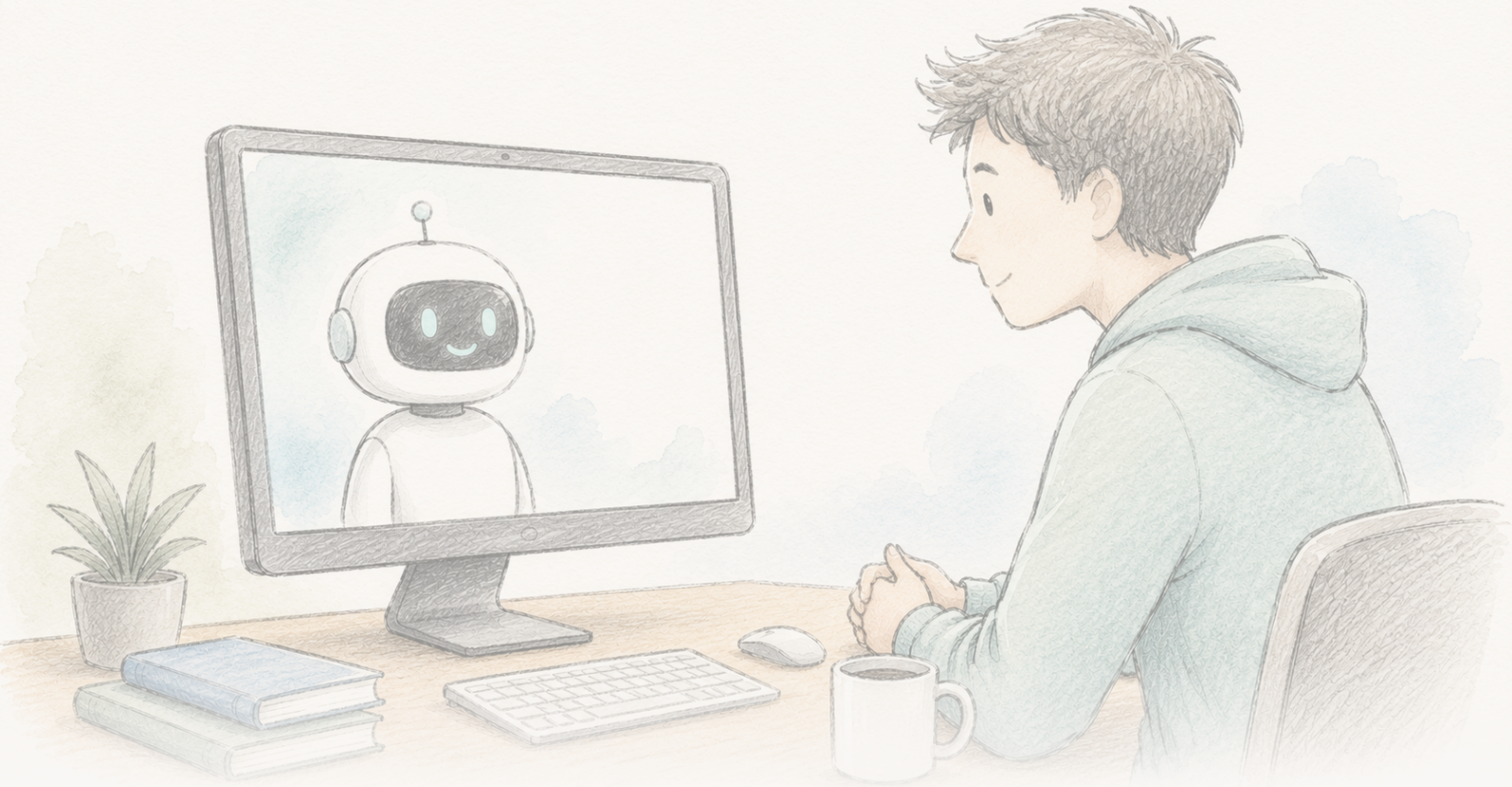


The Clean Paper

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

Pogovor z modelom AI je očitno zmanjšal verovanje v teorije zarote za več mesecev

