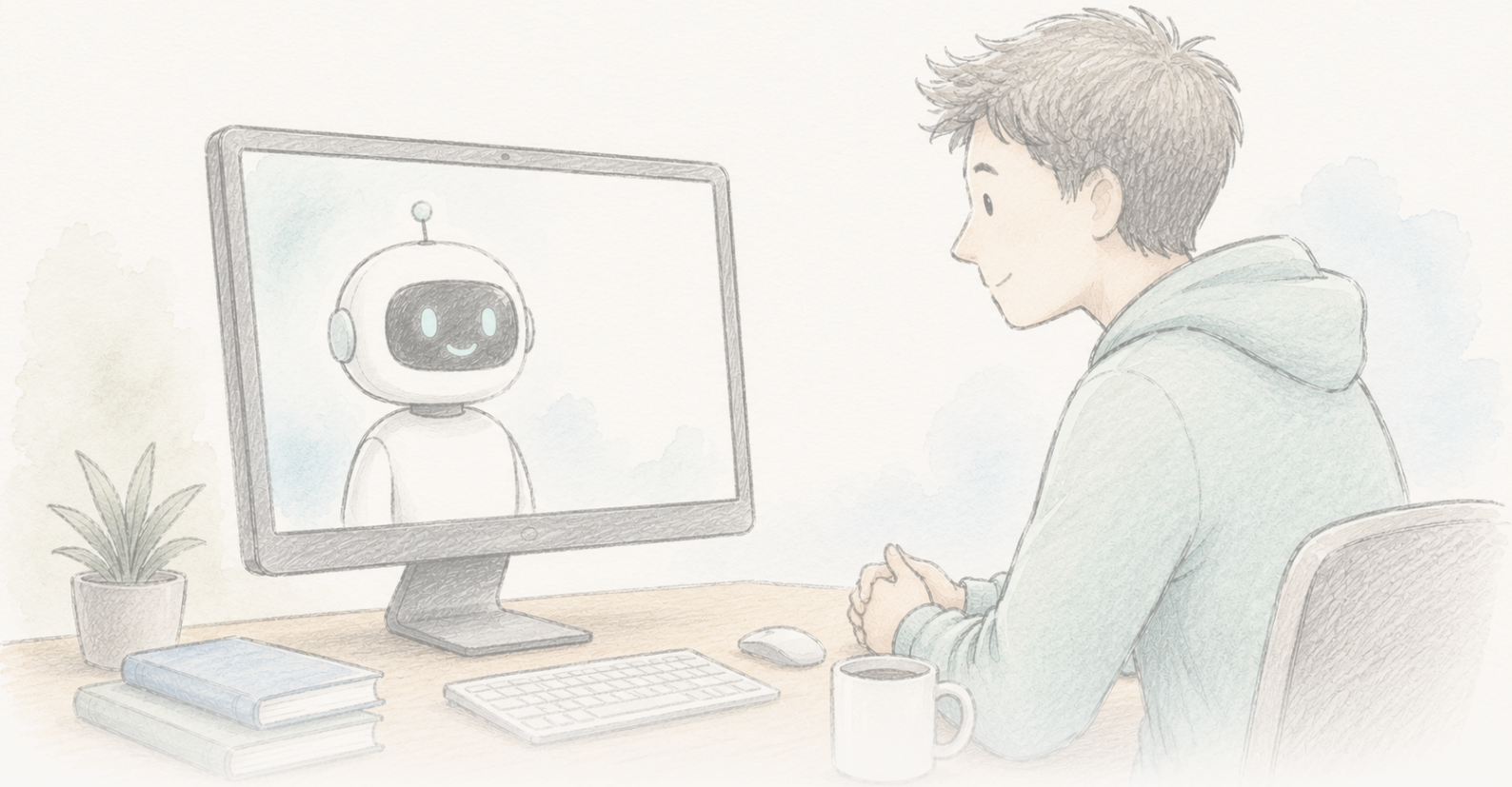


The Clean Paper

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

Wani chatbot ya bayyana ya rage conspiracy beliefs na watanni

Wani bincike na 2024 ya gano cewa tattaunawa rounds uku da GPT-4 Turbo ta rage belief a conspiracy theory da mutum yake riƙe da ita da kusan 20%, effect da ya daɗe watanni biyu kuma ya bayyana a conspiracy types daban-daban, tare da factual claims na AI da aka tantance da accuracy mai yawa. A Yuni 2026 an sanya takardar ƙarƙashin Editorial Expression of Concern saboda data-handling da reproducibility problems; marubutan sun ce corrected analysis ta riƙe direction, significance da size na finding, kuma Science har yanzu tana tantancewa. Result ne mai ban sha'awa kuma carefully built — yanzu under review, ba settled ko debunked ba.

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 ta haɗa Editorial Expression of Concern da wannan takarda yayin da take tantance data-handling da reproducibility questions. Wannan ba retraction ko finding of misconduct ba ne; yana nufin a karanta result a matsayin provisional har review ta warware.

Editorial Expression of Concern →

