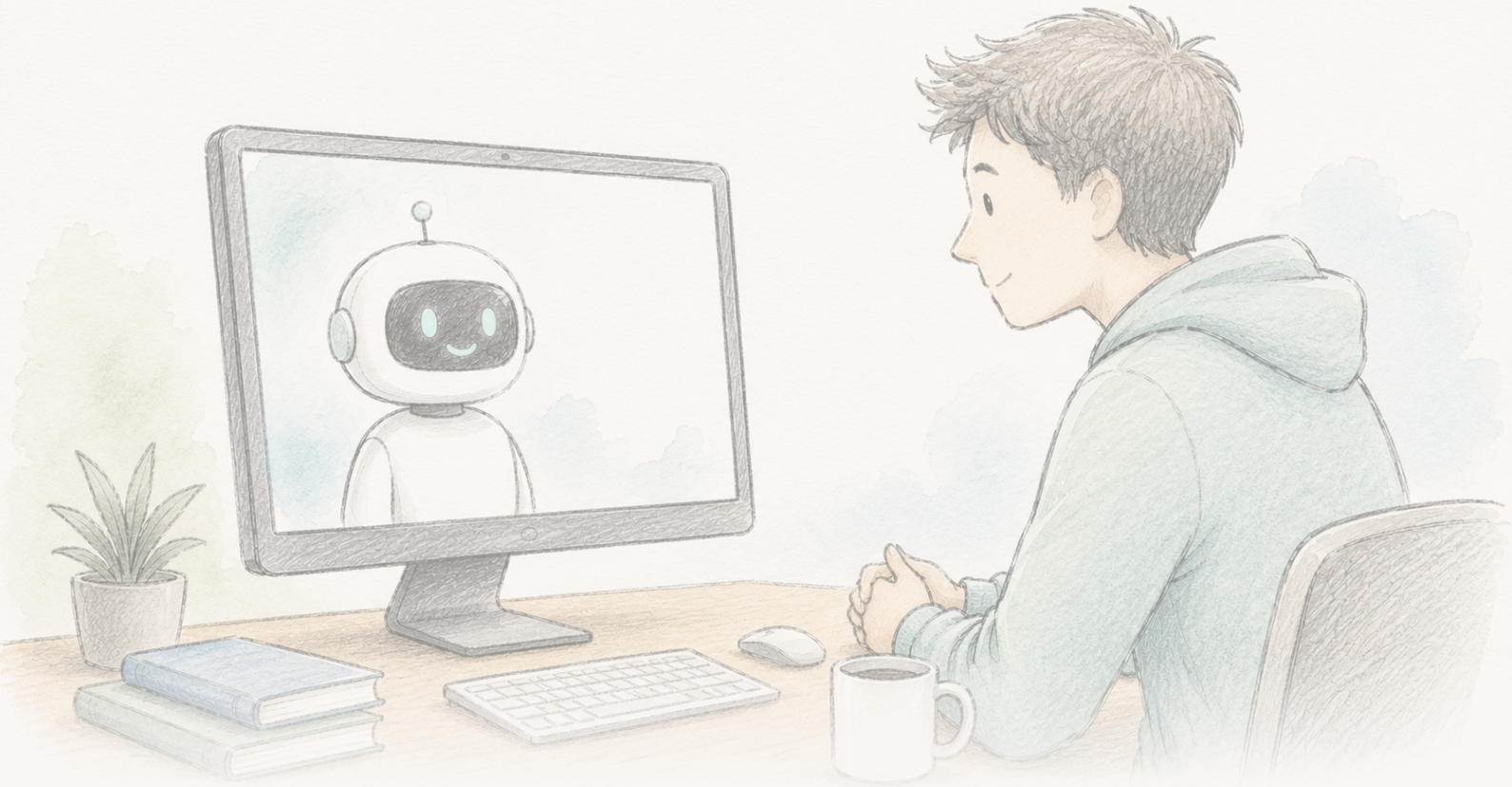


The Clean Paper

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

Tērzēšanas robots, šķiet, uz mēnešiem mazināja ticību sazvērestības teorijām

2024. gada pētījumā trīs kārtu saruna ar GPT-4 Turbo samazināja cilvēku ticību viņu pašu izvēlētai sazvērestības teorijai par aptuveni 20%; efekts saglabājās divus mēnešus, parādījās dažādu sazvērestību gadījumā, un MI apgalvojumi tika vērtēti kā gandrīz pilnībā precīzi. 2026. gada jūnijā Science rakstam pievienoja redakcionālu bažu paziņojumu datu apstrādes un reproducējamības problēmu dēļ; autori saka, ka koriģētā analīze saglabā rezultāta virzienu, nozīmīgumu un lielumu, bet Science vēl turpina izvērtēšanu. Tas ir patiesi interesants, rūpīgi veidots rezultāts, kas pašlaik tiek pārbaudīts — ne galīgi apstiprināts un ne atspēkots.