Študija iz leta 2024 je ugotovila, da je trikraten pogovor z GPT-4 Turbo zmanjšal prepričanje ljudi v teorijo zarote, ki so ji verjeli, za približno 20 %, učinek pa je trajal dva meseca, se pojavljal pri različnih vrstah zarot in temeljil na trditvah AI, ki so bile ocenjene kot skoraj v celoti pravilne. Junija 2026 je revija Science članku dodala Editorial Expression of Concern zaradi težav z obdelavo podatkov in ponovljivostjo; avtorji pravijo, da popravljena analiza ohrani smer, statistično značilnost in velikost učinka, Science pa rezultat še ocenjuje. Resnično zanimiv in skrbno zasnovan rezultat — trenutno v postopku preverjanja, ne potrjen in ne ovržen.

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

Revija Science je temu članku dodala Editorial Expression of Concern, medtem ko ocenjuje vprašanja glede obdelave podatkov in ponovljivosti. To ni umik članka in ne ugotovitev kršitve raziskovalne etike; pomeni, da je treba rezultat do zaključka pregleda obravnavati kot začasen.

Editorial Expression of Concern →

Dejstva vendarle lahko pomagajo — vendar je članek trenutno označen z opozorilom

Leta 2024 je študija poročala o rezultatu, ki nasprotuje precej udobnemu delu splošno sprejete modrosti. Če človeku omogočite kratek pogovor v treh krogih z AI — GPT-4 Turbo, ki dobi navodilo, naj odgovori na konkretne dokaze, s katerimi posameznik utemeljuje teorijo zarote, v katero sam verjame — se njegovo prepričanje v povprečju zmanjša za približno 20 %. Zmanjšanje je bilo še vedno prisotno dva meseca pozneje. Učinek se je pojavil pri klasičnih zarotah (atentat na Johna F. Kennedyja, vesoljci, iluminati) in aktualnih (COVID-19, ameriške volitve leta 2020), celo pri ljudeh, katerih prepričanje je bilo močno in povezano z identiteto. Poklicni preverjevalec dejstev je ocenil 128 trditev, ki jih je izrekla AI: 127 je bilo pravilnih, ena zavajajoča, nobena pa napačna.

Naslov, ki se je razširil, je bil, da je ljudi *mogoče* s pogovorom spraviti iz zajčje luknje — da vernikov v teorije zarote dokazi vendarle lahko dosežejo, če so pravi dokazi dovolj konkretno prilagojeni njihovim argumentom. To je resnično zanimiva trditev, študija pa je bila zasnovana skrbneje kot večina raziskav prepričevanja.

Nato je junija 2026 revija *Science* članku dodala **Editorial Expression of Concern**. Ta del bodo številne obnove preskočile, vendar prav on določa, koliko teže lahko rezultatu trenutno pripišemo.

