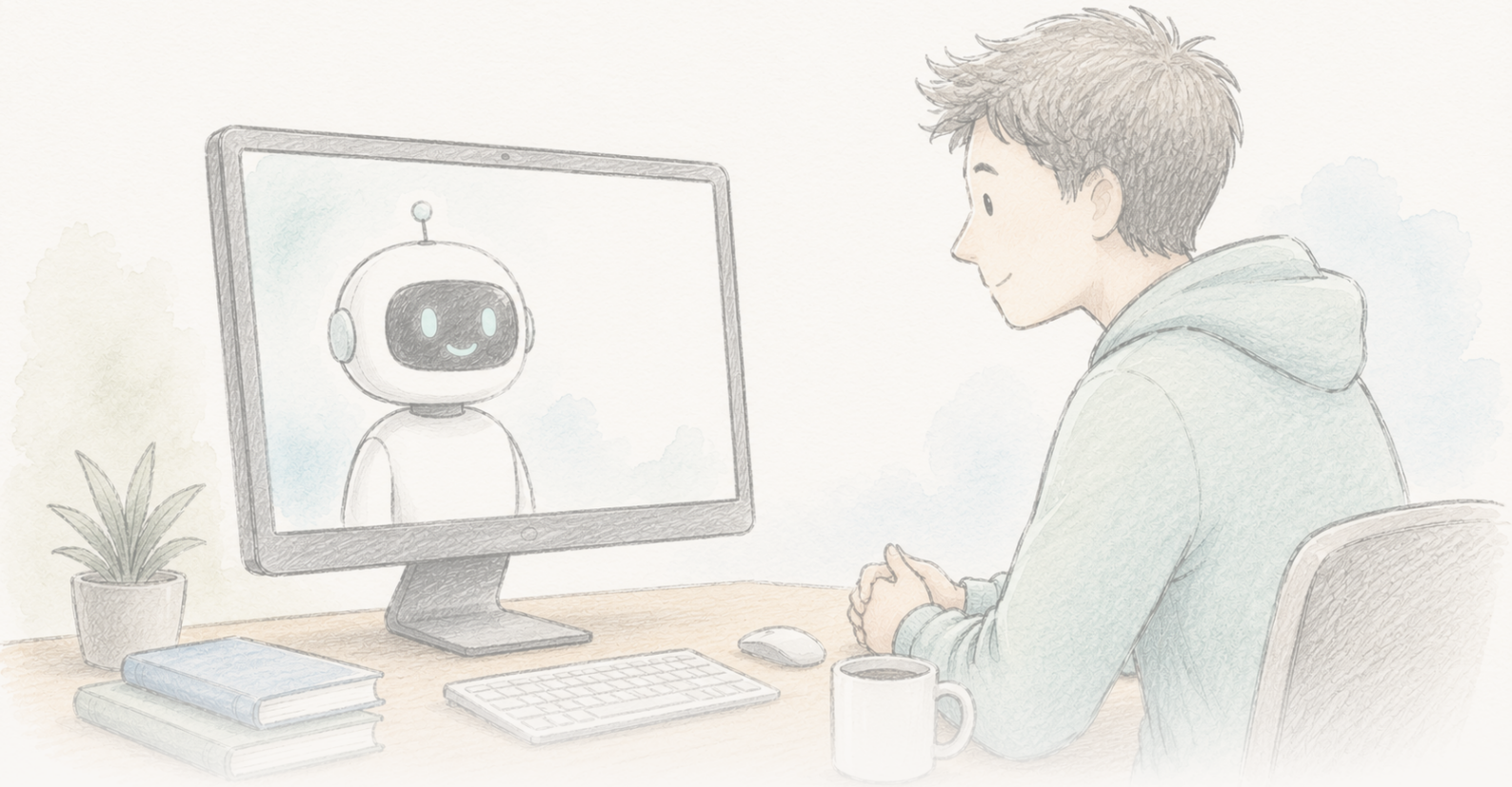


The Clean Paper

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

Chatbot kan dà bí ẹni pé ó dín conspiracy beliefs kù fún ọ̀ṣù

Study 2024 kan rí pé three-round conversation pẹ̀lú GPT-4 Turbo dín belief nínú conspiracy theory tí participants gbà kù ní ayíká 20%, effect tó pẹ́ ọ̀ṣù méjì, tí ó generalize kọjá conspiracy types, pẹ̀lú AI claims tí a rated highly accurate. Ní June 2026, paper náà gba Editorial Expression of Concern fún data-handling àti reproducibility problems; authors ọ̀ sọ corrected analysis pa direction, significance àti size mọ́, Science sì ṣì ń evaluate. Result tó fanimọ́ra, tí a kọ́ pẹ̀lú ìṣọ́ra — ṣùgbọ́n tí review ṣì ń lọ.