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 šim rakstam ir pievienojis redakcionālu bažu paziņojumu, kamēr tiek izvērtēti jautājumi par datu apstrādi un reproducējamību. Tas nav raksta atsaukums un nav konstatēts pārkāpums; tas nozīmē, ka publicētais rezultāts jāuzskata par provizorisku, līdz pārbaude būs pabeigta.

Editorial Expression of Concern →

Fakti tomēr var iedarboties — un pie raksta tagad ir brīdinājuma zīme

2024. gadā publicēts pētījums apstrīdēja visai ērtu priekšstatu. Ja cilvēkam ļauj īsu, trīs kārtu sarunu ar mākslīgo intelektu — GPT-4 Turbo, kam dots uzdevums atbildēt tieši uz pierādījumiem, kurus cilvēks min savas ticētās sazvērestības teorijas atbalstam —, tad vidēji pārliecība par šo teoriju samazinās par aptuveni 20%. Samazinājums bija redzams vēl pēc diviem mēnešiem. Efekts parādījās gan klasisku sazvērestības teoriju gadījumā (Džona F. Kenedija slepkavība, citplanētieši, ilumināti), gan aktuālākās tēmās (Covid-19, 2020. gada ASV vēlēšanas), arī cilvēkiem, kuru pārliecība bija stipra un cieši saistīta ar identitāti. Profesionāls faktu pārbaudītājs novērtēja 128 MI sniegtus faktoloģiskus apgalvojumus: 127 bija patiesi, viens — maldinošs, neviens — nepatiess.

Plaši izplatījās secinājums, ka cilvēku tomēr *var* izrunāt no sazvērestības „truša alas": sazvērestības teoriju piekritēji nav neaizsniedzami pierādījumiem, viņiem vienkārši vajadzīgi tieši viņu argumentiem atbilstoši pierādījumi. Tas ir patiesi interesants apgalvojums, un pētījums bija veidots rūpīgāk nekā liela daļa pārliecināšanas pētījumu.

Taču 2026. gada jūnijā *Science* rakstam pievienoja **redakcionālu bažu paziņojumu** (*Editorial Expression of Concern*). Šo daļu daudzos pārstāstos varētu izlaist, lai gan tieši tā šobrīd nosaka, cik lielu svaru drīkst piešķirt rezultātam.