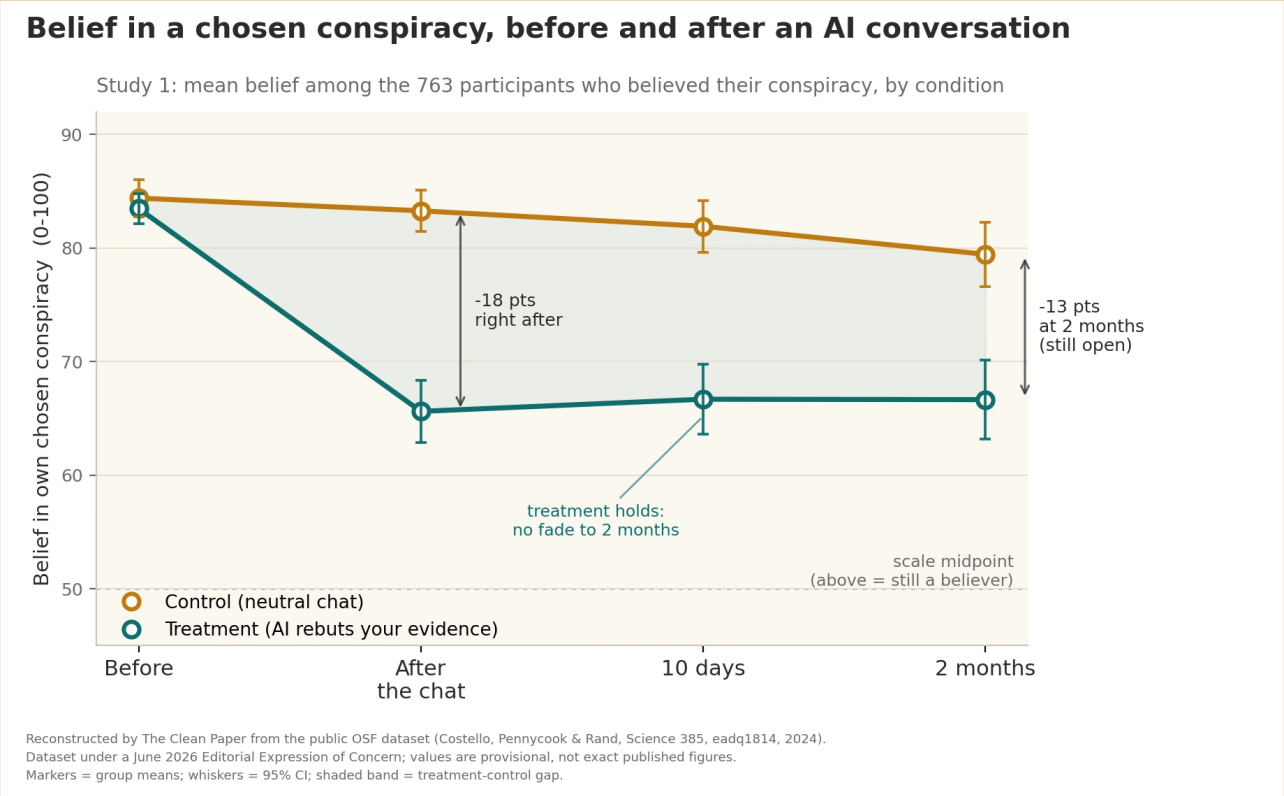
Kaj je Editorial Expression of Concern — in kaj ni

Editorial Expression of Concern je formalno opozorilo, ki ga revija pripne članku, kadar so se pojavila vprašanja, ki še niso razrešena. **Ni** umik članka (članek ni odstranjen) in samo po sebi ni ugotovitev raziskovalne kršitve. Pomeni: berite članek z navedenim zadržkom, saj se lahko znanstveni zapis spremeni. V tem primeru so avtorji po opozorilu na težave izvedli pregled, podrobnosti sporočili reviji Science in posredovali popravljeno analizo.

Označene težave so konkretne. Avtorji so bili opozorjeni na neskladnosti med kriteriji za presejanje udeležencev, kot so bili opisani v rokopisu, in njihovo uporabo v objavljeni analizni kodi. Poleg tega je javni nabor podatkov vseboval odvečne vrstice, ki so se vanj vključile zaradi napake pri združevanju kode — zaradi česar nekaterih poročanih številk iz objavljenih materialov ni bilo mogoče preprosto reproducirati. Avtorji so reviji *Science* predložili popravljeno analizno verigo in posodobljene rezultate, za katere pravijo, da se z izvirnimi ujemajo **po smeri, statistični značilnosti in vsebinski velikosti**. *Science* jih še ocenjuje. Dokler pregled ni končan, so natančne številke začasne in samo revija lahko

to opozorilo odstrani.

Zato sta tu hkrati dve zgodbi: zanimiv rezultat o tem, ali lahko dejstva premaknejo utrjena prepričanja, in živ primer tega, kako se znanstveni zapis javno popravlja. Namen tega besedila je, da ju ne pomešamo.



V študiji naj bi trikraten pogovor z AI, ki je odgovarjal na konkretne dokaze posameznega udeleženca, zmanjšal verovanje v njegovo izbrano teorijo zarote v povprečju za približno 20 %; v nasprotju s kontrolnim pogovorom o nepovezani temi je bilo zmanjšanje še vedno merljivo po 10 dneh in dveh mesecih. Učinek je bil trajen, vendar delen: večina udeležencev je po pogovoru teoriji še vedno verjela — manj, ne pa nič. Članek je trenutno pod Editorial Expression of Concern (*Science*, junij 2026) zaradi neskladnosti v poročanju in napake v javnem naboru podatkov; avtorji pravijo, da popravljena analiza ohrani smer, značilnost in velikost učinka, medtem ko *Science* rezultat še ocenjuje. Graf je The Clean Paper rekonstruiral iz javnega nabora podatkov, zato so prikazane vrednosti začasne in ne dokončne številke članka.

