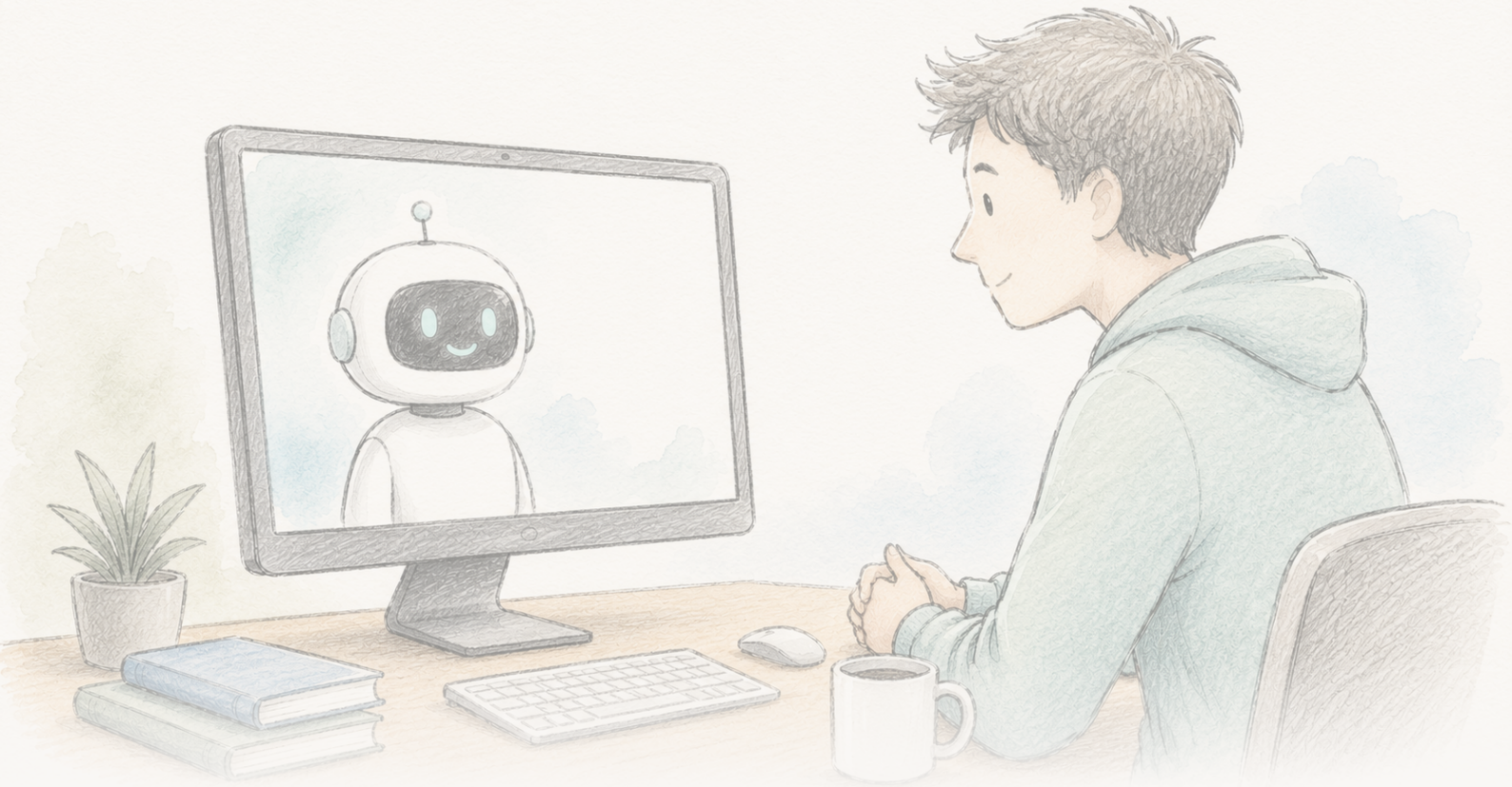


The Clean Paper

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

Vestlusrobot näis vähendavat vandenõuuskumusi kuudeks

2024. aasta uuring leidis, et kolmevooruline vestlus GPT-4 Turboga vähendas inimeste usku nende endi usutud vandenõuteooriasse umbes 20%; mõju püsis kaks kuud ja üldistus eri vandenõutüüpidele ning AI väiteid hinnati valdavalt täpseks. 2026. aasta juunis lisati artiklile andmekäitluse ja reprodutseeritavuse probleemide tõttu Editorial Expression of Concern; autorite sõnul säilitab parandatud analüüs tulemuse suuna, statistilise olulisuse ja suuruse ning Science hindab seda endiselt. Tõeliselt huvitav ja hoolikalt üles ehitatud tulemus — praegu kontrolli all, mitte lõplikult kinnitatud ega ümber lükatud.

