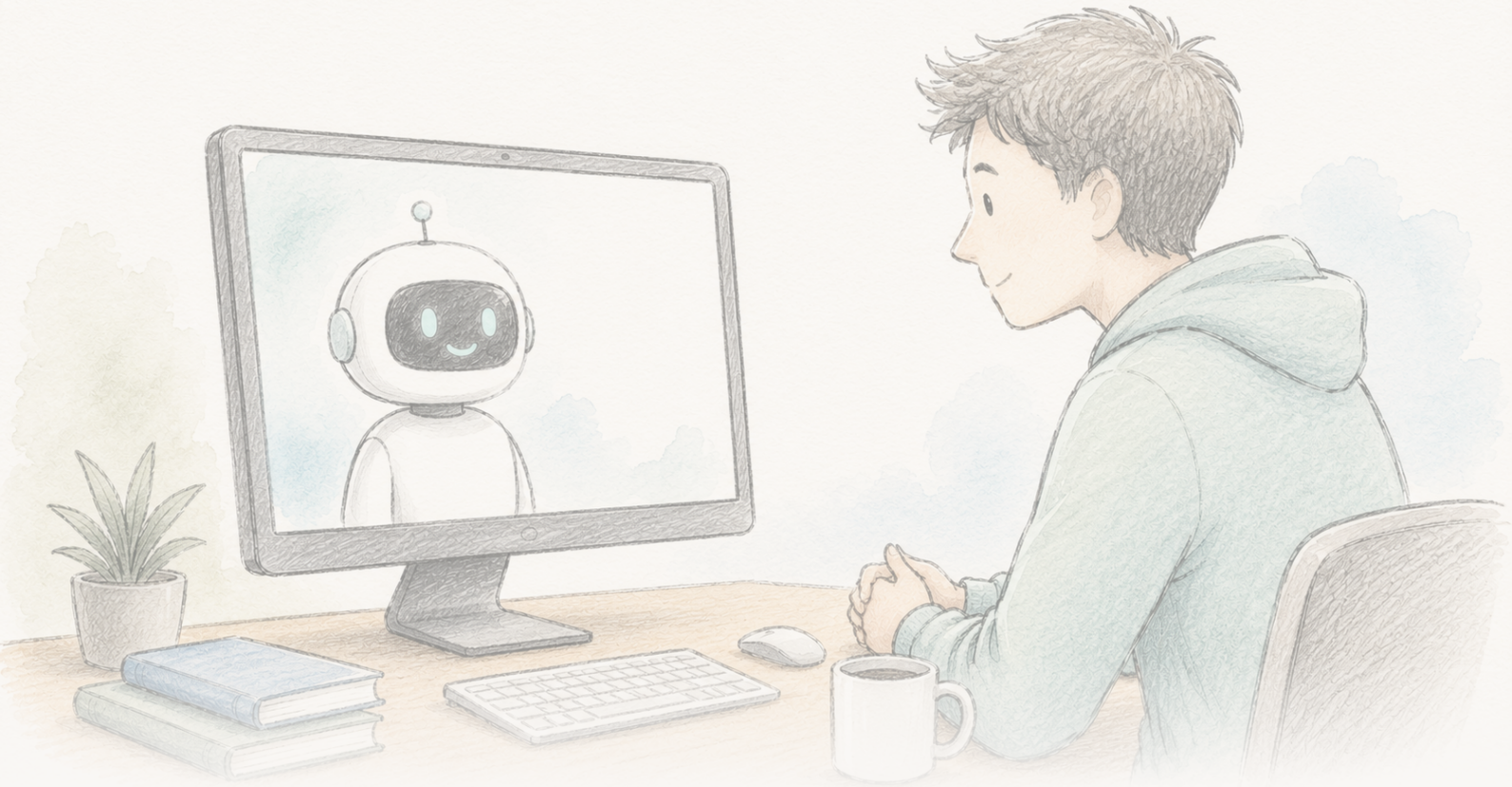


The Clean Paper

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

Chatbot ilionekana kupunguza imani za conspiracy kwa miezi

Utafiti wa 2024 ulipata kwamba mazungumzo ya raundi tatu na GPT-4 Turbo yalipunguza imani ya watu katika conspiracy theory waliyoamini kwa takribani 20%, effect iliyodumu miezi miwili, ikaonekana katika aina tofauti za conspiracies, na kutegemea madai ya AI yaliyohukumiwa kwa kiasi kikubwa kuwa sahihi. Juni 2026 paper iliwekwa chini ya Editorial Expression of Concern kwa matatizo ya data handling na reproducibility; waandishi wanasema corrected analysis inahifadhi direction, significance na size ya finding, na Science bado inatathmini. Result ya kuvutia na iliyojengwa kwa uangalifu — kwa sasa chini ya review, si settled wala debunked.

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 imeambatanisha Editorial Expression of Concern kwenye paper hii huku ikitathmini maswali ya data handling na reproducibility. Hii si retraction wala finding ya misconduct; ina maana matokeo yaliyoripotiwa yanapaswa kusomwa kama provisional mpaka review itatuliwe.

Editorial Expression of Concern →

Ukweli, hata hivyo — na alama ya tahadhari kwenye makala

Mwaka 2024, utafiti uliripoti kitu kinachokwenda kinyume na wazo lililozoeleka. Mpe mtu mazungumzo mafupi ya raundi tatu na AI — GPT-4 Turbo, iliyoelekezwa kujibu ushahidi maalum anaoutaja kwa conspiracy theory anayoamini — na kwa wastani imani yake katika theory hiyo