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 ti fi Editorial Expression of Concern sí paper yìí nígbà tí ó ń evaluate data-handling àti reproducibility questions. Kì í ṣe retraction, kò sì jẹ́ finding of misconduct; ó túmọ̀ sí pé reported result yẹ ká ka gẹ́gẹ́ bí provisional títí review yóò fi parí.

Editorial Expression of Concern →

Nígbà míì, facts lè ṣiṣẹ́ — ṣùgbọ́n ìwé ìwádìí náà ní flag kan

Ní 2024, ìwádìí kan ròyìn ohun tó tako conventional wisdom kan tí ó rọrùn láti gbà. Fún ènìyàn ní short, three-round conversation pẹ̀lú AI — GPT-4 Turbo, tí a prompt láti dáhùn ẹ̀rí pàtó tí ẹni náà ń lò fún conspiracy àbá-ẹ̀kọ́ tí ó gbàgbọ́ — ní apapọ̀, ìgbàgbọ́ ẹni náà nínú àbá-ẹ̀kọ́ náà dín kù ní ayíká 20%. Ìdínkù náà ṣì hàn oṣù méjì lẹ́yìn náà. Ó ṣiṣẹ́ kọjá classic conspiracies (assassination John F. Kennedy, aliens, Illuminati) àti topical ones (COVID-19, 2020 US election), paápàá lára àwọn ènìyàn tí ìgbàgbọ́ wọn lágbára tí ó sì ní ìbáṣepọ̀ pẹ̀lú identity. Nígbà tí professional fact-checker kan grade claims 128 tí AI ṣe, 127 jẹ́ true, ọkan misleading, kò sí false.

àkọlé tó tàn ká ni pé o *lè* bá ènìyàn sọ̀rọ̀ kúrò nínú rabbit hole — pé conspiracy believers kò kọjá ibi tí ẹ̀rí lè dé, wọn kan nílò ẹ̀rí tó tọ́, tí a fi hàn ní ọ̀nà tó specific. Claim yìí fanimọ́ra gidi, àpẹẹrẹ ìwádìí náà sì ṣọ́ra ju ọ̀pọ̀ persuasion research lọ.

Lẹ́yìn náà, ní June 2026, *Science* fi **Editorial Expression of Concern** sí ìwé ìwádìí náà. Èyí ni apá tí ọ̀pọ̀ retellings lè fo kọjá, ṣùgbọ́n apá yìí gan-an ló pinnu iye weight tí èsì lè ru ní báyìí.

