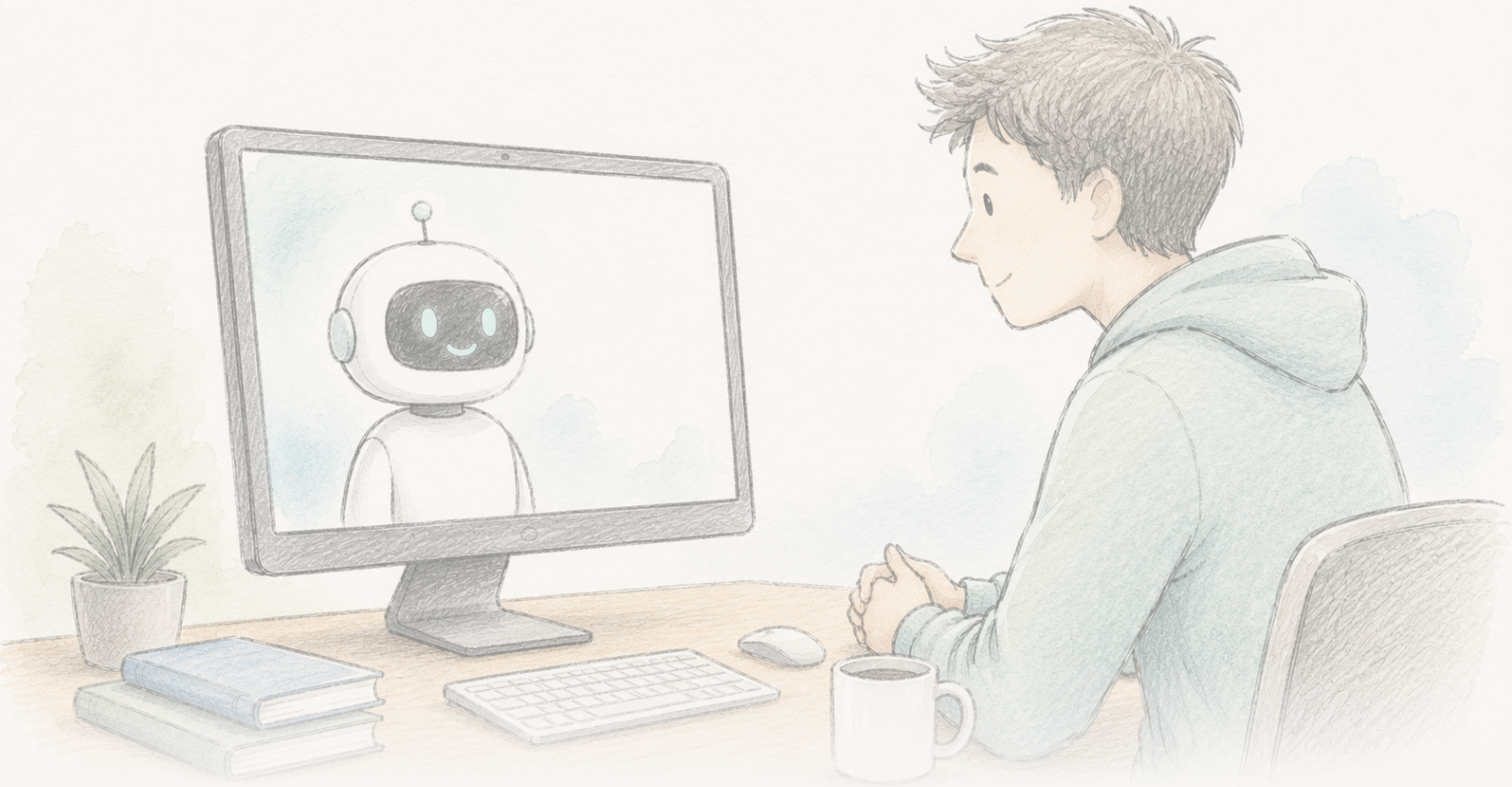


The Clean Paper

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

Čini se da je chatbot mesecima smanjivao verovanje u teorije zavere

Studija iz 2024. pokazala je da je razgovor s GPT-4 Turbo u tri kruga smanjio verovanje ljudi u teoriju zavere koju su zastupali za oko 20%, uz efekat koji je trajao dva meseca, pojavljivao se kroz različite vrste teorija i počivao na činjeničnim tvrdnjama AI-ja ocenjenima gotovo potpuno tačnima. U junu 2026. rad je stavljen pod Editorial Expression of Concern zbog problema u obradi podataka i reproducibilnosti; autori kažu da ispravljena analiza čuva smer, značajnost i veličinu nalaza, a Science ga još procenjuje. Zaista zanimljiv, pažljivo dizajniran rezultat — trenutno pod proverom, 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 procenjuje pitanja obrade podataka i reproducibilnosti. To nije povlačenje rada niti nalaz naučvenog nepoštenja; znači da prijavljeni rezultat treba smatrati privremenim dok se provera ne razreši.

