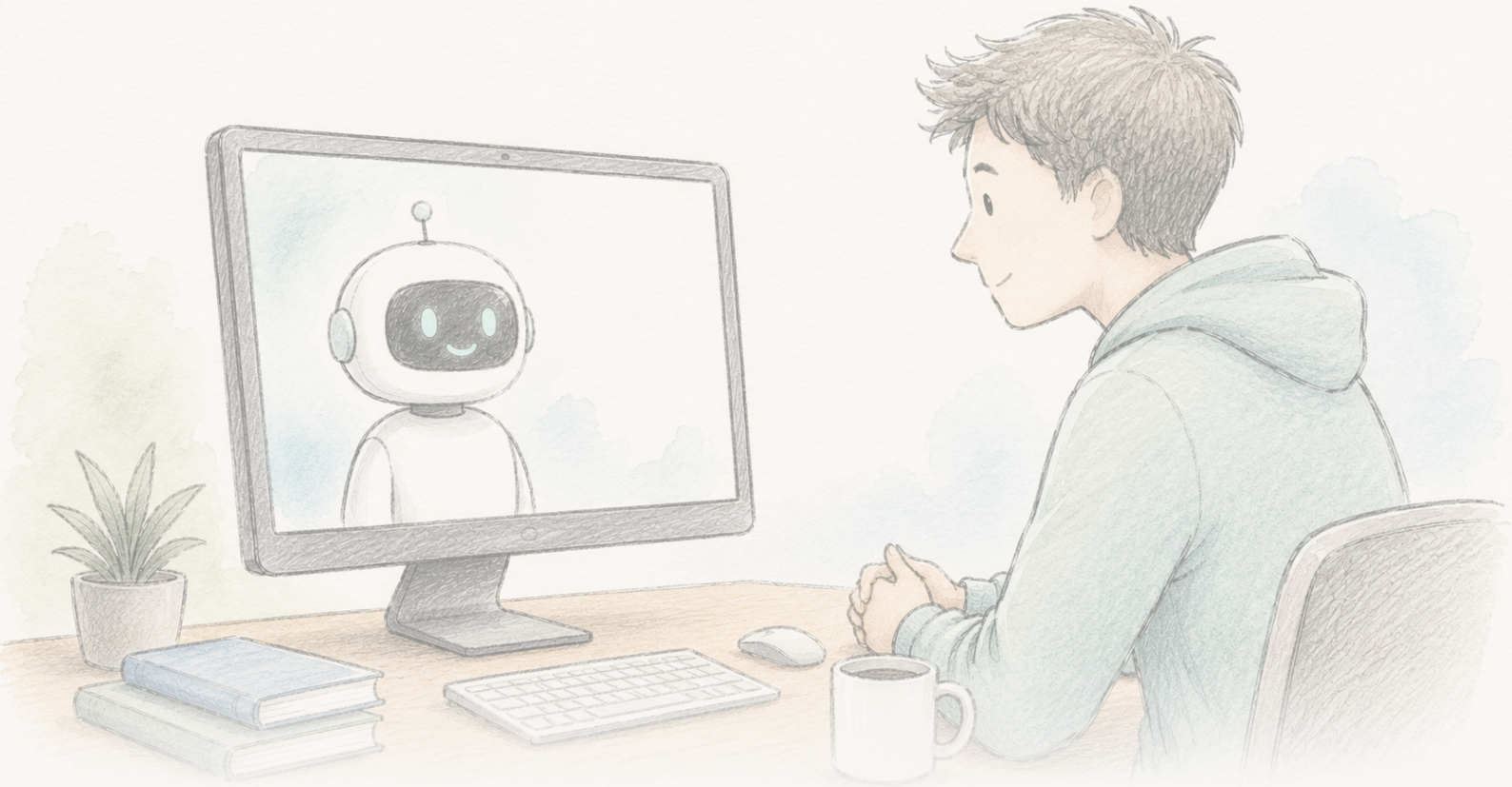


The Clean Paper

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

Një chatbot dukej se uli bindjet konspirative për muaj

Një studim i vitit 2024 gjeti se një bisedë me tri raunde me GPT-4 Turbo uli me rreth 20% bindjen e njerëzve në një teori konspirative që besonin, efekt që zgjati dy muaj, u përgjithësua në lloje të ndryshme konspiracionesh dhe mbështetej në pretendime të AI që u vlerësuan në shumicë dërrmuese të sakta. Në qershor 2026 artikulli mori një Editorial Expression of Concern për probleme me përpunimin e të dhënave dhe riprodhueshmërinë; autorët thonë se analiza e korrigjuar e ruan rezultatin në drejtim, rëndësi dhe madhësi, ndërsa Science ende po e vlerëson. Një rezultat vërtet interesant dhe i ndërtuar me kujdes — aktualisht nën rishikim, as i vendosur dhe as i rrëzuar.