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 on lisanud sellele artiklile Editorial Expression of Concern'i, kuni hindab andmekäitluse ja reprodutseeritavusega seotud küsimusi. See ei ole artikli tagasivõtmine ega rikkumise tuvastamine; see tähendab, et raporteeritud tulemust tuleb kuni kontrolli lõppemiseni käsitleda esialgsena.

Editorial Expression of Concern →

Faktid siiski loevad — ja artikli juures on hoiatuslipp

2024. aastal avaldatud uuring teatas tulemusest, mis läheb vastuollu ühe mugava levinud arusaamaga. Andke inimesele lühike, kolmest voorust koosnev vestlus AI-ga — GPT-4 Turboga, millele on antud ülesanne vastata just neile tõenditele, mida inimene ise oma usutud vandenõuteooria toetuseks toob — ning keskmiselt väheneb tema usk sellesse teooriasse umbes 20%. Vähenemine oli alles ka kaks kuud hiljem. See toimis nii klassikaliste vandenõude puhul (John F. Kennedy mõrv, tulnukad, illuminaadid) kui ka päevakajaliste puhul (COVID-19, USA 2020. aasta valimised) ning isegi inimestel, kelle usk oli tugev ja seotud identiteediga. Kui professionaalne faktikontrollija hindas 128 AI esitatud väidet, osutus 127 tõeseks, üks eksitavaks ja mitte ükski vääraks.

Levima läinud pealkiri oli, et inimesi *saab* jäneseurust välja rääkida — et vandenõuteooriatesse uskujad pole tõendite mõjule immuunsed, vaid vajavad lihtsalt õigeid tõendeid piisavalt isikupärasel kujul. See on tõepoolest huvitav väide ning uuring oli üles ehitatud hoolikamalt kui enamik veenmist käsitlevaid töid.

Siis, 2026. aasta juunis, lisas *Science* artiklile **toimetuse märke kahtluse kohta (Editorial Expression of Concern)**. Just selle osa jätavad paljud ümberjutustused vahele, kuigi praegu otsustab see suuresti, kui palju raskust tulemusele panna saab.

