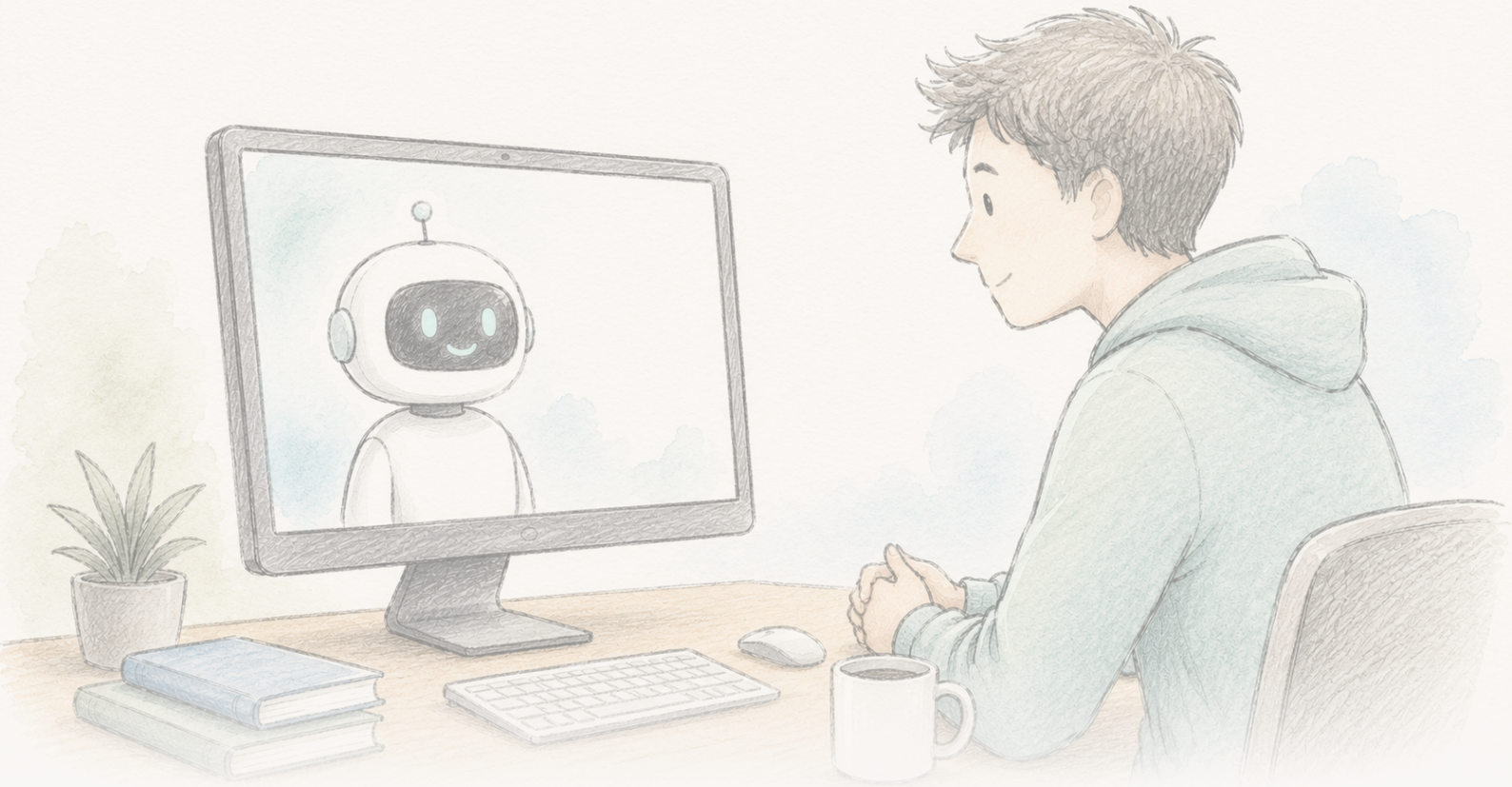


The Clean Paper

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

Pokalbių robotas, regis, keliems mėnesiams susilpnino tikėjimą sąmokslo teorijomis

2024 m. tyrime trijų raundų pokalbis su GPT-4 Turbo žmonių tikėjimą jų pačių pasirinkta sąmokslo teorija sumažino maždaug 20 %; poveikis išliko du mėnesius, pasireiškė įvairioms teorijoms, o DI faktiniai teiginiai buvo įvertinti kaip beveik visiškai tikslūs. 2026 m. birželį prie straipsnio pridėtas redakcinis susirūpinimo pareiškimas dėl duomenų tvarkymo ir atkuriamumo problemų; autoriai teigia, kad pataisyta analizė išlaiko rezultato kryptį, reikšmingumą ir dydį, o „Science“ dar vertina. Tikrai įdomus, kruopščiai sukurtas rezultatas — šiuo metu peržiūrimas, nei galutinai patvirtintas, nei paneigtas.

