'n samesweringsteorie genoem het wat hulle glo, en die bewyse wat dit volgens hulle ondersteun.
- Elke deelnemer het 'n gesprek van drie rondtes met GPT-4 Turbo gevoer. In die **behandelings-**

toestand is die KI opdrag gegee om die deelnemer se spesifieke bewyse te weerlê; in die **kontrole**-toestand het dit oor 'n onverwante onderwerp gesels.

- Geloof in die gekose sameswering is gemeet voor en direk ná die gesprek, en die deelnemers is weer gekontak ná **10 dae en 2 maande**.
- 'n Professionele feitenagaanwyser het die akkuraatheid beoordeel van al 128 feitlike bewerings wat die KI in 'n steekproef gemaak het.
- Daar is getoets of die effek oorgespoel het na ander, onverwante sameswerings — en, as spesifisiteitstoets, of dit ook die geloof verlaag het in sameswerings wat toevallig **waar** is.

Wat hulle gevind het

- **'n Gemiddelde vermindering van ruweg 20%** in geloof in die gekose sameswering in die behandelingsgroep teenoor die kontrole (studie 1: 95%-vertrouensinterval [13,8; 19,7] punte op 'n 100-punt-skaal, $P < 0,001$).
- **Dit het gehou.** In die behandelingsgroep het die daling nie vervaag nie: geloof het ná 10 dae en twee maande naby die vlak van net-ná-die-gesprek gebly eerder as om terug te kruip, terwyl die kontrole deurgaans hoër gebly het — die artikel beskryf die effek op die gekose sameswering as minstens twee maande onverminderd blywend, nie 'n oomblik se wankeling nie.
- **Dit het veralgemeen** oor die reeks sameswerings wat mense genoem het, en het ook gehou by deelnemers wie se oortuiging aanvanklik sterk was.
- **Die KI se bewerings was akkuraat** in hierdie opset: 127 uit 128 (99,2%) as waar beoordeel, 1 (0,8%) as misleidend, geen as vals nie.
- **Dit was spesifiek**, nie blinde twyfel nie: die ingreep het die geloof in ware sameswerings **nie** verminder nie — 'n aanduiding dat dit onondersteunde oortuigings beweeg het eerder as om mense oor alles sinies te maak.
- **Dit het beskeie oorgespoel** — geloof in ander, onverwante sameswerings het met sowat 12% gedaal, en deelnemers het 'n groter voorneme gerapporteer om teen samesweringsbewerings standpunt in te neem. Die hoofeffek het in studie 2 gehou en het in 'n kleiner steekproef sonder kontrolegroep in dieselfde rigting gewys (swakker getuienis, maar konsekwent).

Wat dit nie bewys nie

- Dit staan op die oomblik **nie** onbevraagteken nie. Die artikel dra 'n Editorial Expression of Concern oor probleme met datahantering en reproduseerbaarheid, en die gekorrigeerde syfers word nog deur *Science* beoordeel. Hanteer die spesifieke waardes as voorlopig totdat daardie beoordeling verskyn.
- Dit wys **nie** dat KI samesweringsgeloof “genees” nie. 'n Gemiddelde vermindering van ~20% is 'n betekenisvolle verskuiwing, nie 'n uitwissing nie — die meeste deelnemers het die teorie daarna steeds geglo, net minder.
- Dit is **nie** 'n resultaat uit werklike gebruik nie. Deelnemers het vrywillig by 'n studie aangesluit en te goeder trou met die KI omgegaan; in die wêreld daarbuite is die mense wat die sterkste aan 'n

sameswering vasklou, die minste geneig om 'n kletsbot op te soek wat met hulle sal stry.

- Die 99,2% akkuraatheid is **spesifiek vir hierdie taak, hierdie model en hierdie sorgvuldige aansporing** — geen algemene waarborg dat taalmodelle waar dinge sê nie. Die outeurs sê uitdruklik dat dieselfde gepersonaliseerde oorreringskrag *vals* oortuigings kon aanblaas as dit daarop gerig word.
- “Blywend” beteken hier **twee maande** — indrukwekkend vir 'n enkele gesprek, maar nie dieselfde as permanent nie.

