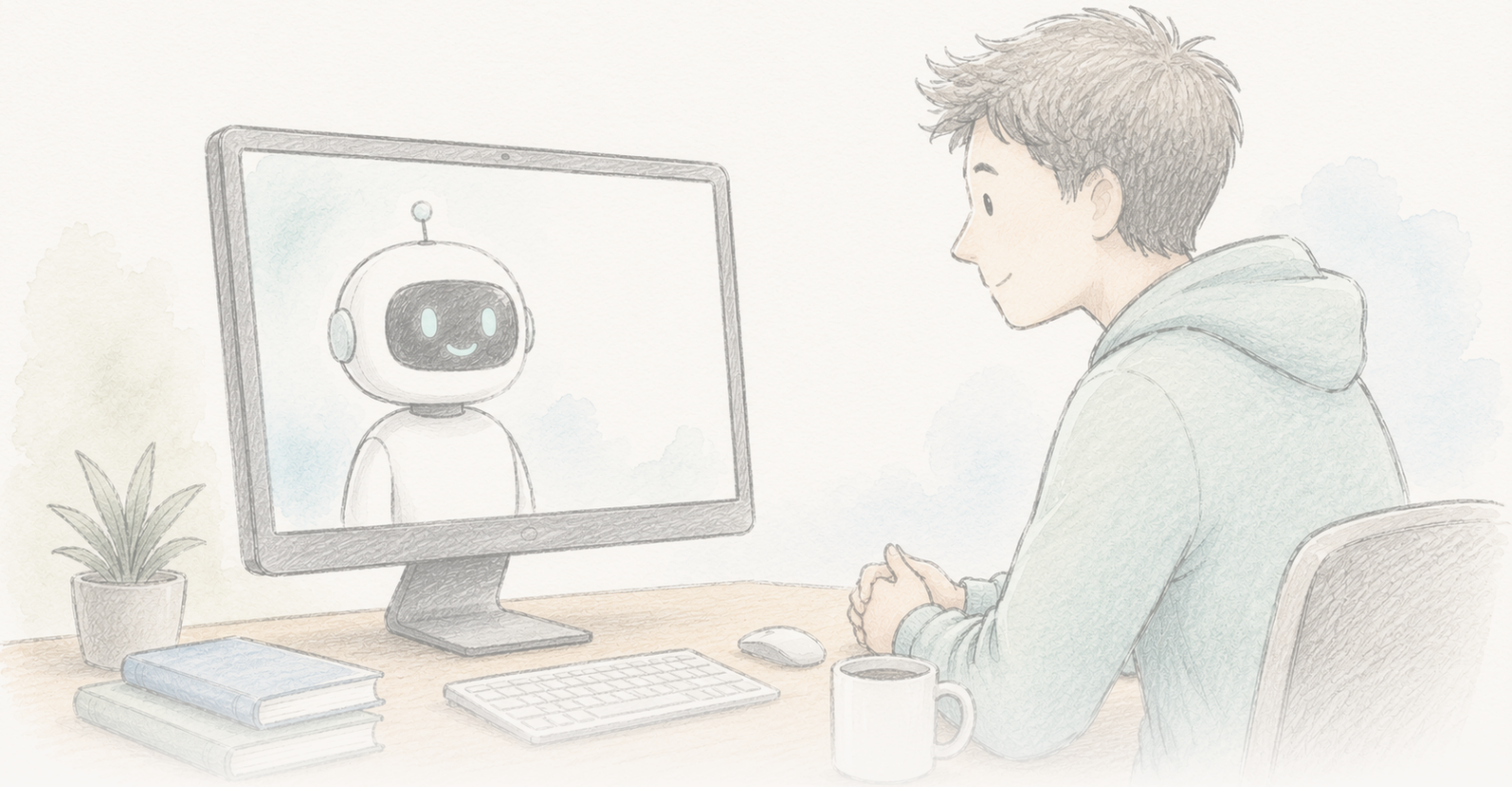


The Clean Paper

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

Čini se da je chatbot mjesecima smanjivao vjerovanje u teorije zavjere

Studija iz 2024. pokazala je da je razgovor s GPT-4 Turbo u tri kruga smanjio vjerovanje ljudi u teoriju zavjere koju su zastupali za oko 20%, uz učinak koji je trajao dva mjeseca, pojavljivao se kroz različite vrste teorija i počivao na činjeničnim tvrdnjama AI-ja ocijenjenima gotovo potpuno točnima. U lipnju 2026. rad je stavljen pod Editorial Expression of Concern zbog problema u obradi podataka i reproducibilnosti; autori kažu da ispravljena analiza čuva smjer, značajnost i veličinu nalaza, a Science ga još procjenjuje. Doista zanimljiv, pažljivo dizajniran rezultat — trenutačno pod provjerom, ni potvrđen ni opovrgnut.

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 je uz ovaj rad objavio Editorial Expression of Concern dok procjenjuje pitanja obrade podataka i reproducibilnosti. To nije povlačenje rada niti nalaz znanstvenog nepoštenja; znači da prijavljeni rezultat treba smatrati privremenim dok se provjera ne razriješi.

Editorial Expression of Concern →

Činjenice ipak mogu djelovati — ali rad sada nosi upozorenje