Kí ni Expression of Concern — àti ohun tí kò jẹ́

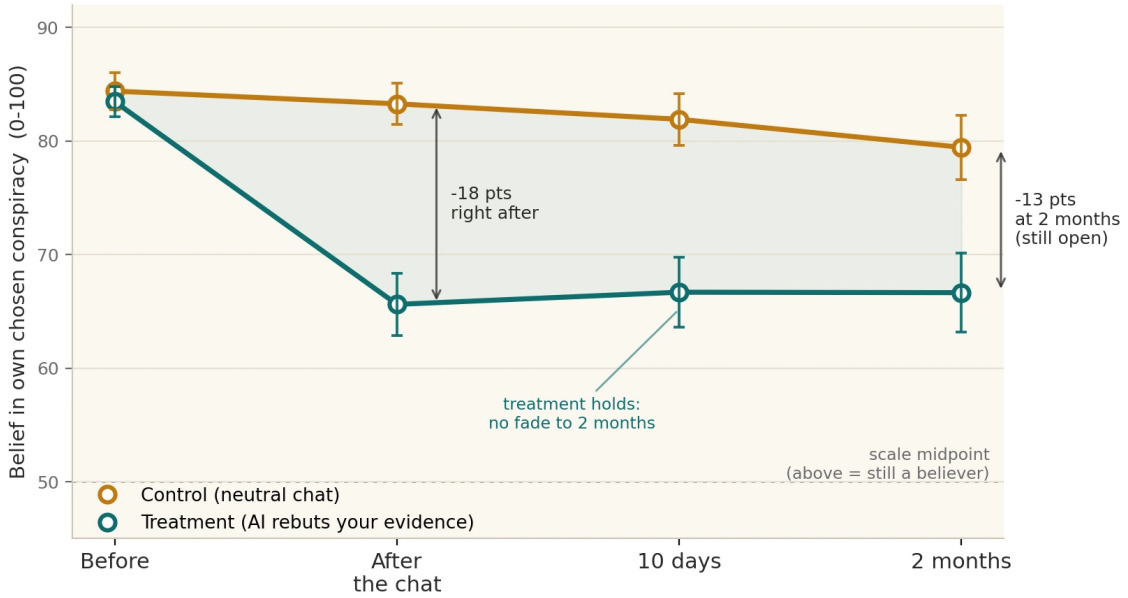
Editorial Expression of Concern jẹ́ formal flag tí journal máa ń fi sí ìwé ìwádìí láti kìlọ̀ fún readers pé ìbéèrè ti dide, nígbà tí a ṣì ń ṣiṣẹ́ lórí wọn. **Kì í ṣe retraction** — a kò yọ ìwé ìwádìí kúrò — kò sì túmọ̀ sí àwárí of misconduct fúnra rẹ̀. Ó túmọ̀ sí: ka ìwé ìwádìí pẹ̀lú caveat yìí, nítorí scientific àkọsílẹ̀ lè yí. Níhìn-ín, lẹ́yìn tí ònkọ̀wé mọ issues náà, wọ́n ṣe investigation, wọ́n ròyìn specifics sí *Science*, wọ́n sì fi corrected ìtúpalẹ̀ ránṣẹ́.

Ohun tí a flag jẹ́ specific. A ṣọ fún ònkọ̀wé pé inconsistencies wà nínú bí participant-screening criteria ṣe apply láàárín manuscript àti published ìtúpalẹ̀ code, àti pé public dataset ní extraneous rows tí code-merging àṣìṣe fi splice wọ inú rẹ̀ — ìṣòro tó mú kí díẹ̀ ninu reported numbers nira láti reproduce láti released ohun èlò. ònkọ̀wé fún *Science* ní corrected ìtúpalẹ̀ pipeline àti updated èsì, tí wọ́n ṣọ pé ó bá original mu **ní ìtọ́ṣọ́nà, statistical significance àti substantive ìwọ̀n**. *Science* ń evaluate wọn. Títí evaluation yẹn yóò parí, exact figures ṣì wà lábẹ́ ìbéèrè mark tí journal nìkan lè yọ.

Nítorí náà, itan méjì ni èyí ní àkókò kan: intriguing èsì nípa bóyá facts lè yí entrenched beliefs, àti live example ti scientific àkọsílẹ̀ tó ń ṣe correction ara rẹ̀ ní gbangba. Kókó article yìí ni láti má ṣe da àwọn itan méjì náà pọ̀.

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.

Nínú ìwádìí náà, three-round AI conversation tó rebut ẹ̀rí tí olukopa funra rẹ̀ sọ ni a ròyìn pé ó dín belief nínú conspiracy tí ẹni náà yan kù ní ayíká 20% ní apapọ̀; unlike control chat lórí unrelated topic, drop náà ṣì measurable ní 10 days àti 2 months. Effect náà durable ṣùgbọ́n partial: ọ̀pọ̀ olukopa ṣì gbà conspiracy náà lẹ́yìn náà — reduction, kì í ṣe cure. ìwé ìwádìí náà wà lábẹ́ Editorial Expression of Concern (*Science*, June 2026) lórí reporting inconsistencies àti àṣìṣe nínú public dataset; òǹkọ̀wé sọ corrected ìtúpalẹ̀ pa ìtọ́ṣọ́nà, significance àti ìwọ̀n mọ́, nígbà tí *Science* ṣì ń evaluate. The Clean ìwé ìwádìí tún chart yìí kọ́ láti public dataset náà, nítorí náà values yìí provisional ni, kì í ṣe final figures ìwé ìwádìí.