hushuka kwa takribani 20%. Kushuka huko bado kulikuwepo miezi miwili baadaye. Kulionekana kwa conspiracies za zamani (mauaji ya John F. Kennedy, aliens, Illuminati) na za wakati huo (COVID-19, uchaguzi wa Marekani wa 2020), hata kwa watu ambao imani ilikuwa kali na imefungamana na utambulisho wao. Professional fact-checker alipokagua madai 128 ya ukweli yaliyotolewa na AI, 127 yalikuwa sahihi, moja lilikuwa misleading, na hakuna lililokuwa false.

kichwa cha habari iliyosafiri ilikuwa kwamba *inawezekana* kumtoa mtu kwenye rabbit hole kwa mazungumzo — kwamba wanaoamini conspiracy hawako nje ya kufikiwa na ushahidi, wanahitaji tu ushahidi unaofaa ulioletwa kwa specificity ya kutosha. Hilo ni dai la kuvutia kweli, na utafiti ulijengwa kwa uangalifu kuliko kazi nyingi za persuasion.

Kisha, Juni 2026, *Science* iliweka **Editorial Expression of Concern** kwenye makala. Hii ndiyo sehemu ambayo retellings nyingi zitaruka, na ndiyo hasa inayoamua uzito gani matokeo yanaweza kubeba kwa sasa.

Expression of Concern ni nini — na si nini

Editorial Expression of Concern ni alama rasmi ambayo journal huambatanisha na makala kuwaambia wasomaji kwamba maswali yameibuliwa huku uchunguzi ukiendelea. **Si** retraction (makala haijaondolewa), na yenyewe si matokeo ya misconduct. Ina maana: soma makala ukiwa na tahadhari hii, kwa sababu rekodi inaweza kubadilika. Hapa, baada ya kuarifiwa kuhusu matatizo, waandishi walichunguza, wakaripoti specifics kwa Science, na wakatoa corrected uchanganuzi.