Gaskiya, bayan komai — amma akwai tutar gargadi a kan takardar

A 2024, wani bincike ya ruwaito wani abu da ya saɓa wa wata ra'ayi da aka saba da shi. Idan ka ba mutum ɗan gajeren tattaunawa na rounds uku da AI — GPT-4 Turbo, an umurce shi ya amsa takamaiman hujjojin da mutumin ya ambata don conspiracy theory da ya yi imani da ita — a matsakaici imanin mutumin a wannan theory yana raguwa da kusan 20%. Raguwar ta ci gaba har bayan watanni biyu. Ta yi aiki a classic conspiracies (kisan John F. Kennedy, aliens, Illuminati) da topical ones (COVID-19, zaɓen Amurka na 2020), har ma ga mutanen da imanin nasu ya yi ƙarfi kuma yana da alaƙa da identity. Lokacin da professional fact-checker ya tantance factual ikirarai 128 da AI ta yi, 127 sun kasance true, ɗaya misleading, babu false.

kanun labarai da ya fi yawo shi ne cewa *za a iya* fitar da mutane daga rabbit hole ta tattaunawa — conspiracy believers ba su wuce ƙarfin shaida ba; abin da suke buƙata shi ne shaida da ya dace da takamaiman hujjar da suke riƙe da ita. Wannan ikirari ne mai ban sha'awa, kuma binciken an tsara shi da kulawa fiye da yawancin persuasion research.