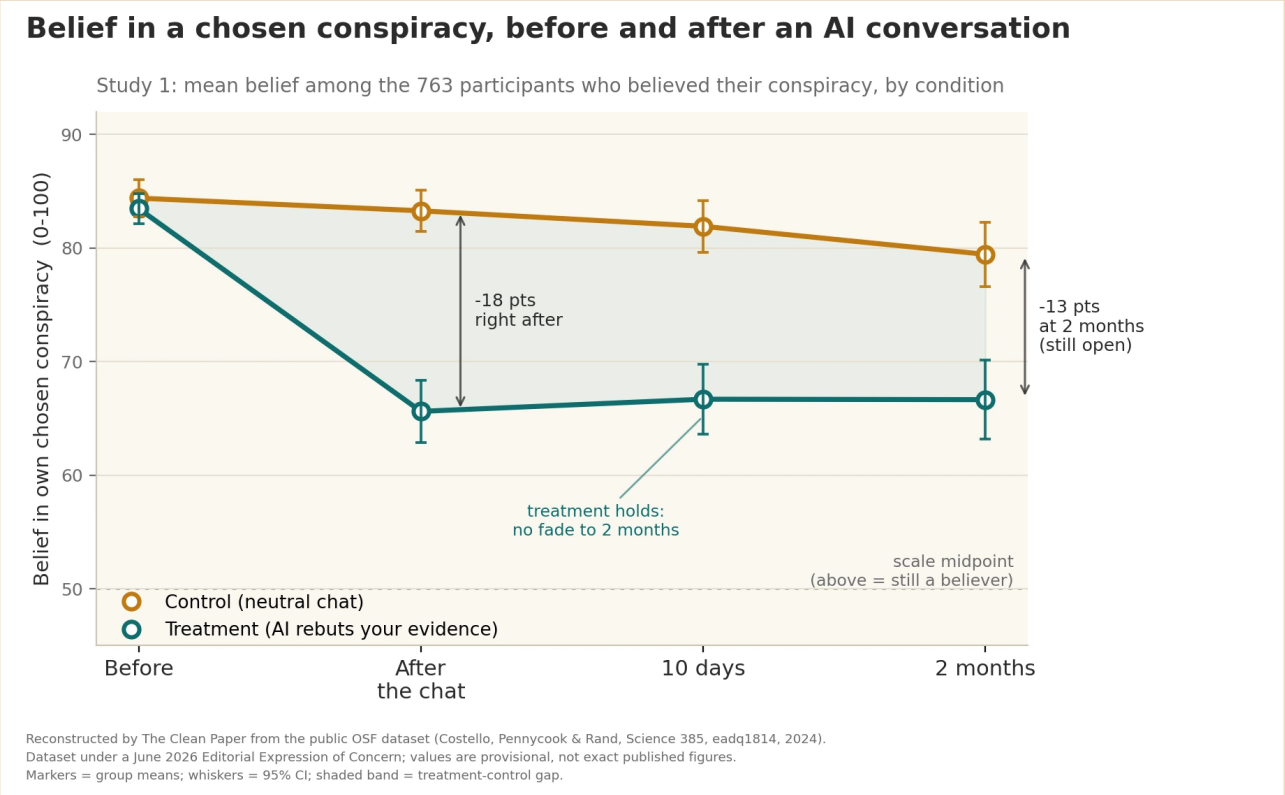
Kas ir — un kas nav — redakcionāls bažu paziņojums

Redakcionāls bažu paziņojums ir oficiāls žurnāla brīdinājums, ka par rakstu radušies jautājumi un tie vēl tiek pārbaudīti. Tas **nav** raksta atsaukums, un pats par sevi tas nav arī secinājums par zinātnisku pārkāpumu. Tas nozīmē: lasiet rezultātu, paturot prātā šo atrunu, jo zinātniskais ieraksts vēl var tikt precizēts. Šajā gadījumā autori, uzzinājuši par problēmām, tās izmeklēja, informēja *Science* par detaļām un iesniedza koriģētu analīzi.

Atzīmētās problēmas ir konkrētas. Autori uzzināja par neatbilstībām starp manuskriptu un publicēto analīzes kodu tajā, kā tika piemēroti dalībnieku atlases kritēriji. Turklāt publiskajā datu kopā bija liekas rindas, kas tajā nonākušas koda apvienošanas kļūdas dēļ. Tāpēc dažus publicētos skaitļus bija grūti atkārtoti iegūt no publiski pieejamajiem datiem un koda. Autori *Science* iesniedza labotu

analīzes procesu un atjaunināt rezultātus, kas, pēc viņu teiktā, sakrīt ar sākotnējiem **efekta virziena, statistiskās nozīmības un praktiskā lieluma ziņā**. *Science* tos vēl vērtē. Kamēr šī pārbaude nav pabeigta, precīzie skaitļi paliek provizoriski.

Tātad šeit vienlaikus ir divi stāsti: interesants rezultāts par to, vai fakti spēj mainīt iesakņojušos uzskatus, un dzīvs piemērs tam, kā zinātniskais ieraksts publiski labo pats sevi. Ir svarīgi šos abus stāstus nesajaukt.



Pētījumā tika ziņots, ka trīs kārtu saruna ar MI, kas atspēkoja paša dalībnieka minētos pierādījumus, vidēji samazināja ticību izvēlētajai sazvērestības teorijai par aptuveni 20%. Atšķirībā no kontroles sarunas par nesaistītu tēmu šis samazinājums bija izmērāms arī pēc 10 dienām un diviem mēnešiem. Efekts bija noturīgs, bet daļējs: vairums dalībnieku sazvērestības teorijai joprojām ticēja — tā bija pārliecības mazināšanās, nevis „izārstēšana". Rakstam pašlaik ir *Science* 2026. gada jūnijā pievienots redakcionāls bažu paziņojums par ziņošanas neatbilstībām un kļūdu publiskajā datu kopā. Autori norāda, ka koriģētā analīze saglabā efekta virzienu, nozīmīgumu un lielumu, kamēr *Science* turpina izvērtēšanu. Šo grafiku The Clean Paper rekonstruēja no publiskās datu kopas, tādēļ tajā redzamās vērtības ir provizoriskas, nevis raksta galīgie skaitļi.