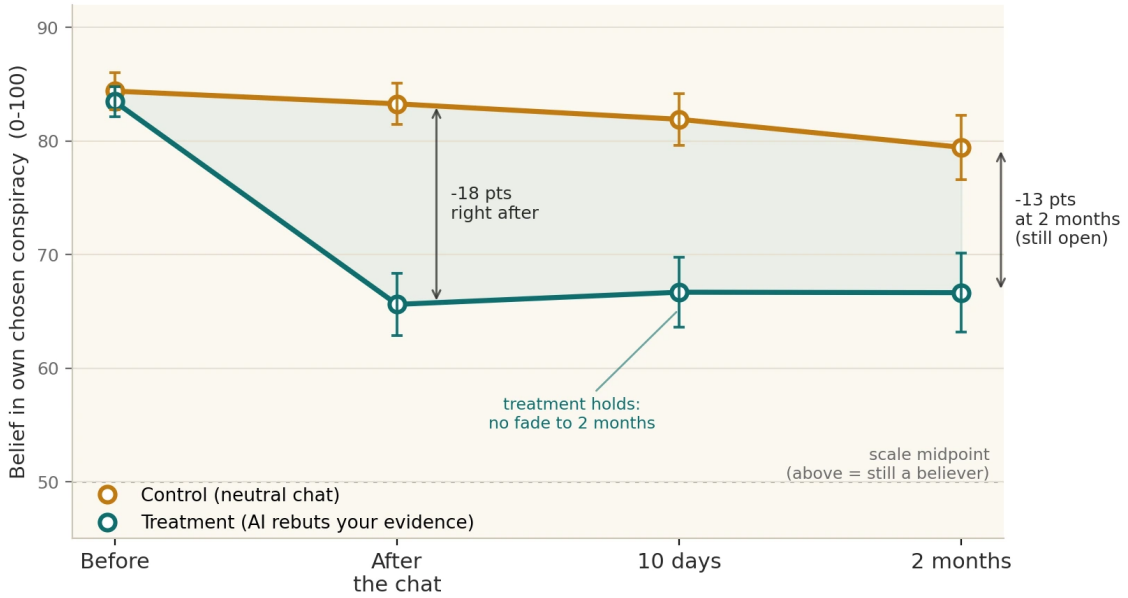
Kilichotiwa alama ni maalum. Waandishi walijulishwa kuhusu inconsistencies katika jinsi participant-screening criteria zilivyotumika kati ya manuscript na published uchanganuzi code, na kwamba hifadhidata ya umma ilikuwa na rows za ziada zilizoingizwa kwa kosa la code merging — matatizo yaliyofanya baadhi ya namba zilizoripotiwa kuwa vigumu kureproduce kutoka materials zilizotolewa. Waandishi waliipa *Science* corrected uchanganuzi pipeline na seti ya updated matokeo, ambazo wanasema zinaendana na original **katika direction, statistical significance na substantive size**.

Science bado inazitathmini. Mpaka tathmini hiyo iishe, exact figures ziko chini ya alama ya swali ambayo journal pekee inaweza kuondoa.

Kwa hiyo hizi ni hadithi mbili kwa wakati mmoja: matokeo ya kuvutia kuhusu kama facts zinaweza kusogeza imani zilizokita mizizi, na mfano unaoendelea sasa wa scientific record kujisahihisha wazi-wazi. Lengo hapa ni kutovichanganya.

Belief in a chosen conspiracy, before and after an AI conversation

Study 1: mean belief among the 763 participants who believed their conspiracy, by condition



Reconstructed by The Clean Paper from the public OSF dataset (Costello, Pennycook & Rand, Science 385, eadq1814, 2024).
Dataset under a June 2026 Editorial Expression of Concern; values are provisional, not exact published figures.
Markers = group means; whiskers = 95% CI; shaded band = treatment-control gap.

Katika utafiti, mazungumzo ya raundi tatu na AI yaliyojibu ushahidi ambao mshiriki mwenyewe aliutaja yaliripotiwa kupunguza imani katika nadharia ya njama iliyochaguliwa kwa takribani 20% kwa wastani. Ikilinganishwa na mazungumzo dhibiti juu ya mada isiyohusiana, tofauti hiyo bado ilionekana baada ya siku 10 na miezi miwili. Athari iliyoripotiwa ilidumu lakini haikuwa kamili: washiriki wengi bado waliamini nadharia hiyo baadaye. Hata hivyo, makala iko chini ya **Editorial Expression of Concern** ya *Science* tangu Juni 2026 kwa sababu ya kutofautiana katika taarifa na kosa katika hifadhidata ya umma. Waandishi wanasema uchanganuzi uliosahihishwa bado unaonyesha mwelekeo, umuhimu wa kitakwimu na ukubwa wa athari sawa, huku *Science* ikiendelea kutathmini suala hilo. Chati hii imejengwa upya na The Clean Paper kutoka hifadhidata hiyo ya umma, kwa hiyo thamani zinazoonyeshwa zinapaswa kuchukuliwa kuwa za muda, si takwimu za mwisho za makala.