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“ prie šio straipsnio pridėjo redakcinį susirūpinimo pareiškimą, kol vertina klausimus dėl duomenų tvarkymo ir atkuriamumo. Tai nėra straipsnio atšaukimas ir nėra nusižengimo nustatymas; tai reiškia, kad rezultatą kol kas reikia laikyti preliminariu, kol peržiūra bus išspręsta.

Editorial Expression of Concern →

Vis dėlto faktai veikia — bet prie straipsnio dabar yra įspėjimo žyma

2024 m. tyrimas pateikė rezultatą, kuris prieštarauja patogiai įsitvirtinusiai nuomonei. Jei žmogus trumpai, tris raundus, pasikalba su DI — GPT-4 Turbo, specialiai paragintu atsakyti į konkrečius įrodymus, kuriais tas žmogus grindžia savo tikimą sąmokslo teoriją — vidutiniškai jo tikėjimas ta teorija sumažėja apie 20 %. Sumažėjimas išliko ir po dviejų mėnesių. Tai veikė tiek klasikinėms sąmokslo teorijoms (Johno F. Kennedy nužudymas, ateiviai, iliuminatai), tiek aktualioms (COVID-19, 2020 m. JAV rinkimai), net kai įsitikinimas buvo stiprus ir susijęs su žmogaus tapatybe. Profesionaliam faktų tikrintojui įvertinus 128 DI pateiktus faktinius teiginius, 127 buvo teisingi, vienas — klaidinantis, nė vienas — klaidingas.

Plačiai pasklidusi išvada skambėjo taip: žmones *galima* iškalbėti iš sąmokslo teorijų „triušio urvo“ — jų nepasiekia ne patys faktai, o per daug bendri faktai; reikia tinkamų įrodymų, pateiktų pakankamai konkrečiai. Tai iš tiesų įdomus teiginys, o tyrimas buvo sukurtas kruopščiau nei dauguma įtikinėjimo darbų.

Tačiau 2026 m. birželį *Science* prie straipsnio pridėjo **redakcinį susirūpinimo pareiškimą (Editorial Expression of Concern)**. Būtent šią dalį dauguma pasakojimų greičiausiai praleis, nors šiandien ji lemia, kiek svorio rezultatui galima suteikti.

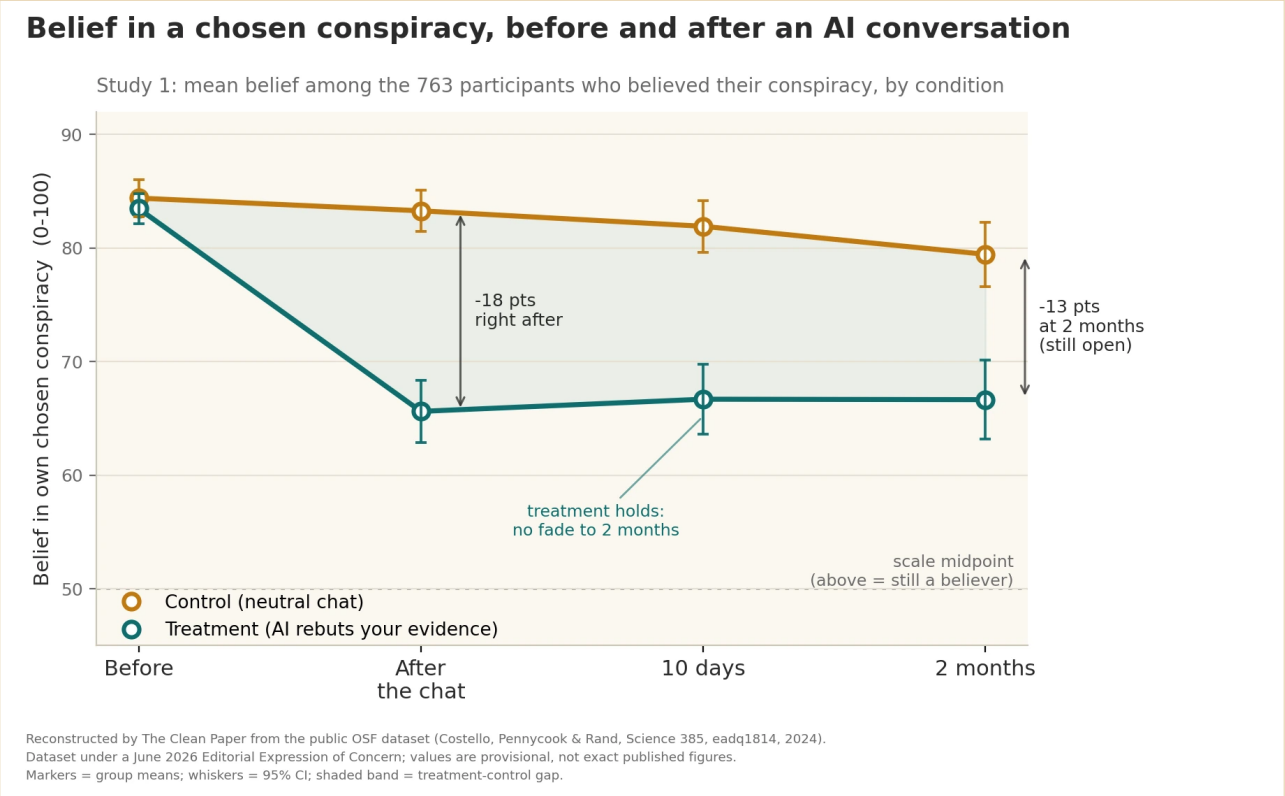
Kas yra „Expression of Concern“ — ir kas tai nėra