Hoe sterk is die getuienis

- **Die ontwerp is 'n werklike sterkte.** Gekontroleerde behandeling-teen-kontrole-eksperimente, 'n skoon placebo-agtige kontrole, replikasie oor twee studies en 'n onafhanklike steekproef, opvolge oor blywendheid, en 'n uitdruklike akkuraatheidstoets op die KI se eie bewerings — dit is sorgvuldiger as die meeste oorreringsnavorsing, en die rigting van die effek is oral konsekwent waar dit getoets is.
- **Maar die Expression of Concern is nie 'n voetnoot nie.** Om die gerapporteerde syfers uit die vrygestelde data en kode te kan herskep, is deel van wat 'n resultaat vertrouenswaardig maak — en dit is presies wat gebreek het: teenstrydighede in keuringskriteria en gedupliseerde rye uit 'n kode-saamvoegfout. Die outeurs se verklaring dat 'n gekorrigeerde pyplyn die resultaat behou, is bemoedigend en aanneemlik, maar dit is die outeurs se eie weergawe, nog nie 'n onafhanklike uitspraak nie. *Science* se beoordeling is die ding om voor te wag.
- **Die eerlike status is “gevlag”, nie “weerlê” en nie “bevestig” nie.** 'n Goed geboude, treffende studie waarvan die presiese syfers onder formele beoordeling staan. Dit is 'n ongemaaklike plek om 'n goeie storie te laat staan, en dit is die akkurate een.

Waarom dit saak maak

Jare lank het 'n invloedryke siening gegeld dat samesweringsgelowiges deur sielkundige behoeftes en identiteit gedryf word, en dus nie met feite uit hulle oortuigings geredeneer kan word nie. 'n Blywende, bewysgebaseerde vermindering — as dit standhou — is 'n betekenisvolle teenwig vir daardie pessimisme: dit dui daarop dat die probleem deels was dat generiese weerlegging nooit die *spesifieke* bewyse aangespreek het wat 'n gelowige werklik aanvoer nie — iets wat 'n taalmodel op skaal kan doen.

Dit maak ook saak as gevallestudie van hoe wetenskap veronderstel is om te werk. Die les van die Expression of Concern is nie dat gevierde resultate waardeloos is nie; dit is dat foute na vore kom, data weer bekyk word en bewerings in die openbaar nagegaan word. Dat daardie masjinerie loop, is 'n kenmerk, nie 'n skandaal nie — en dit is die rede waarom die regte houding teenoor hierdie resultaat geduld is, nie applous en nie afwysing nie.

En dit sny na twee kante by KI. Dieselfde vermoë om 'n vals oortuiging met 'n pasgemaakte argument te laat leegloop, kan een net so doeltreffend opblaas. Die tegniek is neutraal; net die doelwit is nie.

Skoon opsomming

'n Studie van 2024 het gevind dat 'n gesprek van drie rondtes met GPT-4 Turbo mense se geloof in 'n samesweringsteorie wat hulle aanhang met sowat 20% verminder het — 'n effek wat twee maande gehou het, oor samesweringstipes veralgemeen het en gerus het op KI-bewerings wat as oorweldigend akkuraat beoordeel is. In Junie 2026 is die artikel onder 'n Editorial Expression of Concern geplaas oor probleme met datahantering en reproduseerbaarheid; die outeurs rapporteer dat 'n gekorrigeerde ontleding die bevinding behou in rigting, beduidendheid en grootte, en *Science* beoordeel nog. Lees dit as 'n werklik interessante, sorgvuldig geboude resultaat wat tans onder beoordeling staan — nie afgehandel nie, nie weerlê nie. Die eerlikste sin daaroor word tans deur die proses self geskryf: gaan dit weer na.

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>