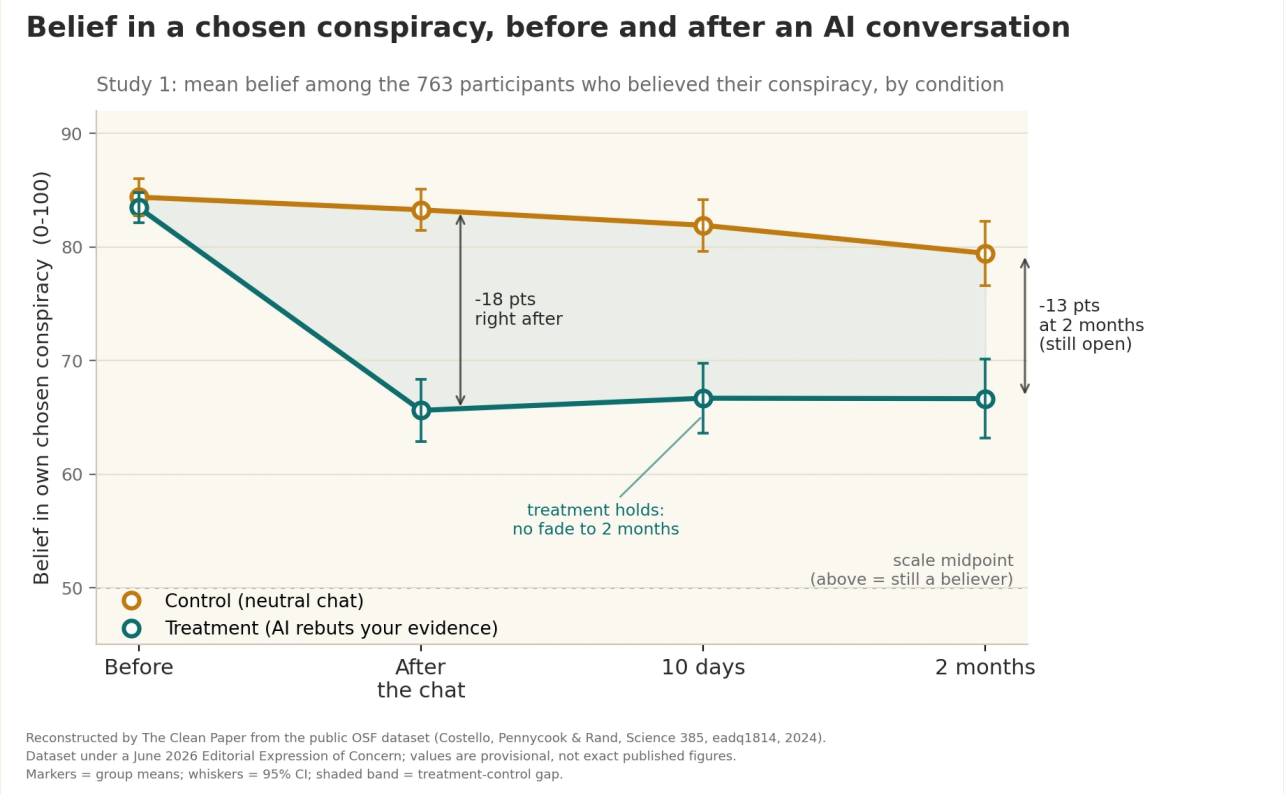
Sai kuma, a Yuni 2026, *Science* ta sanya **Editorial Expression of Concern** a takardar. Wannan shi ne ɓangaren da yawancin retellings za su tsallake, kuma shi ne ainihin abin da ke yanke irin nauyin da sakamakon zai iya ɗauka yanzu.

Menene Expression of Concern — kuma menene ba shi ba

Editorial Expression of Concern formal flag ce da journal ke haɗawa da takarda domin sanar da readers cewa an kawo tambayoyi game da ita yayin da ake bincikensu. **Ba retraction ba ce** — takardar ba a janye ta ba — kuma ba finding of misconduct ba ce da kanta. Ma'anarta ita ce: karanta wannan tare da caveat, domin scientific record na iya canzawa. A nan, bayan an sanar da su matsalolin, marubutan sun bincika, suka kai cikakkun bayanai ga *Science*, suka kuma ba da corrected nazari.

Abin da aka yi flag ɗinsa takamaimai ne. An sanar da marubutan da inconsistencies a yadda participant-screening criteria aka aiwatar tsakanin manuscript da published nazari code, da kuma cewa public dataset yana ɗauke da extra rows da code-merging error ya saka — matsalolin da suka sa wasu lambobin da aka ruwaito suka zama da wuya a reproduce daga released materials. Marubutan sun ba *Science* corrected nazari pipeline da updated sakamako, waɗanda suka ce sun yi daidai da originals **a direction, statistical significance da substantive size**. *Science* har yanzu tana tantance su. Har sai wannan evaluation ta ƙare, exact figures suna ƙarƙashin question mark da journal kaɗai za ta iya cirewa.

Don haka labarai biyu ne a lokaci guda: sakamako mai jan hankali game da ko facts za su iya motsa entrenched beliefs, da live example na scientific record yana gyara kansa a fili. Manufar wannan labari ita ce kada a gauraya su.