Redakcinis susirūpinimo pareiškimas yra oficiali žurnalo žyma, perspėjanti skaitytojus, kad dėl straipsnio iškelta klausimų, kol jie dar tiriami. Tai **nėra** straipsnio atšaukimas — jis nepašalintas — ir savaime nėra nusižengimo nustatymas. Tai reiškia: skaitykite turėdami omenyje šį apribojimą, nes mokslinis įrašas dar gali pasikeisti. Šiuo atveju, sužinoję apie problemas, autoriai jas ištyrė, pateikė detales *Science* ir perdavė pataisytą analizę.

Pažymėtos problemos konkrečios. Autoriams buvo nurodyta, kad dalyvių atrankos kriterijai rankraštyje ir paskelbtame analizės kode taikyti nevienodai, o viešame duomenų rinkinyje dėl kodo sujungimo klaidos atsirado pašalinių eilučių. Dėl to dalį paskelbtų skaičių buvo sunku atkurti iš viešai pateiktos medžiagos. Autoriai perdavė *Science* pataisytą analizės vamzdyną ir atnaujintus rezultatus, kurie, jų teigimu, sutampa su originalais **poveikio kryptimi, statistiniu reikšmingumu ir**

prasmingu dydžiu. *Science* juos vertina. Kol vertinimas nebaigtas, tikslūs skaičiai lieka su klaustuku, kurį gali nuimti tik pats žurnalas.

Taigi čia vienu metu yra dvi istorijos: įdomus rezultatas apie tai, ar faktai gali pajudinti įsitvirtinusius įsitikinimus, ir gyvas pavyzdys, kaip mokslinis įrašas viešai taisomas. Svarbiausia jų nesumaišyti.



Tyrime trijų raundų pokalbis su DI, prieštaravęs konkretiems paties dalyvio nurodytiems įrodymams, vidutiniškai sumažino jo tikėjimą pasirinkta sąmokslo teorija apie 20 %; skirtingai nuo kontrolinio pokalbio kita tema, sumažėjimas tebebuvo matomas po 10 dienų ir po 2 mėnesių. Poveikis buvo ilgalaikis, bet dalinis: dauguma dalyvių ir vėliau tebėjo tikėjo sąmokslu — įsitikinimas susilpnėjo, bet neišnyko. Šiuo metu prie straipsnio pridėtas *Science* redakcinis susirūpinimo pareiškimas (2026 m. birželis) dėl ataskaitų neatitikimų ir klaidos viešame duomenų rinkinyje; autoriai teigia, kad pataisyta analizė išsaugo poveikio kryptį, reikšmingumą ir dydį, o *Science* vertinimą tęsia. Šią diagramą „The Clean Paper“ atkūrė iš to viešo duomenų rinkinio, todėl rodomos reikšmės yra preliminarios, ne galutiniai straipsnio skaičiai.