Studija iz 2024. izvijestila je o rezultatu koji ide protiv jednog ugodnog dijela konvencionalne mudrosti. Dajte osobi kratak razgovor s AI-jem u tri kruga — GPT-4 Turbo, upućen da odgovara baš na dokaze koje ta osoba navodi za teoriju zavjere u koju osobno vjeruje — i njezino vjerovanje u tu teoriju u prosjeku padne za oko 20%. Pad je bio prisutan i dva mjeseca poslije. Učinak se pojavio kod klasičnih teorija zavjere (ubojstvo Johna F. Kennedyja, izvanzemaljci, Iluminati) i aktualnih (COVID-19, američki izbori 2020.), čak i među ljudima čije je uvjerenje bilo snažno i povezano s identitetom. Kada je profesionalni fact-checker ocijenio 128 činjeničnih tvrdnji koje je AI iznio, 127 ih je bilo točno, jedna obmanjujuća, a nijedna netočna.

Naslov koji je obišao svijet glasio je da ljude *možete* izvući iz „zečje rupe" razgovorom — da vjernici u teorije zavjere nisu izvan dosega dokaza, nego trebaju prave dokaze, dovoljno precizno usmjerene na ono u što vjeruju. To je doista zanimljiva tvrdnja, a studija je bila pažljivije osmišljena od većine istraživanja uvjeravanja.

Zatim je u lipnju 2026. *Science* uz rad objavio **Editorial Expression of Concern**. To je dio koji će mnoga prepričavanja preskočiti, a upravo on određuje koliku težinu rezultat trenutačno može nositi.