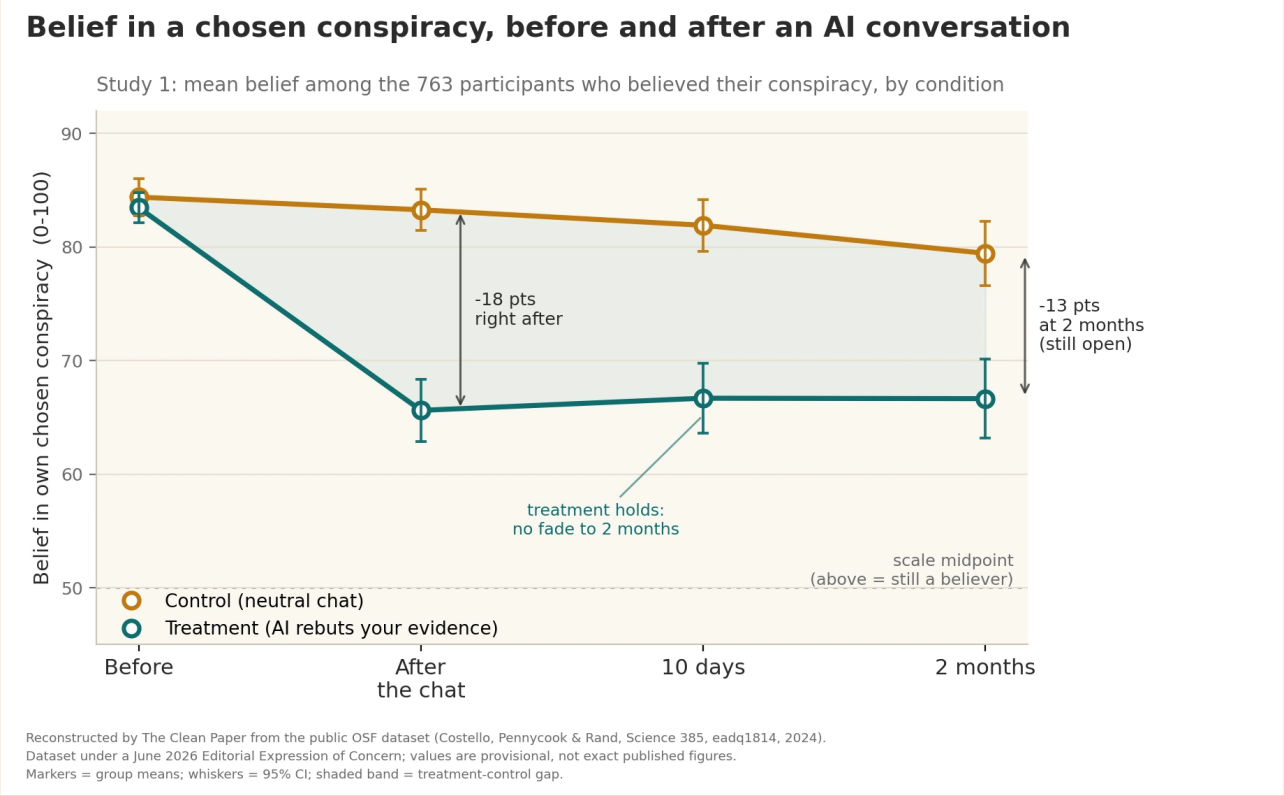
Mis on Expression of Concern — ja mis see ei ole

Editorial Expression of Concern on ametlik märge, mille ajakiri lisab artiklile, et lugejaid teavitada tekkinud küsimustest ajal, mil neid küsimusi veel uuritakse. See **ei ole** tagasivõtmine (artiklit ei ole eemaldatud) ega ole iseenesest rikkumise tuvastamine. See tähendab: lugege tulemust selle reservatsiooniga, sest teaduskirje võib muutuda. Siin uurisid autorid neile teatatud probleeme, edastasid üksikasjad *Science'ile* ning esitasid parandatud analüüsi.

Esile toodud probleemid on konkreetsed. Autoritele juhiti tähelepanu sellele, et osalejate sõelumiskriteeriume rakendati käsikirjas ja avaldatud analüüsikoodis ebajärjekindlalt ning et avalik andmestik sisaldas üleliigseid ridu, mis sattusid sinna koodi ühendamisel tehtud vea tõttu — probleemid, mis muutsid osa raporteeritud arvudest avaldatud materjalide põhjal raskesti reprodutseeritavaks. Autorid andsid *Science'ile* parandatud analüüsiahela ja uuendatud tulemused, mis nende sõnul vastavad algsetele **suuna, statistilise olulisuse ja sisulise suuruse poolest**. *Science* hindab neid

praegu. Seni, kuni see hindamine lõpeb, jäävad täpsed arvud küsimärgi alla, mille saab eemaldada ainult ajakiri.

Seega on siin korraga kaks lugu: intrigeeriv tulemus selle kohta, kas faktid suudavad juurdunud uskumusi nihutada, ning elav näide sellest, kuidas teaduskirje end avalikult parandab. Selle loo mõte on hoida need kaks asja lahus.



Uuringu järgi vähendas kolmest voorust koosnev AI-vestlus, mis lükkas ümber osaleja enda nimetatud tõendeid, keskmiselt umbes 20% selle inimese usku valitud vandenõuteooriasse; erinevalt kontrollvestlusest seosetul teemal oli langus mõõdetav ka 10 päeva ja 2 kuu pärast. Mõju oli püsiv, kuid osaline: enamik osalejaid uskus vandenõuteooriat ka hiljem — lihtsalt vähem. Artiklil on praegu *Science'i* Editorial Expression of Concern (juuni 2026), sest raporteerimises leiti ebakõlasid ja avalikus andmestikus viga; autorite sõnul säilitab parandatud analüüs mõju suuna, statistilise olulisuse ja suuruse, samal ajal kui *Science* jätkab hindamist. Selle graafiku rekonstrueeris The Clean Paper avalikust andmestikust, seega on näidatud väärtused esialgsed, mitte artikli lõplikud arvud.