Original figure — The Clean Paper CC BY 4.0

Ką padarė autoriai

- Atliko du eksperimentus su 2190 dalyvių, kurių kiekvienas savais žodžiais įvardijo sąmokslo teoriją, kuria tiki, ir įrodymus, kurie, jo nuomone, ją pagrindžia.
- Kiekvienas dalyvis turėjo trijų raundų pokalbį su GPT-4 Turbo. **Intervencijos** sąlygoje DI buvo paragintas paneigti konkrečius dalyvio įrodymus; **kontrolinėje** sąlygoje kalbėta nesusijusia tema.
- Tikėjimą pasirinkta sąmokslo teorija matavo prieš pokalbį ir iškart po jo, o tada dalyvius pakar-

totinai apklausė po **10 dienų ir 2 mėnesių**.
- Profesionalus faktų tikrintojas įvertino visų 128 faktinių teiginių, kuriuos DI pateikė atrinktame pavyzdyje, tikslumą.
- Tikrino, ar poveikis persikelia į kitas, nesusijusias sąmokslo teorijas — ir kaip specifiškumo testą, ar intervencija nesumažina tikėjimo sąmokslais, kurie iš tikrųjų yra **teisingi**.

Ką jie nustatė

- **Vidutiniškai apie 20 % sumažėjo** tikėjimas pasirinkta sąmokslo teorija intervencijos grupėje, palyginti su kontrole (1 tyrimas: 95 % pasikliautinasis intervalas [13,8; 19,7] punkto 100 punktų skalėje, P < 0,001).
- **Poveikis išliko.** Intervencijos grupėje sumažėjimas neišnyko: po 10 dienų ir po dviejų mėnesių įsitikinimas liko arti lygio iškart po pokalbio, užuot pamažu grįžęs, o kontrolinėje grupėje visą laiką buvo aukštesnis. Straipsnyje teigiama, kad poveikis pasirinktai sąmokslo teorijai be pastebimo susilpnėjimo išliko bent du mėnesius, o ne buvo tik trumpalaikis svyravimas.
- **Poveikis apėmė įvairias teorijas** ir išliko net tiems dalyviams, kurių pradinis įsitikinimas buvo stiprus.
- **DI faktiniai teiginiai šiame kontekste buvo tikslūs:** 127 iš 128 (99,2 %) įvertinti kaip teisingi, 1 (0,8 %) — kaip klaidinantis, nė vienas — kaip klaidingas.
- **Poveikis buvo specifinis**, o ne bendras nepasitikėjimo skatinimas: intervencija **nesumažino** tikėjimo teisingais sąmokslais. Tai leidžia manyti, kad ji veikė nepagrįstus įsitikinimus, o ne darė žmones ciniškus visko atžvilgiu.
- **Buvo nedidelis persidavimo efektas:** tikėjimas kitomis, nesusijusiomis sąmokslo teorijomis sumažėjo apie 12 %, o dalyviai teigė labiau ketinantys prieštarauti sąmokslo teiginiams. Pagrindinis poveikis pasikartojo 2 tyrime ir ta pačia kryptimi buvo matomas mažesnėje imtyje be kontrolinės grupės (silpnesni, bet suderinami įrodymai).

Ko tai neįrodo

- Šiuo metu rezultatas **nėra be klausimų**. Prie straipsnio pridėtas redakcinis susirūpinimo pareiškimas dėl duomenų tvarkymo ir atkuriamumo problemų, o pataisytus skaičius *Science* dar vertina. Konkrečias reikšmes reikėtų laikyti preliminariomis, kol šis vertinimas bus baigtas.
- Tyrimas **neparodo**, kad DI „išgydo“ tikėjimą sąmokslais. Vidutinis ~20 % sumažėjimas yra prasmingas pokytis, ne išnykimas — dauguma dalyvių po pokalbio vis dar tikėjo teorija, tik silpniau.
- Tai **ne realaus pasaulio diegimo rezultatas**. Dalyviai savanoriškai prisijungė prie tyrimo ir sąžiningai bendravo su DI; realybėje labiausiai įsitikinę sąmokslo šalininkai gali būti mažiausiai linkę kreiptis į pokalbių robotą, kuris jiems prieštaraus.
- 99,2 % tikslumas yra **specifinis šiai užduočiai, modeliui ir kruopščiam promptui** — tai nėra bendras pažadas, kad kalbos modeliai sako tiesą. Autoriai aiškiai pažymi, kad tokia pati individualizuoto įtikinėjimo galia, nukreipta priešinga kryptimi, galėtų stiprinti ir *klaidingus* įsitikinimus.
- „Ilgalaikis“ čia reiškia **du mėnesius**. Vienam pokalbiui tai įspūdinga, bet tai nėra tas pats, kas

nuolatinis poveikis.