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 i ka bashkëngjitur këtij artikulli një Editorial Expression of Concern ndërsa vlerëson pyetje mbi përpunimin e të dhënave dhe riprodhueshmërinë. Kjo nuk është tërheqje e artikullit dhe as gjetje për shkelje shkencore; do të thotë se rezultati i raportuar duhet lexuar si provizor derisa shqyrtimi të përfundojë.

Editorial Expression of Concern →

Faktet, megjithatë — dhe një paralajmërim mbi punimin

Në vitin 2024, një studim raportoi diçka që shkon kundër një ideje të rehatshme dhe të përhapur. Jepi dikujt një bisedë të shkurtër me tri raunde me një AI — GPT-4 Turbo, të udhëzuar të përgjigjet ndaj provave konkrete që ai person sjell për një teori konspirative që e beson vetë — dhe, mesatarisht, bindja në atë teori bie me rreth 20%. Ulja ishte ende e pranishme dy muaj më vonë. Funksionoi si për konspiracione klasike (vrasja e John F. Kennedy-t, alienët, Illuminati), ashtu edhe për ato të kohës (COVID-19, zgjedhjet amerikane të vitit 2020), madje edhe te njerëz me bindje të fortë dhe të lidhur me identitetin e tyre. Kur një verifikues profesionist faktesh vlerësoi 128 pretendime që AI kishte bërë, 127 dolën të vërteta, një ishte mashtruese dhe asnjë e rreme.

Titulli që u përhap ishte se njerëzit *mund* të nxirren nga "vrima e lepurit" me argumente — se besimtarët e teorive konspirative nuk janë jashtë arritjes së provave, por kanë nevojë për provën e duhur, të adresuar mjaftueshëm konkretisht. Ky është një pretendim vërtet interesant, dhe studimi ishte ndërtuar më me kujdes se shumica e punës mbi bindjen.

Pastaj, në qershor 2026, *Science* vendosi mbi artikullin një **Editorial Expression of Concern**. Kjo është pjesa që shumica e rrëfimeve do ta kalojnë, dhe është pikërisht pjesa që përcakton sa peshë mund të mbajë rezultati sot.