Original figure — The Clean Paper CC BY 4.0

Waandishi walifanya nini

- Walifanya experiments mbili zenye washiriki 2,190 ambao kila mmoja alitaja — kwa maneno yake mwenyewe — conspiracy theory anayoamini na ushahidi anaodhani unaiunga mkono.
- Kila mshiriki alikuwa na mazungumzo ya raundi tatu na GPT-4 Turbo. Katika **matibabu** AI iliag-

izwa kupinga ushahidi maalum wa mshiriki; katika **control** ilijadili mada isiyohusiana.
- Walipima imani katika conspiracy iliyochaguliwa kabla na mara baada ya mazungumzo, kisha wakawasiliana tena na washiriki baada ya **siku 10 na miezi 2**.
- Professional fact-checker alitathmini usahihi wa madai yote 128 ya factual yaliyotolewa na AI katika sample.
- Walikagua kama athari ilienea kwa conspiracies nyingine zisizohusiana — na, kama specificity test, kama pia ilipunguza imani katika conspiracies ambazo kwa kweli **ni za kweli**.

Walichokuta

- **Kupungua kwa wastani kwa takribani 20%** katika imani ya nadharia ya njama iliyochaguliwa kwa kundi la matibabu ikilinganishwa na kundi dhibiti (utafiti wa 1: 95% CI [13.8, 19.7] pointi kwenye kipimo cha pointi 100, $P < 0.001$).
- **Kulidumu.** Katika kundi la matibabu kushuka hakukufifia: imani ilibaki karibu na kiwango cha mara baada ya chat siku 10 na miezi miwili baadaye badala ya kupanda tena, huku control ikibaki juu. makala inaeleza athari kwenye conspiracy iliyochaguliwa kama iliyodumu bila kupungua kwa angalau miezi miwili, si mtikisiko wa muda mfupi.
- **Kulionekana kwa aina mbalimbali** za conspiracies ambazo watu walitaja, na hata kwa washiriki walioanza na imani kali.
- **Madai ya AI yalikuwa sahihi** katika setting hii: 127 kati ya 128 (99.2%) yalihukumiwa true, 1 (0.8%) misleading, hakuna false.
- **athari ilikuwa specific**, si kuleta mashaka kwa kila kitu: intervention **haikupunguza** imani katika conspiracies za kweli, ikidokeza kwamba ilihamisha unsupported beliefs badala ya kuwafanya watu kuwa cynical kwa kila dai.
- **Kulikuwa na athari ndogo nje ya nadharia iliyolengwa** — imani katika nadharia nyingine za njama zisizohusiana ilishuka kwa takribani 12%, na washiriki waliripoti nia kubwa zaidi ya kupinga madai ya njama. Athari kuu ilijirudia katika utafiti wa 2 na ikaelekea upande uleule katika sampuli ndogo isiyokuwa na kundi dhibiti; ushahidi huo wa mwisho ni dhaifu zaidi, lakini unaendana na mwelekeo uleule.

Kile ambacho utafiti huu hauthibitishi

- Kwa sasa **haisomiwi bila swali**. makala iko chini ya Editorial Expression of Concern kutokana na matatizo ya data handling na reproducibility, na corrected numbers bado zinatathminiwa na *Science*. Chukulia values maalum kama provisional mpaka mapitio hiyo ikamilike.
- **Haionyeshi** AI "inaponya" imani za conspiracy. Kupungua kwa wastani ~20% ni mabadiliko ya maana, si kufuta imani — washiriki wengi bado waliamini theory baadaye, lakini kwa kiwango kidogo.
- **Si** matokeo ya deployment ya dunia halisi. Washiriki walikubali kuingia utafiti na kuzungumza na AI kwa good faith; nje ya maabara, watu waliojikita zaidi katika conspiracy huenda ndiyo walio na uwezekano mdogo wa kutafuta chatbot itakayobishana nao.