Original figure — The Clean Paper CC BY 4.0

Ko autori darīja

- Veica divus eksperimentus ar 2190 dalībniekiem. Katrs saviem vārdiem nosauca sazvērestības teoriju, kurai ticēja, un pierādījumus, kas, viņaprāt, to atbalstīja.
- Katrs dalībnieks trīs kārtās sarunājās ar GPT-4 Turbo. **Intervences** grupā MI tika uzdots atspēkot

dalībnieka konkrētos pierādījumus; **kontroles** grupā tas runāja par nesaistītu tēmu.
- Izmērīja pārliecību par izvēlēto sazvērestību pirms sarunas un tūlīt pēc tās, pēc tam atkārtoti sazinājās ar dalībniekiem pēc **10 dienām un diviem mēnešiem**.
- Profesionālam faktu pārbaudītājam lika novērtēt visu 128 faktoloģisko apgalvojumu precizitāti atlasītā MI atbilžu paraugā.
- Pārbaudīja, vai efekts pāriet arī uz citām, nesaistītām sazvērestības teorijām, un kā specifiskuma pārbaudi — vai tas mazinātu ticību arī sazvērestībām, kas patiešām ir **notikušas**.

Ko viņi atklāja

- **Aptuveni 20% vidēju samazinājumu** ticībā izvēlētajai sazvērestības teorijai intervences grupā salīdzinājumā ar kontroli (1. pētījumā: 95% ticamības intervāls [13.8, 19.7] punkti 100 punktu skalā, $P < 0.001$).
- **Efekts saglabājās.** Intervences grupā pārliecība pēc sarunas neatgriezās iepriekšējā līmenī: pēc 10 dienām un diviem mēnešiem tā palika aptuveni tūlīt pēc sarunas sasniegtajā līmenī, kamēr kontroles grupā tā bija augstāka. Raksts šo efektu raksturo kā vismaz divus mēnešus nesamazinājušos, nevis īslaicīgu svārstību.
- **Efekts parādījās dažādu sazvērestības teoriju gadījumā** un arī cilvēkiem, kuru sākotnējā pārliecība bija stipra.
- **MI apgalvojumi šajā uzdevumā bija precīzi:** 127 no 128 (99.2%) tika vērtēti kā patiesi, 1 (0.8%) — kā maldinošs, neviens — kā nepatiess.
- **Efekts bija specifisks**, nevis vispārēja neuzticēšanās: intervence **nemazināja** ticību patiesiem sazvērestību gadījumiem. Tas liecina, ka tika mazināti nepamatoti uzskati, nevis vienkārši vairots cinisms pret visu.
- **Bija arī neliels pārneses efekts:** ticība citām, nesaistītām sazvērestības teorijām samazinājās par aptuveni 12%, un dalībnieki biežāk pauda gatavību iebilst pret sazvērestību apgalvojumiem. Galvenais efekts atkārtojās 2. pētījumā un bija tajā pašā virzienā arī mazākā izlasē bez kontroles grupas — tas ir vājāks pierādījums, taču saskanīgs.

Ko tas nepierāda

- Šis rezultāts pašlaik **nav bez jautājuma zīmes**. Rakstam ir redakcionāls bažu paziņojums par datu apstrādes un reproducējamības problēmām, un *Science* vēl vērtē koriģētos skaitļus. Līdz pārbaudes beigām konkrētās vērtības jāuzskata par provizoriskām.
- Tas **neparāda**, ka MI „izārstē" ticību sazvērestības teorijām. Aptuveni 20% vidējs samazinājums ir nozīmīga izmaiņa, nevis pārliecības izzušana — vairums dalībnieku teorijai pēc sarunas joprojām ticēja, tikai mazāk.
- Tas **nav** rezultāts no reālas plaša mēroga ieviešanas. Dalībnieki brīvprātīgi piedalījās pētījumā un godprātīgi iesaistījās sarunā ar MI; ārpus pētījuma vispārliecinātākie sazvērestības teoriju piekritēji, iespējams, vismazāk vēlēsies meklēt tērzēšanas robotu, kas viņiem iebilst.
- 99.2% precizitāte ir **specifiska šim uzdevumam, modelim un rūpīgajai uzvednei/pamudināju-**

mam, nevis vispārēja garantija, ka valodas modeļi saka patiesību. Autori skaidri norāda, ka tā pati personalizētā pārliecināšanas spēja varētu nostiprināt arī *nepatiesu* uzskatu, ja to izmantotu šādam mērķim.
- „Noturīgs" šeit nozīmē **divus mēnešus** — iespaidīgi vienai sarunai, taču ne tas pats, kas pastāvīgs efekts.