Ohun tí àwọn olùkọ̀wé ṣe

- Wọ́n ṣe ìdánwò méjì pẹ̀lú olukopa 2190, tí ọkọọkan sọ — ní ọ̀rọ̀ tirẹ̀ — conspiracy àbá-ẹ̀kọ́ tí ó gbàgbọ́ àti ẹ̀rí tí ó rò pé ó ṣe atilẹyin rẹ̀.
- Participant kọọkan ní three-round conversation pẹ̀lú GPT-4 Turbo. Nínú **ìtọ́ju** condition, AI ni prompt láti rebut participant's specific ẹ̀rí; nínú **control** condition, ó jíròrò unrelated topic.
- Wọ́n wọn belief nínú chosen conspiracy ṣáájú àti lẹ́ṣẹ̀kẹṣẹ̀ lẹ́yìn conversation, wọ́n sì tún contact olukopa ní **10 days àti 2 months**.
- Professional fact-checker kan evaluate ìpéye gbogbo 128 factual claims tí AI ṣe nínú àpẹẹrẹ kan.
- Wọ́n ṣàyẹ̀wò bóyá ipa náà spill over sí other unrelated conspiracies — àti, gẹ́gẹ́ bí specificity ìdánwò, bóyá ó tún dín belief nínú conspiracies tí ó jẹ́ **true** kù.

Ohun tí wọ́n rí

- **Ayíká 20% average reduction** nínú belief nínú chosen conspiracy fún ìtọju ẹgbẹ́ relative sí control (ìwádìí 1: 95% ìgbẹ́kẹ̀lé interval [13.8, 19.7] kókó lórí 100-kókó scale, $P < 0.001$).
- **Ó pẹ́.** Nínú ìtọju ẹgbẹ́, drop náà kò fade: belief dúró nítòsí just-after-chat ìpele ní 10 days àti 2 months dípò kó padà sókè, control sì dúró higher. ìwé ìwádìí náà sọ ipa lórí chosen conspiracy tẹ̀síwájú láì dín kù fún o kere tán oṣù méjì, kì í ṣe momentary wobble.
- **Ó generalize** kọjá ààlà ti conspiracies tí people named, ó sì ṣiṣẹ́ paápàá fún olukopa tí initial belief wọn lágbára.
- **AI claims jẹ́ accurate** nínú setting yìí: 127 nínú 128 (99.2%) ni rated true, 1 (0.8%) misleading, kò sí false.
- **Ó specific**, kì í ṣe blanket doubt: intervention kò dín belief nínú true conspiracies kù, tó ń daba pé ó yí unsupported beliefs padà dípò kó mú people cynical nípa ohun gbogbo.
- **Spillover díẹ̀ wà** — belief nínú other unrelated conspiracies dín kù ní ayíká 12%, olukopa sì ròyìn greater intention láti push back lórí conspiracy claims. Main ipa dúró nínú ìwádìí 2, ó sì tọ́ sí ìtọ́ṣọ́nà kan náà nínú smaller àpẹẹrẹ tí kò ní control ẹgbẹ́ (weaker ẹ̀rí, ṣùgbọ́n consistent).

Ohun tí èyí kò fi ẹ̀rí múlẹ̀

- Ní báyìí ìwé ìwádìí náà kò dúró **láìsí ìbéèrè**. Ó wà lábẹ́ Editorial Expression of Concern fún data-handling àti reproducibility ìṣòro, corrected numbers sì ṣì ń jẹ́ evaluated nipasẹ *Science*. Ka exact values bí provisional tití review yẹn yóò parí.
- Kò fi hàn pé AI "**cures**" conspiracy belief. ~20% average reduction jẹ́ meaningful shift, kì í ṣe erasure — ọ̀pọ̀ olukopa ṣì gbà àbá-ẹ̀kọ́ náà lẹ́yìn, wọ́n kan gbà á kéré síi.
- Kì í ṣe **ayé gidi deployment èsì**. Participants yan láti wọ ìwádìí, wọ́n sì engage pẹ̀lú AI ní good faith; ní gidi world, àwọn tó committed jù sí conspiracy lè jẹ́ àwọn tí kò fẹ́ bá chatbot kan sọ̀rọ̀ tí yóò tako wọn.
- 99.2% ìpéye jẹ́ **specific sí iṣẹ́-ṣíṣe yìí, àwòṣe yìí àti careful prompting** — kì í ṣe gbogbogbò guarantee pé èdè àwòṣe máa ń sọ truth. òǹkọ̀wé sọ kedere pé kan náà personalized persuasive power lè push *false* beliefs tí a bá lo o sí i.
- "Durable" níhìn-ín túmọ̀ sí **oṣù méjì**, ohun tó impressive fún conversation kan, ṣùgbọ́n kì í ṣe permanent.