A binciken, tattaunawa rounds uku da AI da ta amsa hujjojin da participant kansa ya bayar an ruwaito ta rage imani a chosen conspiracy da kusan 20% a matsakaici; sabanin control chat kan topic da ba shi da alaƙa, raguwar har yanzu ana iya auna ta bayan kwana 10 da watanni 2. Effect ɗin ya daɗe amma bai kawar da imanin gaba ɗaya ba: yawancin participants har yanzu suna da imanin bayan tattaunawar — reduction ce, ba cure ba. Takardar yanzu tana ƙarkashin Editorial Expression of Concern (*Science*, Yuni 2026) saboda reporting inconsistencies da error a public dataset; marubutan sun ce corrected nazari ta riƙe direction, significance da size na effect, yayin da *Science* ke ci gaba da evaluation. The Clean takardar ta sake gina wannan chart daga public dataset, saboda haka values ɗin provisional ne, ba final figures na takardar ba.

Original figure — The Clean Paper CC BY 4.0

Abin da marubutan suka yi

- Sun gudanar da gwaje-gwaje biyu da participants 2190, kowannensu ya faɗi — da kalmominsa — conspiracy theory da ya yi imani da ita da shaida da yake ganin yana goyon bayan ta.
- Kowane participant ya yi rounds uku na tattaunawa da GPT-4 Turbo. A **magani** condition, AI an umurce ta ta ƙaryata takamaiman shaida na participant; a **control** condition, ta tattauna unrelated topic.
- Sun auna belief a chosen conspiracy kafin da nan bayan tattaunawar, sannan suka sake tuntubar participants bayan **kwana 10 da watanni 2**.
- Professional fact-checker ya tantance accuracy na factual ikirarai 128 da AI ta yi a sample.

- Sun duba ko effect ɗin ya bazu zuwa sauran unrelated conspiracies — kuma, a specificity test, ko ya rage belief a conspiracies da **gaskiya** ne.

Abin da suka gano

- **Kusan 20% matsakaici reduction** a belief a chosen conspiracy a magani group idan aka kwatanta da control (bincike 1: 95% confidence interval [13.8, 19.7] points a scale 100, $P < 0.001$).
- **Ya daɓe.** A magani group, raguwar ba ta bace ba: belief ya ci gaba kusa da level ɗin nan take bayan chat a kwana 10 da watanni biyu, yayin da control ya tsaya sama. Takardar ta bayyana effect ɗin a chosen conspiracy a matsayin ya ci gaba ba tare da raguwa ba na ƙalla watanni biyu.
- **Ya generalize** a conspiracy types daban-daban, har ga participants da initial belief ɗinsu ya yi ƙarfi.
- **ikirarai na AI sun yi accurate** a wannan setting: 127 cikin 128 (99.2%) an tantance true, 1 (0.8%) misleading, babu false.
- **Specific ne**, ba blanket doubt ba: intervention ba ta rage belief a true conspiracies ba, abin da ya nuna ta motsa unsupported beliefs maimakon ta sa mutane su yi cynicism ga komai.
- **Ya ɗan spill over** — belief a sauran unrelated conspiracies ya ragu kusan 12%, kuma participants sun ce sun fi son tunkarar conspiracy ikirarai. babban effect ya tsaya a bincike 2, kuma ya nuna direction ɗaya a smaller sample da ba control group (shaida ta fi rauni, amma consistent).

Abin da wannan bai tabbatar ba

- Sakamakon **ba ya tsaye babu tambaya a yanzu**. Takardar tana ƙarƙashin Editorial Expression of Concern saboda bayanai-handling da reproducibility problems, kuma corrected numbers har yanzu *Science* na tantance su. A ɗauki exact values a matsayin provisional har review ta ƙare.
- Ba ya nuna AI "**tana warkar da**" conspiracy belief. ~20% matsakaici reduction meaningful shift ce, ba shafewa gaba ɗaya ba — yawancin participants har yanzu sun yi imani, kawai ƙasa da da.
- Wannan **ba rayuwar yau da kullum deployment sakamako ba ne**. Participants sun yarda su shiga bincike kuma sun yi hulɗa da AI cikin good faith; a duniya ta gaske, mutanen da suka fi manne wa conspiracy theory su ne ma mafi ƙarancin yiwuwar neman chatbot da zai yi musu gardama.
- Accuracy na 99.2% **na wannan task, model da careful prompting ne** — ba general guarantee cewa language models suna faɗin gaskiya ba. Marubutan sun bayyana cewa personalized persuasive power ɗaya na iya tura *false* beliefs idan an nufa ta haka.
- "Durable" a nan na nufin **watanni biyu**, abin da yake ban sha'awa ga tattaunawa ɗaya, amma ba permanent ba.