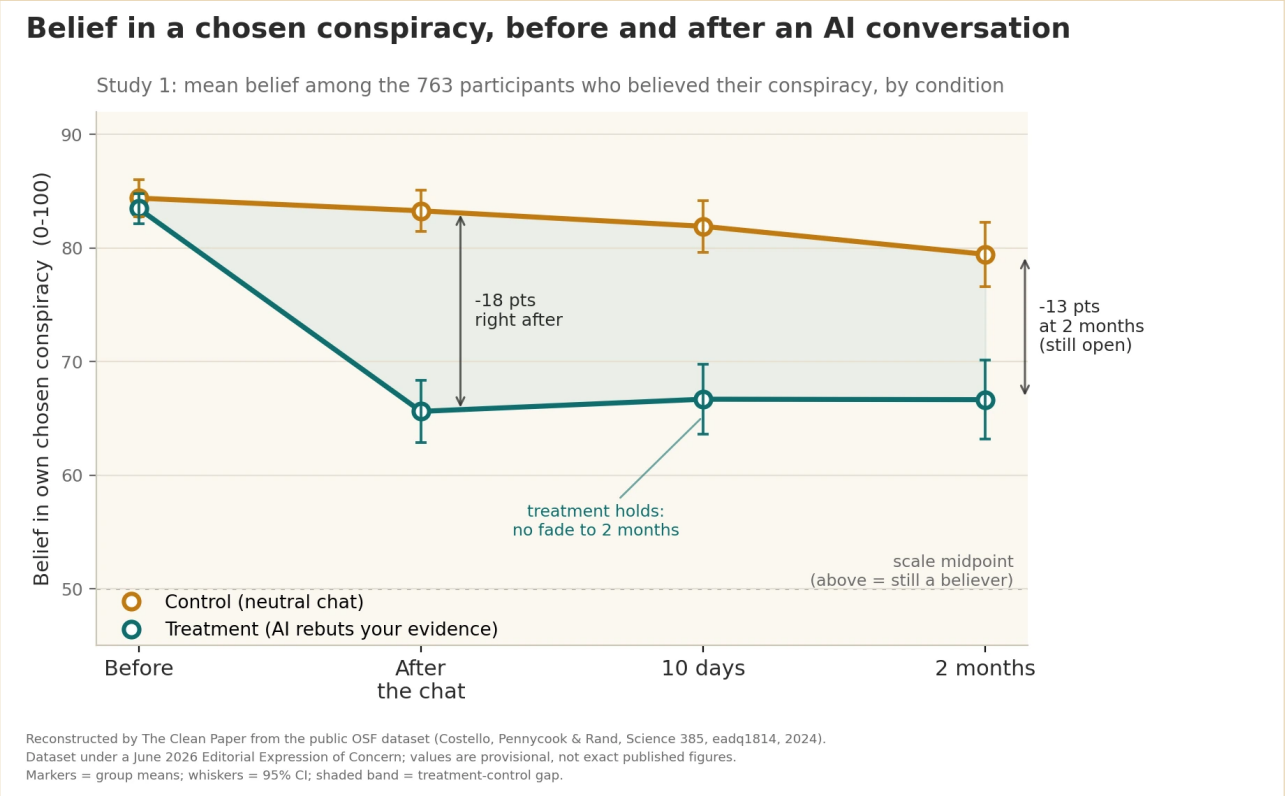
Što je Expression of Concern — a što nije

Editorial Expression of Concern formalno je upozorenje koje časopis pridružuje radu kako bi čitateljima dao do znanja da su otvorena pitanja koja se još razjašnjavaju. To **nije** povlačenje rada (rad nije uklonjen) i samo po sebi nije nalaz znanstvenog nepoštenja. Znači: čitajte uz ovu ogradu jer se znanstveni zapis možda promijeni. Ovdje su autori, nakon što su upozoreni na probleme, proveli istragu, prijavili pojedinosti časopisu *Science* i dostavili ispravljenu analizu.

Ono što je označeno vrlo je konkretno. Autori su upozoreni na nedosljednosti u tome kako su kriteriji za uključivanje sudionika primijenjeni između rukopisa i objavljenog koda za analizu te na činjenicu da je javni skup podataka sadržavao suvišne retke umetnute pogreškom pri spajanju koda — probleme zbog kojih je dio prijavljenih brojeva bilo teško reproducirati iz objavljenih materijala. Autori su *Scienceu* dostavili ispravljeni analitički postupak i skup ažuriranih rezultata za koje tvrde da se s izvornima podudaraju **u smjeru, statističkoj značajnosti i praktičnoj veličini učinka**. *Science* ih još procjenjuje. Dok ta procjena ne završi, nad točnim brojkama stoji upitnik koji može ukloniti samo

časopis.

Zato su ovo dvije priče istodobno: intrigantan rezultat o tome mogu li činjenice pomaknuti ukorijenjena uvjerenja i živi primjer znanstvenog zapisa koji se javno ispravlja. Poanta ovog teksta jest ne pomiješati ih.