Báwo ni ẹ̀rí ṣe lágbára tó?

- **Design jẹ́ gidi strength.** Controlled treatment-versus-control ìdánwò, clean placebo-style control, replication kọjá ìwádìí méjì àti independent àpẹẹrẹ, durability follow-ups, explicit ìpéye check lórí AI claims — èyí ṣọ́ra ju ọ̀pọ̀ persuasion research lọ, ìtọ́ṣọ́nà ti ipa sì consistent ní gbogbo ibi tí a ìdánwò.
- **Ṣùgbọ́n Expression of Concern kì í ṣe footnote.** Ability láti regenerate reported numbers láti released dátà àti code jẹ́ apá trustworthiness, ìyẹn gan-an ni ohun tí ó fọ — screening-criteria inconsistencies àti duplicated rows láti code-merging àṣìṣe. Statement òǹkọ̀wé pé corrected pipeline pa èsì mọ́ jẹ́ encouraging, ó sì ṣeé gbà, ṣùgbọ́n ó ṣì jẹ́ account òǹkọ̀wé, kì í ṣe independent verdict. Evaluation *Science* ni ohun tó yẹ ká dúró de.
- **Status tó tọ́ ni "flagged," kì í ṣe "debunked" tàbí "confirmed."** ìwádìí tó kọ́ dáadáa, striking, ṣùgbọ́n exact numbers rẹ̀ wà lábẹ́ formal review. Ó lè jẹ́ uncomfortable láti fi ìtàn rere sí ipo bẹ́ẹ̀, ṣùgbọ́n ipo tó pe ni.

Kí nìdí tí ó fi ṣe pàtàkì?

Fún ọdún púpọ̀, view tó influential ni pé conspiracy believers ni psychological needs àti identity ń darí, nítorí náà facts kò lè yí belief wọn. Durable, evidence-based reduction — tí ó bá survive review — jẹ́ counterweight tó meaningful sí pessimism yẹn: ó daba pé apá ìṣòro ni pé generic debunking kò ń bá *specific* ẹ̀rí tí believer gan-an ń cite ṣiṣẹ́, ohun tí LLM lè ṣe at scale.

Ó tún ṣe pàtàkì gẹ́gẹ́ bí ọ̀ràn ìwádìí ní bí science ṣe kí ó ṣiṣẹ́. Ẹ̀kọ́ Expression of Concern kì í ṣe pé celebrated èsì kò ní iye; ó jẹ́ pé àṣìṣe ń hàn, dátà ń jẹ́ re-examined, claims sì ń jẹ́ re-checked ní gbangba. Machinery yìí tó ń ṣiṣẹ́ jẹ́ àbùdá, kì í ṣe scandal — ìdí ni pé posture tó tọ́ sí èsì yìí ni patience, kì í ṣe applause lẹ́ṣẹ̀kẹṣẹ̀ tàbí dismissal.

AI aspect náà ní edge méjì. Capacity kan náà láti dín false belief kù pẹ̀lú tailored argument lè pọ̀ false belief síi pẹ̀lú effectiveness kan náà. Technique jẹ́ neutral; aim kì í ṣe neutral.

Àkótán kedere

ìwádìí 2024 kan rí pé three-round conversation pẹ̀lú GPT-4 Turbo dín belief àwọn ènìyàn nínú conspiracy àbá-ẹ̀kọ́ tí wọ́n gbà kù ní ayíká 20%, ipa tó tẹ̀síwájú fún oṣù méjì, tó sì generalize kọjá conspiracy types, pẹ̀lú AI factual claims tí a rated overwhelmingly accurate. Ní June 2026, ìwé ìwádìí náà wà lábẹ́ Editorial Expression of Concern fún data-handling àti reproducibility ìṣòro; òǹkọ̀wé ròyìn pé corrected ìtúpalẹ̀ pa ìtọ́ṣọ́nà, significance àti ìwọ̀n ti àwárí mọ́, *Science* sì ṣì ń evaluate. Ka a gẹ́gẹ́ bí genuinely interesting, carefully built èsì tí a ṣì ń review — kì í ṣe settled, kì í ṣe debunked. Gbolohun tó honest jù nípa rẹ̀ ni ohun tí ìlànà náà funra rẹ̀ ń kọ: ṣàyẹ̀wò lẹ́ẹ̀kan sí i.

Àwọn orísun

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`