Yaya ƙarfin shaidar yake

- **Design babban ƙarfi ne.** Controlled magani-versus-control gwaje-gwaje, clean placebo-style control, replication a bincike-bincike biyu da independent sample, durability follow-ups, da explicit accuracy check ga ikirarai na AI — wannan ya fi yawancin persuasion research kulawa, kuma direction na effect consistent ne inda aka

gwada shi.

- **Amma Expression of Concern ba footnote ba ce.** Iya sake samar da reported numbers daga released bayanai da code ɓangare ne na amincin sakamako, kuma a nan ne abin ya karye — screening inconsistencies da duplicated rows daga code-merging error. Maganar marubutan cewa corrected pipeline ta riƙe sakamako encouraging ce, amma har yanzu account ɗinsu ne, ba independent verdict ba. Evaluation na *Science* shi ne abin da ake jira.
- **Matsayin gaskiya shi ne "flagged," ba "debunked" ko "confirmed" ba.** Bincike mai kyau da striking sakamako wanda exact numbers ɗinsa ke ƙarƙashin formal review. Wannan ba matsayi mai daɗi ba ne ga labari mai kyau, amma shi ne daidai.

Me ya sa wannan yake da muhimmanci

Shekaru da yawa, wani influential view ya ce conspiracy believers suna motsawa ne da psychological needs da identity, don haka facts ba za su iya fitar da su daga beliefs ɗinsu ba. Idan durable, shaida-based reduction ɗin nan ya tsaya bayan review, zai zama counterweight mai muhimmanci ga wannan pessimism: yana nuna matsalar watakila wani ɓangare shi ne generic debunking ba ya shiga *takamaiman* shaida da believer ke ambata, abin da LLM zai iya yi a scale.

Haka kuma case bincike ne na yadda science ya kamata ta yi aiki. Darasin Expression of Concern ba cewa celebrated sakamako ba su da amfani ba ne; shi ne errors suna fitowa, bayanai ana sake dubawa, ikirarai kuma ana sake checking a fili. Wannan machinery yana aiki feature ne, ba scandal ba — shi ya sa posture da ya dace ga sakamakon patience ne, ba applause ko dismissal nan take ba.

Kuma AI ɗin tana da bangarori biyu. Capacity ɗin da zai rage false belief da tailored argument na iya ƙara wani false belief da irin wannan ƙarfi. Technique ɗin neutral ce; aim ɗin ba neutral ba ne.

Taƙaitaccen bayani

Wani bincike na 2024 ya gano cewa rounds uku na tattaunawa da GPT-4 Turbo sun rage belief na mutane a conspiracy theory da suke riƙe da ita da kusan 20%, effect da ya ci gaba watanni biyu kuma ya generalize a conspiracy types, yayin da factual ikirarai na AI aka tantance da accuracy mai yawa. A Yuni 2026, takardar ta shiga Editorial Expression of Concern saboda bayanai-handling da reproducibility problems; marubutan sun ce corrected nazari ta riƙe direction, significance da size na finding, kuma *Science* har yanzu tana tantancewa. A karanta shi a matsayin interesting, carefully built sakamako da yanzu ke under review — ba settled ba, ba debunked ba. Jumlar da ta fi gaskiya ita ce abin da scientific process ɗin kanta ke rubutawa: a sake dubawa.

Majiyoyi

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