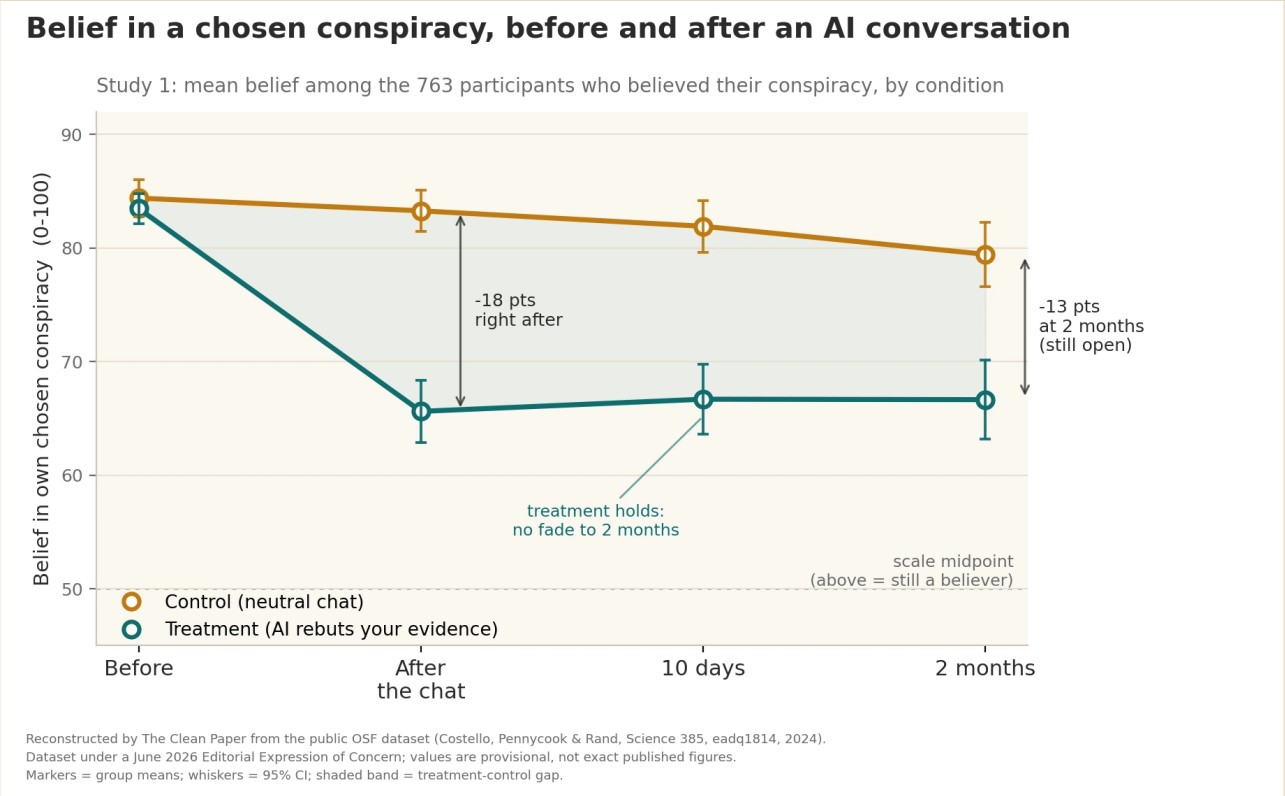
Çfarë është — dhe çfarë nuk është — një Expression of Concern

Një Editorial Expression of Concern është një shenjë formale që një revistë i vendos një punimi për t'i paralajmëruar lexuesit se janë ngritur pyetje, ndërsa ato ende po hetohen. **Nuk** është tërheqje e artikullit (punimi nuk është hequr) dhe nuk është vetvetiu gjetje për shkelje etike. Do të thotë: lexoje duke mbajtur parasysh këtë rezervë, sepse rekordi mund të ndryshojë. Këtu, pasi u vunë në dijeni për problemet, autorët hetuan, i raportuan hollësitë te Science dhe dorëzuan një analizë të korrigjuar.

Problemet e sinjalizuara janë konkrete. Autorët u vunë në dijeni për mospërputhje mes mënyrës si kriteret e filtrimit të pjesëmarrësve përshkruheshin në dorëshkrim dhe si zbatoheshin në kodin e analizës së publikuar, si dhe se dataset-i publik përmbante rreshta të tepërt të futur nga një gabim gjatë bashkimit të kodit — probleme që e bënin të vështirë riprodhimin e disa shifrave të raportuara nga materialet e lëshuara. Autorët i dhanë *Science* një pipeline të korrigjuar dhe rezultate të përditësuara, të cilat sipas tyre ruajnë rezultatin **në drejtim, rëndësi statistikore dhe madhësi thelbësore**.

Science po i vlerëson. Derisa ai vlerësim të mbarojë, shifrat e sakta mbeten provizore.

Pra këtu kemi dy histori njëherësh: një rezultat intrigues mbi pyetjen nëse faktet mund të lëvizin bindje të rrënjosura, dhe një shembull të gjallë të rekordit shkencor që po korrigjohet hapur. Qëllimi është t'i mbajmë të ndara.



Në studim, një bisedë me tri raunde me AI që kundërshtonte provat e deklaruara nga vetë pjesëmarrësi u raportua se uli bindjen në konspiracionin e zgjedhur me rreth 20% mesatarisht; ndryshe nga një bisedë kontrolli mbi temë të palidhur, ulja mbeti e matshme pas 10 ditësh dhe 2 muajsh. Efekti i raportuar ishte i qëndrueshëm, por i pjesshëm: shumica e pjesëmarrësve ende e besonin teorinë më pas — një ulje, jo një kurë. Punimi është aktualisht nën një Editorial Expression of Concern (*Science*, qershor 2026) për mospërputhje në raportim dhe një gabim në dataset-in publik; autorët thonë se analiza e korrigjuar ruan drejtimin, rëndësinë dhe madhësinë e efektit, ndërsa *Science* vazhdon vlerësimin. Ky grafik është rindërtuar nga The Clean Paper mbi dataset-in publik, ndaj vlerat e treguara janë provizore, jo shifrat përfundimtare të punimit.