Original figure — The Clean Paper CC BY 4.0

Mida autorid tegid

- Viisid läbi kaks katset 2190 osalejaga, kellest igaüks nimetas oma sõnadega vandenõuteooria, millesse ta uskus, ning tõendid, mis tema arvates seda toetasid.
- Lasid igal osalejal pidada GPT-4 Turboga kolmevoorulise vestluse. **Sekkumisrühmas** pidi AI vastama osaleja konkreetsetele tõenditele ja neid ümber lükkama; **kontrollrühmas** arutati seosetut

teemat.
- Mõõtsid usku valitud vandenõuteooriasse enne vestlust ja vahetult pärast seda ning võtsid osalejatega uuesti ühendust **10 päeva ja 2 kuu** pärast.
- Lasid professionaalsel faktikontrollijal hinnata valimis kõigi 128 AI esitatud faktiväite täpsust.
- Kontrollisid, kas mõju kandus üle teistele, seosetutele vandenõuteooriatele, ning spetsiifilisuse testina ka seda, kas see vähendas usku vandenõudesse, mis juhtumisi on **tõesed**.

Mida nad leidsid

- **Ligikaudu 20% keskmine vähenemine** usus valitud vandenõuteooriasse sekkumisrühmas võrreldes kontrolliga (uuring 1: 95% usaldusvahemik [13.8, 19.7] punkti 100-punktilisel skaalal, $P < 0.001$).
- **Mõju püsis.** Sekkumisrühmas ei kadunud langus ära: usk jäi 10 päeva ja kahe kuu järel vestlusjärgse taseme lähedale, selle asemel et tagasi tõusta, samal ajal kui kontrollrühma tase püsis kogu aeg kõrgem. Artikkel kirjeldab mõju valitud vandenõule kui vähemalt kaks kuud vaibumatult püsinut, mitte hetkelist kõikumist.
- **Mõju üldistus** eri tüüpi vandenõuteooriatele, mida inimesed nimetasid, ning püsis ka nende osalejate seas, kelle esialgne usk oli tugev.
- **AI väited olid selles olukorras täpsed:** 127 väidet 128-st (99,2%) hinnati tõeseks, 1 (0,8%) eksitavaks ja mitte ühtegi vääraks.
- **Mõju oli spetsiifiline**, mitte üldine kahtlustamine: sekkumine **ei** vähendanud usku tõestesse vandenõudesse, mis viitab sellele, et see nihutas toetamatuid uskumusi, mitte ei muutnud inimesi kõige suhtes küüniliseks.
- **Mõju kandus mõõdukalt üle** — usk teistesse, seosetutesse vandenõudesse langes umbes 12% ning osalejad teatasid suuremast kavatsusest vandenõuväidetele vastu vaielda. Põhimõju kordus uuringus 2 ning sama suunda näitas väiksem kontrollrühmata valim (nõrgem tõendus, kuid kooskõlaline).

Mida see ei tõesta

- Tulemus **ei ole** praegu vaidlustest vaba. Artikli juures on andmekäitluse ja reprodutseeritavuse probleemide tõttu Editorial Expression of Concern ning parandatud arve hindab *Science* endiselt. Konkreetseid väärtusi tuleb käsitleda esialgsetena, kuni see kontroll lõpeb.
- See **ei** näita, et AI „ravib“ vandenõuuskumusi. Umbes 20% keskmine vähenemine on tähenduslik nihe, mitte uskumuse kustutamine — enamik osalejaid uskus teooriat pärast vestlust endiselt, lihtsalt vähem.
- See **ei ole** päriselu kasutuselevõtu tulemus. Osalejad nõustusid uuringus osalema ja suhtlesid AI-ga heas usus; päriselus on kõige pühendunumad vandenõuteooriauskujad tõenäoliselt just need, kes kõige vähem otsivad vestlusrobotit, mis nendega vaidleks.
- 99,2% täpsus on **spetsiifiline sellele ülesandele, mudelile ja hoolikale promptimisele** — see ei ole üldine garantii, et keelemudelid esitavad tõeseid väiteid. Autorid rõhutavad, et sama

isikupärastatud veenmisjõud võiks sama hästi tugevdada *valesid* uskumusi, kui see sellele suunata.
- „Püsiv“ tähendab siin **kahte kuud**, mis on ühe vestluse kohta muljetavaldav, kuid ei võrdu püsiva muutusega kogu eluks.