Editorial Expression of Concern →

Činjenice ipak mogu delovati — ali rad sada nosi upozorenje

Studija iz 2024. izvestila je o rezultatu koji ide protiv jednog prijatnog dela 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 zavere u koju lično veruje — i njeno verovanje u tu teoriju u proseku padne za oko 20%. Pad je bio prisutan i dva meseca posle. Efekat se pojavio kod klasičnih teorija zavere (ubistvo Džona F. Kenedija, vanzemaljci, Iluminati) i aktuelnih (COVID-19, američki izbori 2020.), čak i među ljudima čije je uverenje bilo snažno i povezano s identitetom. Kada je profesionalni fact-checker ocenio 128 činjeničnih tvrdnji koje je AI izneo, 127 ih je bilo tačno, jedna obmanjujuća, a nijedna netočna.

Naslov koji je obišao svet glasio je da ljude *možete* izvući iz „zečje rupe" razgovorom — da vernici u teorije zavere nisu izvan dometa dokaza, nego trebaju prave dokaze, dovoljno precizno usmerene na ono u što veruju. To je zaista zanimljiva tvrdnja, a studija je bila pažljivije osmišljena od većine istraživanja uveravanja.

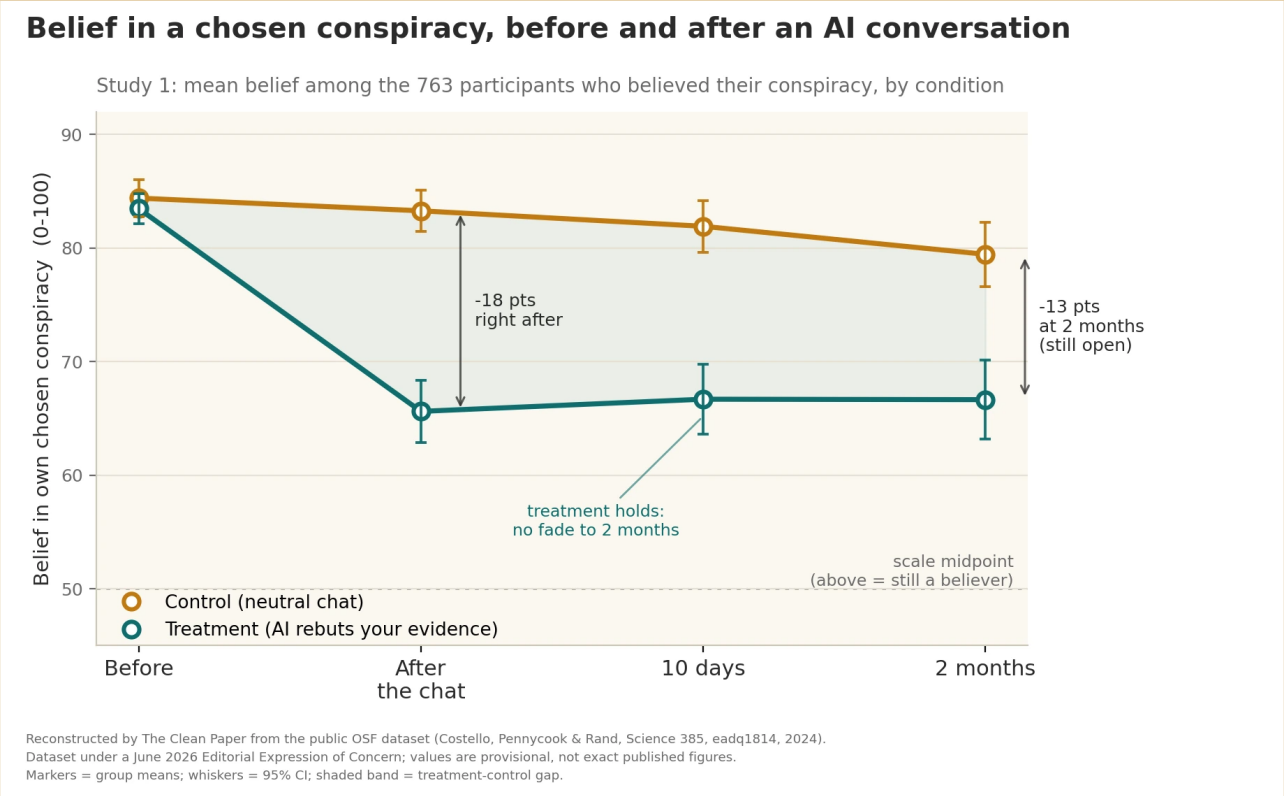
Zatim je u junu 2026. *Science* uz rad objavio **Editorial Expression of Concern**. To je deo koji će mnoga prepričavanja preskočiti, a upravo on određuje koliku težinu rezultat trenutno može nositi.