Original figure — The Clean Paper CC BY 4.0

Çfarë bënë autorët

- Zhvilluan dy eksperimente me 2190 pjesëmarrës, secili prej të cilëve emërtoi — me fjalët e veta — një teori konspirative që besonte dhe provat që mendonte se e mbështesnin.
- Çdo pjesëmarrës bëri një bisedë me tri raunde me GPT-4 Turbo. Në kushtin e **trajtimit**, AI u udhëzua të kundërshtonte provat specifike të pjesëmarrësit; në kushtin e **kontrollit**, diskutoi një temë të palidhur.

- Matën bindjen në konspiracionin e zgjedhur para dhe menjëherë pas bisedës, pastaj i rikontaktuan pjesëmarrësit pas **10 ditësh dhe 2 muajsh**.
- Një verifikues profesionist faktesh vlerësoi saktësinë e të 128 pretendimeve faktike që AI bëri në një kampion.
- Kontrolluan nëse efekti kalonte edhe te konspiracione të tjera të palidhura dhe, si test specifik, nëse ulte edhe besimin në konspiracione që rastësisht janë **të vërteta**.

Çfarë gjetën

- **Rreth 20% ulje mesatare** të bindjes në konspiracionin e zgjedhur në grupin e trajtimit krahasuar me kontrollin (studimi 1: interval besimi 95% [13.8, 19.7] pikë në një shkallë 100-pikëshe, $P < 0.001$).
- **Efekti zgjati.** Në grupin e trajtimit ulja nuk u zhduk: bindja mbeti afër nivelit menjëherë pas bisedës edhe pas 10 ditësh e dy muajsh, ndërsa grupi kontroll mbeti më lart. Punimi e përshkruan efektin mbi konspiracionin e zgjedhur si të pandryshuar për të paktën dy muaj, jo si luhatje të çastit.
- **U përgjithësua** në lloje të ndryshme konspiracionesh dhe u pa edhe te pjesëmarrës me bindje fillestare të fortë.
- **Pretendimet e AI ishin të sakta** në këtë mjedis: 127 nga 128 (99.2%) u vlerësuan të vërteta, 1 (0.8%) mashtruese, asnjë e rreme.
- **Efekti ishte specifik**, jo skepticizëm i përgjithshëm: ndërhyrja **nuk** uli besimin në konspiracione të vërteta, duke sugjeruar se lëvizi bindjet e pambështetura dhe jo besimin në gjithçka.
- **Pati edhe një efekt modest më gjerë** — bindja në konspiracione të tjera të palidhura ra rreth 12%, dhe pjesëmarrësit raportuan më shumë gatishmëri për të kundërshtuar pretendimet konspirative. Efekti kryesor u përsërit në studimin 2 dhe tregoi në të njëjtin drejtim në një kampion më të vogël pa grup kontrolli (provë më e dobët, por e përputhshme).

Çfarë nuk provon

- Aktualisht **nuk qëndron pa pikëpyetje**. Punimi është nën Editorial Expression of Concern për probleme me përpunimin e të dhënave dhe riprodhueshmërinë, dhe shifrat e korrigjuara ende po vlerësohen nga *Science*. Vlerat konkrete duhen trajtuar si provizore derisa ky proces të përfundojë.
- **Nuk** tregon se AI "shëron" bindjet konspirative. Një ulje mesatare ~20% është ndryshim domethënës, jo zhdukje — shumica ende e besonin teorinë më pas, vetëm më pak.
- **Nuk** është rezultat përdorimi në botën reale. Pjesëmarrësit pranuan të hynin në studim dhe u angazhuan me AI në mirëbesim; jashtë laboratorit, personat më të përkushtuar ndaj një teorie mund të jenë pikërisht më pak të prirurit të kërkojnë një chatbot që do t'i kundërshtojë.
- Saktësia 99.2% është **specifike për këtë detyrë, model dhe promptim të kujdesshëm** — jo garanci e përgjithshme se modelet gjuhësore thonë të vërtetën. Autorët theksojnë se e njëjta fuqi bindëse e personalizuar mund të shtyjë edhe bindje *të rreme* nëse përdoret në atë drejtim.
- "I qëndrueshëm" këtu do të thotë **dy muaj**, gjë mbresëlënëse për një bisedë të vetme, por jo e

njëjta gjë me përhershmërinë.