U studiji je razgovor s AI-jem u tri kruga, koji je pobijao dokaze koje je sudionik sam naveo, prema objavljenim rezultatima smanjio vjerovanje u odabranu teoriju zavjere za oko 20% u prosjeku; za razliku od kontrolnog razgovora o nepovezanoj temi, pad je bio mjerljiv i nakon 10 dana i nakon 2 mjeseca. Učinak je bio trajan, ali djelomičan: većina sudionika i dalje je vjerovala u teoriju — smanjenje, ne „izlječenje". Rad je trenutačno pod Editorial Expression of Concernom (*Science*, lipanj 2026.) zbog nedosljednosti u izvještavanju i pogreške u javnom skupu podataka; autori kažu da ispravljena analiza čuva smjer, značajnost i veličinu učinka, dok *Science* nastavlja procjenu. Ovaj grafikon rekonstruirao je The Clean Paper iz tog javnog skupa podataka, pa su prikazane vrijednosti privremene, a ne konačne brojke rada.

Original figure — The Clean Paper CC BY 4.0

Što su autori napravili

- Proveli su dva eksperimenta s 2190 sudionika; svaki je vlastitim riječima naveo teoriju zavjere u koju vjeruje i dokaze za koje smatra da je podupiru.
- Svaki je sudionik vodio razgovor s GPT-4 Turbo u tri kruga. U **tretmanskom** uvjetu AI je dobio zadatak pobijati konkretne dokaze sudionika; u **kontrolnom** je razgovarao o nepovezanoj temi.
- Mjerili su vjerovanje u odabranu teoriju prije i neposredno nakon razgovora, a zatim ponovno

kontaktirali sudionike nakon **10 dana i 2 mjeseca**.
- Profesionalni fact-checker procijenio je točnost svih 128 činjeničnih tvrdnji koje je AI iznio u jednom poduzorku.
- Provjerili su prelijeva li se učinak na druge, nepovezane teorije zavjere — i, kao test specifičnosti, smanjuje li i vjerovanje u zavjere koje su se doista dogodile i stoga su **istinite**.

Što su pronašli

- **Prosječno smanjenje od približno 20%** u vjerovanju u odabranu teoriju u tretmanskoj skupini u odnosu na kontrolu (studija 1: interval pouzdanosti od 95% [13.8, 19.7] bodova na ljestvici od 100 bodova, $P < 0.001$).
- **Učinak je trajao.** U tretmanskoj skupini pad nije nestao: vjerovanje je i nakon 10 dana i dva mjeseca ostalo blizu razine neposredno nakon razgovora umjesto da se vrati prema početnoj, dok je kontrolna skupina cijelo vrijeme ostala više. Rad opisuje učinak na odabranu teoriju kao nesmanjen najmanje dva mjeseca, a ne kao trenutačan kolebljivi pomak.
- **Učinak se generalizirao** kroz širok raspon teorija koje su ljudi naveli i pojavio se i kod sudionika čije je početno uvjerenje bilo snažno.
- **Tvrdnje AI-ja bile su točne** u ovom okruženju: 127 od 128 (99.2%) ocijenjeno je točnima, 1 (0.8%) obmanjujućom, nijedna netočnom.
- **Učinak je bio specifičan**, a ne opći skepticizam: intervencija **nije** smanjila vjerovanje u istinite zavjere, što upućuje na pomak neutemeljenih uvjerenja, a ne na to da su ljudi postali cinični prema svemu.
- **Došlo je i do skromnog prelijevanja** — vjerovanje u druge, nepovezane teorije palo je za oko 12%, a sudionici su prijavili veću namjeru suprotstaviti se tvrdnjama o zavjerama. Glavni se učinak ponovio u studiji 2 i išao u istom smjeru u manjem uzorku bez kontrolne skupine (slabiji dokaz, ali dosljedan).

Što ovo ne dokazuje

- Rezultat trenutačno **nije** bez otvorenih pitanja. Rad je pod Editorial Expression of Concernom zbog problema u obradi podataka i reproducibilnosti, a ispravljene brojke *Science* još procjenjuje. Konkretne vrijednosti treba smatrati privremenima dok ta procjena ne završi.
- **Ne** pokazuje da AI „liječi” vjerovanje u teorije zavjere. Prosječno smanjenje od ~20% smislen je pomak, ne brisanje uvjerenja — većina sudionika i dalje je vjerovala u teoriju, samo manje.
- Ovo **nije** rezultat stvarne primjene u svijetu. Sudionici su dobrovoljno ušli u studiju i razgovarali s AI-jem u dobroj vjeri; izvan laboratorija ljudi najčvršće vezani uz teoriju zavjere vjerojatno su upravo najmanje skloni potražiti chatbot koji će im proturječiti.
- Točnost od 99.2% **specifična je za ovaj zadatak, model i pažljivo oblikovane upute** — nije opće jamstvo da jezični modeli govore istinu. Autori izričito navode da bi ista personalizirana moć uvjeravanja mogla gurati i *lažna* uvjerenja ako bi bila usmjerena na to.
- „Trajno” ovdje znači **dva mjeseca**, što je impresivno za jedan razgovor, ali nije isto što i trajno