Šta je Expression of Concern — a što nije

Editorial Expression of Concern formalno je upozorenje koje časopis pridružuje radu kako bi čitalacima dao do znanja da su otvorena pitanja koja se još razjašnjavaju. To **nije** povlačenje rada (rad nije uklonjen) i samo po sebi nije nalaz naučvenog nepoštenja. Znači: čitajte uz ovu ogradu jer se naučni zapis možda promeni. Ovde 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 nedoslednosti u tome kako su kriterijumi za uključivanje učesnika primenjeni između rukopisa i objavljenog koda za analizu te na činjenicu da je javni skup podataka sadržavao suvišne retke umetnute greškom pri spajanju koda — probleme zbog kojih je deo 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 originalima podudaraju **u smeru, statističkoj značajnosti i praktičnoj veličini efekta**. *Science* ih još procenjuje. Dok ta procena ne završi, nad tačnim brojkama stoji upitnik koji može ukloniti samo časopis.

Zato su ovo dve priče istovremeno: intrigantan rezultat o tome mogu li činjenice pomeriti ukorenjena uverenja i živi primer naučvenog zapisa koji se javno ispravlja. Poenta ovog teksta je ne pomešati ih.



U studiji je razgovor s AI-jem u tri kruga, koji je pobijao dokaze koje je učesnik sam naveo, prema objavljenim rezultatima smanjio verovanje u odabranu teoriju zavere za oko 20% u proseku; za razliku od kontrolnog razgovora o nepovezanoj temi, pad je bio merljiv i nakon 10 dana i nakon 2 meseca. Efekat je bio trajan, ali delomičan: većina učesnika i dalje je verovala u teoriju — smanjenje, ne „izlječenje". Rad je trenutno pod Editorial Expression of Concernom (*Science*, jun 2026.) zbog nedoslednosti u izveštavanju i greške u javnom skupu podataka; autori kažu da ispravljena analiza čuva smer, značajnost i veličinu efekta, dok *Science* nastavlja procenu. Ovaj grafikon rekonstruisao je The Clean Paper iz tog javnog skupa podataka, pa su prikazane vrednosti privremene, a ne konačne brojke rada.

Original figure — The Clean Paper CC BY 4.0

Šta su autori uradili

- Proveli su dva eksperimenta s 2190 učesnika; svaki je sopstvenim rečima naveo teoriju zavere u koju veruje i dokaze za koje smatra da je podržavaju.
- Svaki je učesnik vodio razgovor s GPT-4 Turbo u tri kruga. U **tretmanskom** uslovu AI je dobio zadatak pobijati konkretne dokaze učesnika; u **kontrolnom** je razgovarao o nepovezanoj temi.
- Merili su verovanje u odabranu teoriju pre i neposredno nakon razgovora, a zatim ponovo kontaktirali učesnike nakon **10 dana i 2 meseca**.
- Profesionalni fact-checker procenio je tačnost svih 128 činjeničnih tvrdnji koje je AI izneo u jednom poduzorku.

- proverili su da li se preliva efekat na druge, nepovezane teorije zavere — i, kao test specifičnosti, da li smanjuje i verovanje u zavere koje su se zaista dogodile i stoga su **istinite**.

Šta su pronašli

- **Prosečno smanjenje od približno 20%** u verovanju u odabranu teoriju u tretmanskoj grupi u odnosu na kontrolu (studija 1: interval pouzdanosti od 95% [13.8, 19.7] bodova na skali od 100 bodova, $P < 0.001$).
- **Efekat je trajao.** U tretmanskoj grupi pad nije nestao: verovanje je i nakon 10 dana i dva meseca ostalo blizu nivoi neposredno nakon razgovora umesto da se vrati prema početnoj, dok je kontrolna grupa celo vreme ostala više. Rad opisuje efekat na odabranu teoriju kao nesmanjen najmanje dva meseca, a ne kao trenutan kolebljivi pomak.
- **Efekat se generalizirao** kroz širok raspon teorija koje su ljudi naveli i pojavio se i kod učesnika čije je početno uverenje bilo snažno.
- **Tvrdnje AI-ja bile su tačne** u ovom okruženju: 127 od 128 (99.2%) ocenjeno je tačnima, 1 (0.8%) obmanjujućom, nijedna netačnom.
- **Efekat je bio specifičan**, a ne opšti skepticizam: intervencija **nije** smanjila verovanje u istinite zavere, što upućuje na pomak nezasnovanih uverenja, a ne na to da su ljudi postali cinični prema svemu.
- **Došlo je i do skromnog prelivanja** — verovanje u druge, nepovezane teorije palo je za oko 12%, a učesnici su prijavili veću nameru suprotstaviti se tvrdnjama o zavjerama. Glavni se efekat ponovio u studiji 2 i išao u istom smeru u manjem uzorku bez kontrolne grupe (slabiji dokaz, ali dosledan).