Sa e fortë është prova

- **Dizajni është një pikë e fortë reale.** Eksperimente të kontrolluara trajtim-kundrejt-kontrollit, kontroll i pastër në stil placebo, përsëritje në dy studime dhe një kampion të pavarur, ndjekje në kohë dhe kontroll i veçantë i saktësisë së pretendimeve të AI — më shumë kujdes se shumica e kërkimeve mbi bindjen, dhe drejtimi i efektit është i njëjtë kudo ku u testua.
- **Por Expression of Concern nuk është fusnotë.** Aftësia për të rigjeneruar shifrat e raportuara nga të dhënat dhe kodi i publikuar është pjesë e besueshmërisë së një rezultati, dhe pikërisht kjo u prish — mospërputhje në kriteret e filtrimit dhe rreshta të dyfishuar/të tepërt nga një gabim në bashkimin e kodit. Deklarata e autorëve se pipeline-i i korrigjuar e ruan rezultatin është inkurajuese dhe e mundshme, por mbetet rrëfimi i vetë autorëve, jo verdikt i pavarur. Vlerësimi i *Science* është ai që ka peshën përfundimtare.
- **Statusi i ndershëm është "i sinjalizuar për shqyrtim", jo "i rrëzuar" ose "i konfirmuar".** Një studim i ndërtuar mirë dhe i fortë, shifrat e sakta të të cilit janë nën rishikim formal. Është vend i pakëndshëm për ta lënë një histori të mirë, por është vendi i saktë.

Pse ka rëndësi

Për vite, një pikëpamje me ndikim thoshte se besimtarët e teorive konspirative drejtohen nga nevoja psikologjike dhe identiteti dhe, për këtë arsye, nuk mund të largohen nga bindjet e tyre me fakte. Nëse qëndron pas rishikimit, një ulje e qëndrueshme e bazuar në prova është kundërpeshë me kuptim ndaj këtij pesimizmi: sugjeron se një pjesë e problemit mund të ketë qenë se kundërshtimi i përgjithshëm nuk i përgjigjej kurrë *provës konkrete* që një person citon — diçka që një LLM mund ta bëjë në shkallë të madhe.

Ka rëndësi edhe si shembull i mënyrës si duhet të funksionojë shkenca. Mësimi i Expression of Concern nuk është se rezultatet e famshme janë të pavlera; është se gabimet dalin në dritë, të dhënat rishikohen dhe pretendimet kontrollohen publikisht. Fakti që kjo makineri funksionon është veçori, jo skandal — dhe arsyeja pse qëndrimi i duhur ndaj këtij rezultati është durimi, jo duartrokitja ose hedhja poshtë.

Dhe ka dy tehe për AI. E njëjta aftësi për të ulur një bindje të rreme me argument të përshtatur mund ta rrisë një bindje të rreme po aq efektivisht. Teknika është neutrale; qëllimi jo.

Përmbledhje e shkurtër

Një studim i vitit 2024 gjeti se një bisedë me tri raunde me GPT-4 Turbo uli bindjen e njerëzve në një teori konspirative që ata e besonin me rreth 20%, efekt që vazhdoi për dy muaj dhe u pa në lloje të ndryshme konspiracionesh, ndërsa pretendimet faktike të AI u vlerësuan në shumicë dërrmuese të sakta. Në qershor 2026 punimi u vendos nën Editorial Expression of Concern për probleme në përpunimin e të dhënave dhe riprodhueshmërinë; autorët raportojnë se analiza e korrigjuar e ruan rezultatin në drejtim, rëndësi dhe madhësi, dhe *Science* ende po e vlerëson. Lexoje si një rezultat vërtet interesant dhe të ndërtuar me kujdes që është aktualisht nën shqyrtim — as i vendosur, as i rrëzuar. Fjalia më e ndershme për të është ajo që po shkruan vetë procesi: kontrolloje përsëri.

Burimet

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