Kui tugev on tõendus

- **Katsekujundus on tõeline tugevus.** Kontrollitud sekkumis- ja kontrollrühma katsed, puhas platseebolaadne kontroll, kordamine kahes uuringus ja sõltumatus valimis, järelmõõtmised püsivuse hindamiseks ning AI enda väidete selge täpsuskontroll — see on hoolikam kui suur osa veenmisuuringutest ning mõju suund oli kooskõlaline kõikjal, kus seda testiti.
- **Kuid Expression of Concern ei ole joonealune märkus.** Usaldusväärse tulemuse osa on võimalus avaldatud andmetest ja koodist raporteeritud arvud uuesti saada, ning just see läks katki — sõelumiskriteeriumide ebakõlad ja koodi ühendamisveast tekkinud duplikaatread. Autorite väide, et parandatud analüüs säilitab tulemuse, on julgustav ja usutav, kuid praegu on see autorite enda kirjeldus, mitte sõltumatu otsus. *Science'i* hindamine on see, mida tuleb oodata.
- **Aus staatus on „märgistatud“, mitte „ümber lükatud“ ega „kinnitatud“.** Hästi üles ehitatud ja silmatorkav uuring, mille täpsed arvud on ametliku kontrolli all. See on ebamugav koht, kuhu hea lugu jätta, kuid praegu on see õige koht.

Miks see oluline on

Aastaid on olnud mõjukas seisukoht, et vandenõuteooriatesse uskujaid juhivad psühholoogilised vajadused ja identiteet ning seetõttu ei saa fakte kasutades neid uskumustest välja argumenteerida. Kui püsiv tõendipõhine vähenemine kontrollile vastu peab, on see sellele pessimismile oluline vastukaal: probleem võis osaliselt olla selles, et üldine ümberlükkamine ei tegele kunagi *konkreetsete* tõenditega, mida uskuja tegelikult nimetab — midagi, mida LLM suudab teha suurel skaalal.

See on oluline ka juhtumiuuringuna sellest, kuidas teadus peaks toimima. Expression of Concern'i õppetund ei ole, et kuulsad tulemused on väärtusetud; see on, et vead tulevad nähtavale, andmeid uuritakse uuesti ja väiteid kontrollitakse avalikult. Selle mehhanismi töötamine on süsteemi omadus, mitte skandaal — ning just seepärast on õige hoiak selle tulemuse suhtes kannatlikkus, mitte aplaus ega mahakandmine.

Ja AI puhul lõikab see mõlemas suunas. Sama võime kohandatud argumendiga valeuskumust nõrgendada võiks sama tõhusalt mõnda valeuskumust tugevdada. Tehnika on neutraalne; eesmärk ei ole.

Lühidalt

2024. aasta uuring leidis, et kolmevooruline vestlus GPT-4 Turboga vähendas inimeste usku nende endi usutud vandenõuteooriasse umbes 20%; mõju püsis kaks kuud ja üldistus eri tüüpi vandenõudele ning AI faktiväiteid hinnati valdavalt täpseks. 2026. aasta juunis lisati artiklile andmekäitluse ja reprodutseeritavuse probleemide tõttu Editorial Expression of Concern; autorite sõnul säilitab parandatud analüüs tulemuse suuna, statistilise olulisuse ja suuruse ning *Science* hindab seda endiselt. Seda tuleks lugeda kui tõeliselt huvitavat ja hoolikalt üles ehitatud tulemust, mis on praegu kontrolli all — mitte lõplikult kinnitatud ega ümber lükatud. Kõige ausam lause on see, mida protsess ise praegu kirjutab: kontrollige uuesti.

Allikad

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