Šta ovo ne dokazuje

- Rezultat trenutno **nije** bez otvorenih pitanja. Rad je pod Editorial Expression of Concernom zbog problema u obradi podataka i reproducibilnosti, a ispravljene brojke *Science* još procenjuje. Konkretne vrednosti treba smatrati privremenima dok ta procena ne završi.
- **Ne** pokazuje da AI „leči" verovanje u teorije zavere. Prosečno smanjenje od ~20% smislen je pomak, ne brisanje uverenja — većina učesnika i dalje je verovala u teoriju, samo manje.
- Ovo **nije** rezultat stvarne primene u svetu. Učesnici su dobrovoljno ušli u studiju i razgovarali s AI-jem u dobroj veri; izvan laboratorija ljudi najčvršće vezani uz teoriju zavere verovatno su upravo najmanje skloni potražiti chatbot koji će im protivrečiti.
- Tačnost od 99.2% **specifična je za ovaj zadatak, model i pažljivo oblikovane uputstva** — nije opšte garancija da jezički modeli govore istinu. Autori izričito navode da bi ista personalizovana moć uveravanja mogla gurati i *lažna* uverenja ako bi bila usmerena na to.
- „Trajno" ovde znači **dva meseca**, što je impresivno za jedan razgovor, ali nije isto što i trajno zauvek.

Koliko su dokazi jaki

- **Dizajn je prava snaga studije.** kontrolisani eksperimenti tretman nasuprot kontroli, čista placebo-kontrola, ponavljanje kroz dve studije i nezavisni uzorak, praćenje trajnosti te eksplicitna provera tačnosti tvrdnji samog AI-ja — sve je to pažljivije od većine istraživanja uveravanja, a smer efekta dosledan je gde god je testiran.
- **Ali Expression of Concern nije fusnota.** Mogućnost da se prijavljene brojke ponovo dobiju iz objavljenih podataka i koda deo je onoga što rezultat čini pouzdanim, a upravo je to zakazalo — zbog nedosljednih kriterijuma uključivanja i dupliciranih redaka iz greške pri spajanju koda. Tvrdnja autora da ispravljeni postupak čuva rezultat ohrabrujuća je i moguća, ali još je to njihov sopstveni izveštaj, ne nezavisna presuda. Treba pričekati procenu *Sciencea*.
- **Pošten status je „označeno upozorenjem", ne „opovrgnuto" ni „potvrđeno".** Dobro osmišljena, upečatljiva studija čije su tačne brojke pod formalnom proverom. To je neprijatno mesto za ostaviti dobru priču, ali je tačno.

Zašto je to važno

Godinama je uticajno stajalište tvrdilo da vernike u teorije zavere pokreću psihološke potrebe i identitet te da ih se zato činjenicama ne može razuveriti. Trajno smanjenje zasnovano na dokazima — ako se održi — važna je protuteža tom pesimizmu: upućuje da je deo problema možda bio u tome što uopšteno pobijanje nikada nije zahvaćalo *konkretne* dokaze koje pojedini vernik zaista navodi, a LLM to može raditi u velikom opsegu.

Važno je i kao studija slučaja kako bi nauka trebala raditi. Pouka Expression of Concerna nije da su proslavljeni rezultati bezvredni; pouka je da greške izlaze na videlo, podaci se ponovo ispituju, a tvrdnje javno proveravaju. To što taj mehanizam radi karakteristika je sistema, 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 smera. Ista sposobnost da prilagođenim argumentom oslabi lažno uverenje može ga jednako efikasno i ojačati. Tehnika je neutralna; cilj nije.

Kratak rezime

Studija iz 2024. pokazala je da je razgovor s GPT-4 Turbo u tri kruga smanjio verovanje ljudi u teoriju zavere koju su sami zastupali za oko 20%, uz efekat koji je trajao dva meseca i pojavljivao se kroz različite vrste teorija, dok su činjenične tvrdnje AI-ja ocenjene gotovo potpuno tačnima. U junu 2026. rad je stavljen pod Editorial Expression of Concern zbog problema u obradi podataka i reproducibilnosti; autori izveštavaju da ispravljena analiza čuva smer, značajnost i veličinu nalaza, a *Science* još procenjuje. Treba ga čitati kao zaista zanimljiv i pažljivo dizajniran rezultat koji je trenutno pod proverom — nije zaključen, ali nije ni opovrgnut. Najpoštenija rečenica o njemu ona je koju sam proces upravo piše: proverite ponovo.

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