Kiek tvirti yra įrodymai

- **Tyrimo dizainas yra tikra stiprybė.** Kontroliuojami intervencijos ir kontrolės eksperimentai, aiški placebo tipo kontrolė, pakartojimas dviejuose tyrimuose ir nepriklausomoje imtyje, ilgalaikio poveikio sekimas ir atskiras DI faktinių teiginių tikslumo patikrinimas — tai kruopščiau nei dauguma įtikinėjimo tyrimų, o poveikio kryptis buvo nuosekli visur, kur tikrinta.
- **Tačiau susirūpinimo pareiškimas nėra smulki išnaša.** Galimybė iš paskelbtų duomenų ir kodo atkurti nurodytus skaičius yra dalis to, kas daro rezultatą patikimą, ir būtent čia atsirado problema — atrankos kriterijų neatitikimai ir dėl kodo sujungimo klaidos pasikartojusios ar pašalinės eilutės. Autorių teiginys, kad pataisytas vamzdynas išlaiko rezultatą, yra padrąsinantis ir tikėtinas, bet kol kas tai pačių autorių ataskaita, o ne nepriklausomas sprendimas. Reikia sulaukti *Science* vertinimo.
- **Sąžiningas statusas yra „pažymėta“, ne „paneigta“ ar „patvirtinta“.** Tai gerai suprojektuotas, ryškus tyrimas, kurio tikslūs skaičiai šiuo metu oficialiai peržiūrimi. Gerai istorijai tai nepatogi būsena, bet būtent ji yra tiksli.

Kodėl tai svarbu

Daugelį metų įtakinga nuomonė teigė, kad sąmokslo teorijų šalininkus pirmiausia veda psichologiniai poreikiai ir tapatybė, todėl vien faktais jų įtikinti nepavyks. Jei ilgalaikis, įrodymais pagrįstas sumažėjimas išliks po patikros, tai bus svarbus kontrargumentas tokiam pesimizmui: galbūt problema iš dalies buvo ta, kad bendri paneigimai niekada neatsakydavo į *konkrečius* įrodymus, kuriuos pats žmogus laiko svarbiais, o LLM gali tai daryti dideliu mastu.

Tyrimas taip pat svarbus kaip pavyzdys, kaip turi veikti mokslas. Susirūpinimo pareiškimo pamoka nėra ta, kad garsūs rezultatai beverčiai. Ji tokia: klaidos iškyla į paviršių, duomenys tikrinami iš naujo, o teiginiai viešai persvarstomi. Tai, kad ši sistema veikia, yra mokslo savybė, ne skandalas — todėl teisinga reakcija į šį rezultatą dabar yra kantrybė, o ne nei plojimai, nei atmetimas.

Ir DI požiūriu tai veikia į abi puses. Tas pats gebėjimas individualizuotu argumentu susilpninti klaidingą įsitikinimą lygiai taip pat galėtų jį sustiprinti. Metodas neutralus; tikslas — ne.

Trumpa santrauka

2024 m. tyrimas parodė, kad trijų raundų pokalbis su GPT-4 Turbo sumažino žmonių tikėjimą jų pačių pasirinkta sąmokslo teorija maždaug 20 %, poveikis išliko du mėnesius, pasireiškė įvairioms sąmokslo teorijoms, o DI faktiniai teiginiai buvo įvertinti kaip beveik visiškai tikslūs. 2026 m. birželį prie

straipsnio pridėtas redakcinis susirūpinimo pareiškimas dėl duomenų tvarkymo ir atkuriamumo problemų; autoriai teigia, kad pataisyta analizė išlaiko rezultato kryptį, reikšmingumą ir dydį, o *Science* dar vertina. Tai iš tiesų įdomus ir kruopščiai sukurtas rezultatas, kuris šiuo metu yra peržiūrimas — nei galutinai patvirtintas, nei paneigtas. Sąžiningiausia išvada tokia, kokią dabar rašo pats procesas: patikrinkime dar kartą.

Šaltiniai

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