Original figure — The Clean Paper CC BY 4.0

Kaj so naredili avtorji

- Izvedli so dva poskusa z 2190 udeleženci, ki so vsak s svojimi besedami poimenovali teorijo zarote, v katero verjamejo, in dokaze, za katere menijo, da jo podpirajo.
- Vsak udeleženec je opravil pogovor v treh krogih z GPT-4 Turbo. V **intervencijski** skupini je AI dobila navodilo, naj zavrne konkretne dokaze udeleženca; v **kontrolni** skupini je z njim razpravljala o nepovezani temi.

- Verovanje v izbrano teorijo zarote so izmerili pred pogovorom in takoj po njem, nato pa udeležence ponovno kontaktirali po **10 dneh in dveh mesecih**.
- Poklicni preverjevalec dejstev je ocenil pravilnost vseh 128 dejanskih trditev, ki jih je AI izrekla v izbranem vzorcu.
- Preverili so, ali se učinek prenese tudi na druge nepovezane teorije zarote — ter kot test specifičnosti, ali bi zmanjšal tudi verovanje v zarote, za katere je znano, da so **resnične**.

Kaj so ugotovili

- **Približno 20-odstotno povprečno zmanjšanje** verovanja v izbrano teorijo zarote v intervencijski skupini glede na kontrolo (študija 1: 95-odstotni interval zaupanja [13,8, 19,7] točke na 100-točkovni lestvici, $P < 0{,}001$).
- **Učinek je trajal.** V intervencijski skupini se zmanjšanje ni izgubilo: verovanje je po 10 dneh in dveh mesecih ostalo blizu ravni takoj po pogovoru, namesto da bi se vračalo navzgor; kontrolna skupina je ves čas ostala višje. Članek učinek na izbrano teorijo opisuje kot nezmanjšan vsaj dva meseca, ne kot kratkotrajno nihanje.
- **Učinek se je pojavil pri različnih teorijah zarote** in tudi pri udeležencih, katerih začetno prepričanje je bilo močno.
- **Trditve AI so bile v tem okolju pravilne:** 127 od 128 (99,2 %) jih je bilo ocenjenih kot pravilnih, 1 (0,8 %) kot zavajajoča, nobena pa kot napačna.
- **Učinek je bil specifičen**, ne splošno povečanje dvoma: intervencija **ni** zmanjšala verovanja v resnične zarote, kar kaže, da je premikala neutemeljena prepričanja, ne pa preprosto povzročila cinizma do vsega.
- **Pojavil se je tudi manjši prenos na druga prepričanja** — verovanje v druge, nepovezane teorije zarote se je zmanjšalo za okoli 12 %, udeleženci pa so poročali o večji pripravljenosti, da nasprotujejo zarotniškim trditvam. Glavni učinek se je ponovil v drugi študiji in kazal v isto smer tudi v manjšem vzorcu brez kontrolne skupine (šibkejši dokaz, vendar skladen).

Česa to ne dokazuje

- Trenutno rezultat **ni brez vprašanj**. Članek ima Editorial Expression of Concern zaradi težav z obdelavo podatkov in ponovljivostjo, popravljene številke pa *Science* še ocenjuje. Konkretne vrednosti je do zaključka pregleda treba obravnavati kot začasne.
- Študija **ne** kaže, da AI »ozdravi« verovanje v teorije zarote. Povprečno zmanjšanje okoli 20 % je pomemben premik, ne izbris — večina udeležencev je teoriji po pogovoru še vedno verjela, le manj.
- To **ni** rezultat uporabe v resničnem svetu. Udeleženci so se prostovoljno vključili v študijo in z AI sodelovali v dobri veri; zunaj laboratorija so ljudje, ki so teoriji najbolj predani, morda prav najmanj pripravljeni poiskati model, ki jim bo ugovarjal.
- 99,2-odstotna pravilnost velja **za to nalogo, ta model in skrbno oblikovana navodila** — ni splošno zagotovilo, da jezikovni modeli govorijo resnico. Avtorji izrecno opozarjajo, da bi bila ista person-

alizirana prepričevalna moč lahko uporabljena tudi za krepitev *napačnih* prepričanj.
- »Trajno« tukaj pomeni **dva meseca**, kar je veliko za en sam pogovor, vendar ne pomeni stalno.