zauvijek.

Koliko su dokazi jaki

- **Dizajn je prava snaga studije.** Kontrolirani eksperimenti tretman nasuprot kontroli, čista placebo-kontrola, ponavljanje kroz dvije studije i neovisni uzorak, praćenje trajnosti te eksplicitna provjera točnosti tvrdnji samog AI-ja — sve je to pažljivije od većine istraživanja uvjeravanja, a smjer učinka dosljedan je gdje god je testiran.
- **Ali Expression of Concern nije fusnota.** Mogućnost da se prijavljene brojke ponovno dobiju iz objavljenih podataka i koda dio je onoga što rezultat čini pouzdanim, a upravo je to zakazalo — zbog nedosljednih kriterija uključivanja i dupliciranih redaka iz pogreške pri spajanju koda. Tvrdnja autora da ispravljeni postupak čuva rezultat ohrabrujuća je i moguća, ali još je to njihov vlastiti izvještaj, ne neovisna presuda. Treba pričekati procjenu *Sciencea*.
- **Pošten status jest „označeno upozorenjem", ne „opovrgnuto" ni „potvrđeno".** Dobro osmišljena, upečatljiva studija čije su točne brojke pod formalnom provjerom. To je neugodno mjesto za ostaviti dobru priču, ali je točno.

Zašto je to važno

Godinama je utjecajno stajalište tvrdilo da vjernike u teorije zavjere pokreću psihološke potrebe i identitet te da ih se zato činjenicama ne može razuvjeriti. Trajno smanjenje temeljeno na dokazima — ako se održi — važna je protuteža tom pesimizmu: upućuje da je dio problema možda bio u tome što općenito pobijanje nikada nije zahvaćalo *konkretne* dokaze koje pojedini vjernik doista navodi, a LLM to može raditi u velikom opsegu.

Važno je i kao studija slučaja kako bi znanost trebala raditi. Pouka Expression of Concerna nije da su proslavljeni rezultati bezvrijedni; pouka je da pogreške izlaze na vidjelo, podaci se ponovno ispituju, a tvrdnje javno provjeravaju. To što taj mehanizam radi značajka je sustava, ne skandal — i razlog zašto je pravi stav prema ovom rezultatu strpljenje, a ne ni pljesak ni odbacivanje.

I priča o AI-ju ide u oba smjera. Ista sposobnost da prilagođenim argumentom oslabi lažno uvjerenje može ga jednako učinkovito i ojačati. Tehnika je neutralna; cilj nije.

Sažetak bez uljepšavanja

Studija iz 2024. pokazala je da je razgovor s GPT-4 Turbo u tri kruga smanjio vjerovanje ljudi u teoriju zavjere koju su sami zastupali za oko 20%, uz učinak koji je trajao dva mjeseca i pojavljivao se kroz različite vrste teorija, dok su činjenične tvrdnje AI-ja ocijenjene gotovo potpuno točnima. U lipnju 2026. rad je stavljen pod Editorial Expression of Concern zbog problema u obradi podataka i re-

producibilnosti; autori izvještavaju da ispravljena analiza čuva smjer, značajnost i veličinu nalaza, a *Science* još procjenjuje. Treba ga čitati kao doista zanimljiv i pažljivo dizajniran rezultat koji je trenutačno pod provjerom — nije zaključen, ali nije ni opovrgnut. Najpoštenija rečenica o njemu ona je koju sam proces upravo piše: provjerite ponovno.

Izvori

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