- Usahihi wa 99.2% **ni maalum kwa task hii, modeli hii na prompting makini** — si guarantee ya jumla kwamba lugha modeli husema kweli. Waandishi wanaweka wazi kwamba uwezo uleule wa persuasion ya binafsi unaweza kusukuma *imani za uongo* ukiwekwa katika mwelekeo huo.
- "Durable" hapa ina maana **miezi miwili**, ambayo ni ya kuvutia kwa mazungumzo moja, lakini si sawa na permanent.

Ushahidi una nguvu kiasi gani

- **Muundo wa utafiti ni nguvu halisi.** Kulikuwa na majaribio yaliyodhibitiwa ya matibabu dhidi ya kundi dhibiti, kundi dhibiti linalofanana na placebo, kurudiwa kwa matokeo katika tafiti mbili na sampuli huru, ufuatiliaji wa kudumu kwa athari, na ukaguzi wa usahihi wa madai ya AI. Huo ni muundo makini kuliko tafiti nyingi za ushawishi, na mwelekeo wa athari ulikuwa thabiti kila ulipojaribiwa.
- **Lakini Expression of Concern si footnote.** Uwezo wa kuzalisha upya namba zilizoripotiwa kutoka data na code zilizotolewa ni sehemu ya uaminifu wa matokeo, na hapo ndipo kitu kilivunjika — inconsistencies za screening criteria na duplicated/extraneous rows kutoka code-merging error. Kauli ya waandishi kwamba corrected pipeline inahifadhi matokeo inatia moyo na inawezekana, lakini bado ni account ya waandishi wenyewe, si huru verdict. Evaluation ya *Science* ndiyo inayoamua.
- **Hali ya uaminifu ni "flagged," si "debunked" wala "confirmed."** utafiti iliyojengwa vizuri na yenye matokeo makali, lakini exact numbers zake ziko chini ya formal mapitio. Ni sehemu isiyofurahisha kuacha simulizi nzuri, na ndiyo sahihi.

Kwa nini ni muhimu

Kwa miaka, mtazamo wenye ushawishi ulikuwa kwamba conspiracy believers husukumwa na psychological needs na identity, na kwa hiyo hawawezi kubadilishwa na facts. Reduction inayodumu na inayotegemea ushahidi — kama itasimama — ni counterweight ya maana kwa pessimism hiyo: inapendekeza tatizo lilikuwa kwa sehemu kwamba generic debunking haijawahi kushughulikia *ushahidi maalum* ambao believer anataja, kitu ambacho LLM inaweza kufanya kwa scale.

Pia ni muhimu kama case utafiti wa jinsi science inavyopaswa kufanya kazi. Somo la Expression of Concern si kwamba celebrated matokeo hazina thamani; ni kwamba errors zinaibuliwa, data inachunguzwa upya, na madai zinakaguliwa tena hadharani. Machinery hiyo kufanya kazi ni feature, si scandal — na ndiyo sababu msimamo sahihi kwa matokeo haya ni subira badala ya applause au dismissal.

Na kwenye AI ina pande mbili. Uwezo uleule wa kupunguza false belief kwa argument iliyolengwa unaweza kuiongeza kwa ufanisi sawa. Technique ni neutral; lengo si neutral.

Muhtasari safi

Utafiti wa 2024 ulipata kwamba mazungumzo ya raundi tatu na GPT-4 Turbo yalipunguza imani ya watu katika conspiracy theory waliyokuwa wanaamini kwa takribani 20%, athari iliyodumu miezi miwili na kuonekana katika aina tofauti za conspiracies, huku madai ya factual ya AI yakihukumiwa kwa kiasi kikubwa kuwa sahihi. Juni 2026 makala iliwekwa chini ya Editorial Expression of Concern kwa matatizo ya data handling na reproducibility; waandishi wanaripoti kwamba corrected uchanganuzi inahifadhi matokeo katika direction, significance na size, na *Science* bado inatathmini. Isome kama matokeo ya kuvutia kweli na iliyojengwa kwa uangalifu ambayo kwa sasa iko chini ya mapitio — si settled, si debunked. Sentensi ya uaminifu zaidi ni ile process yenyewe inaandika: ikague tena.

Vyanzo

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