Kako močni so dokazi

- **Zasnova je prava prednost.** Kontrolirani poskusi intervencija proti kontroli, čista placebo podobna kontrola, ponovitev v dveh študijah in dodatnem neodvisnem vzorcu, spremljanje trajnosti ter izrecno preverjanje pravilnosti trditev AI — to je skrbneje od večine raziskav prepričevanja, smer učinka pa je povsod, kjer so jo preverili, skladna.
- **Toda Editorial Expression of Concern ni opomba pod črto.** Zmožnost ponovnega izračuna objavljenih številk iz izdanih podatkov in kode je del zanesljivosti rezultata, in prav tam je prišlo do napake — pri kriterijih presejanja in podvojenih oziroma odvečnih vrsticah zaradi združevanja kode. Trditev avtorjev, da popravljena analiza ohrani rezultat, je spodbudna in verjetna, vendar je še vedno njihova lastna ocena, ne neodvisna razsodba. Odločilen bo pregled revije *Science*.
- **Pošten status je »označeno za preverjanje«, ne »ovrženo« ali »potrjeno«.** Gre za dobro zasnovano in presenetljivo študijo, katere natančne številke so trenutno v formalnem pregledu. To ni udoben konec zgodbe, je pa natančen.

Zakaj je pomembno

Dolga leta je bilo vplivno stališče, da vernike v teorije zarote vodijo psihološke potrebe in identiteta ter da jih zato z dejstvi ni mogoče prepričati. Trajno zmanjšanje, ki temelji na dokazih — *če se rezultat potrdi* — je pomembna protiutež takemu pesimizmu: lahko pomeni, da je bil del težave v tem, da splošno razbijanje mitov ni nikoli odgovorilo na *konkretne* dokaze, ki jih posameznik dejansko navaja, medtem ko lahko LLM to počne v velikem obsegu.

Rezultat je pomemben tudi kot primer, kako naj bi znanost delovala. Nauka Editorial Expression of Concern ne gre razumeti tako, da so odmevni rezultati brez vrednosti; nauk je, da se napake pojavijo na površju, podatki ponovno pregledajo in trditve javno znova preverijo. Delovanje tega mehanizma je prednost, ne škandal — zato je do rezultata primerneje imeti potrpljenje kot pa ga nemudoma slaviti ali zavreči.

In pri AI ima zgodba dve strani. Ista sposobnost, ki lahko s prilagojenim argumentom zmanjša napačno prepričanje, bi ga lahko prav tako učinkovito okrepila. Tehnika je nevtralna; cilj ni.

Kratek povzetek

Študija iz leta 2024 je ugotovila, da je trikraten pogovor z GPT-4 Turbo zmanjšal verovanje ljudi v teorijo zarote, ki so ji verjeli, za približno 20 %, učinek pa je vztrajal dva meseca in se pojavljal pri

različnih vrstah zarot, medtem ko so bile dejanske trditve AI ocenjene kot skoraj v celoti pravilne. Junija 2026 je revija *Science* članku dodala Editorial Expression of Concern zaradi težav z obdelavo podatkov in ponovljivostjo; avtorji poročajo, da popravljena analiza ohrani smer, statistično značilnost in velikost rezultata, *Science* pa jo še ocenjuje. Berite ga kot resnično zanimiv in skrbno zasnovan rezultat, ki je trenutno v postopku preverjanja — ne kot dokončno potrjenega in ne kot ovrženega. Najbolj pošten stavek o njem je tisti, ki ga trenutno piše sam znanstveni proces: preverite še enkrat.

Viri

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