Cik pārliecinoši ir pierādījumi?

- **Pētījuma dizains ir īsts pluss.** Kontrolēti intervences un kontroles eksperimenti, skaidra placebo tipa kontrole, atkārtojums divos pētījumos un neatkarīgā izlasē, ilgtermiņa mērījumi un tieša MI apgalvojumu precizitātes pārbaude — tas ir rūpīgāk nekā daudzos pārliecināšanas pētījumos, un efekta virziens bija konsekvents visur, kur tas tika testēts.
- **Taču bažu paziņojums nav zemsvītras piezīme.** Spēja no publiskotajiem datiem un koda atkārtoti iegūt publicētos rezultātus ir daļa no tā, kas pētījumu padara uzticamu, un tieši tur radās problēma: neatbilstības atlases kritērijos un dublētas rindas koda apvienošanas kļūdas dēļ. Autoru paziņojums, ka labotais analīzes process saglabā rezultātu, ir iedrošinošs un ticams, taču tas pagaidām ir pašu autoru vērtējums, nevis neatkarīgs spriedums. Jāgaida *Science* izvērtējums.
- **Godīgais statuss ir „atzīmēts pārbaudei", nevis „atspēkots" vai „apstiprināts".** Tas ir labi veidots, pārsteidzošs pētījums, kura precīzie skaitļi pašlaik tiek formāli pārbaudīti. Tas nav tik apmierinošs noslēgums kā skaidrs „jā" vai „nē", bet tas ir precīzs.

Kāpēc tas ir svarīgi

Gadiem ilgi ietekmīgs skatījums bija tāds, ka sazvērestības teoriju piekritējus vada psiholoģiskas vajadzības un identitāte, tāpēc ar faktiem viņus pārliecināt nevar. Ja noturīgs, uz pierādījumiem balstīts samazinājums izturēs pārbaudi, tas būs nozīmīgs pretarguments šim pesimismam. Tas liecinātu, ka daļa problēmas bija tajā, ka vispārīga faktu pārbaude neatbild uz *konkrētajiem* pierādījumiem, kurus cilvēks pats min — bet LLM var to darīt lielā mērogā.

Šis gadījums ir svarīgs arī kā piemērs tam, kā zinātnei vajadzētu darboties. Redakcionāla bažu paziņojuma mācība nav tāda, ka ievērojami rezultāti ir bezvērtīgi. Mācība ir, ka kļūdas tiek pamanītas, dati tiek pārbaudīti no jauna un apgalvojumi publiski tiek atkārtoti vērtēti. Tas, ka šis mehānisms darbojas, nav skandāls — tā ir sistēmas īpašība. Tāpēc pareizā attieksme pret šo rezultātu pašlaik ir pacietība, nevis neapdomīgi aplausi vai tikpat neapdomīga noraidīšana.

Un MI gadījumā šī spēja darbojas abos virzienos. Tā pati personalizētā argumentācija, kas var vājināt nepatiesu uzskatu, varētu tikpat efektīvi to nostiprināt. Tehnika pati par sevi ir neitrāla; mērķis — ne.

Īss kopsavilkums

2024. gada pētījumā tika konstatēts, ka trīs kārtu saruna ar GPT-4 Turbo samazināja cilvēku ticību viņu pašu izvēlētai sazvērestības teorijai par aptuveni 20%; efekts saglabājās divus mēnešus un parādījās dažādu sazvērestības teoriju gadījumā, bet MI faktoloģiskie apgalvojumi tika vērtēti kā gandrīz pilnībā precīzi. 2026. gada jūnijā *Science* rakstam pievienoja redakcionālu bažu paziņojumu datu apstrādes un reproducējamības problēmu dēļ. Autori apgalvo, ka koriģētā analīze saglabā rezultāta virzienu, statistisko nozīmīgumu un lielumu, bet *Science* izvērtējums vēl turpinās. Tāpēc tas jālasa kā patiesi interesants un rūpīgi veidots rezultāts, kas pašlaik tiek pārbaudīts — ne kā galīgi apstiprināts un ne kā atspēkots. Precīzākā attieksme ir vienkārša: pārbaudīt vēlreiz.

Avoti

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