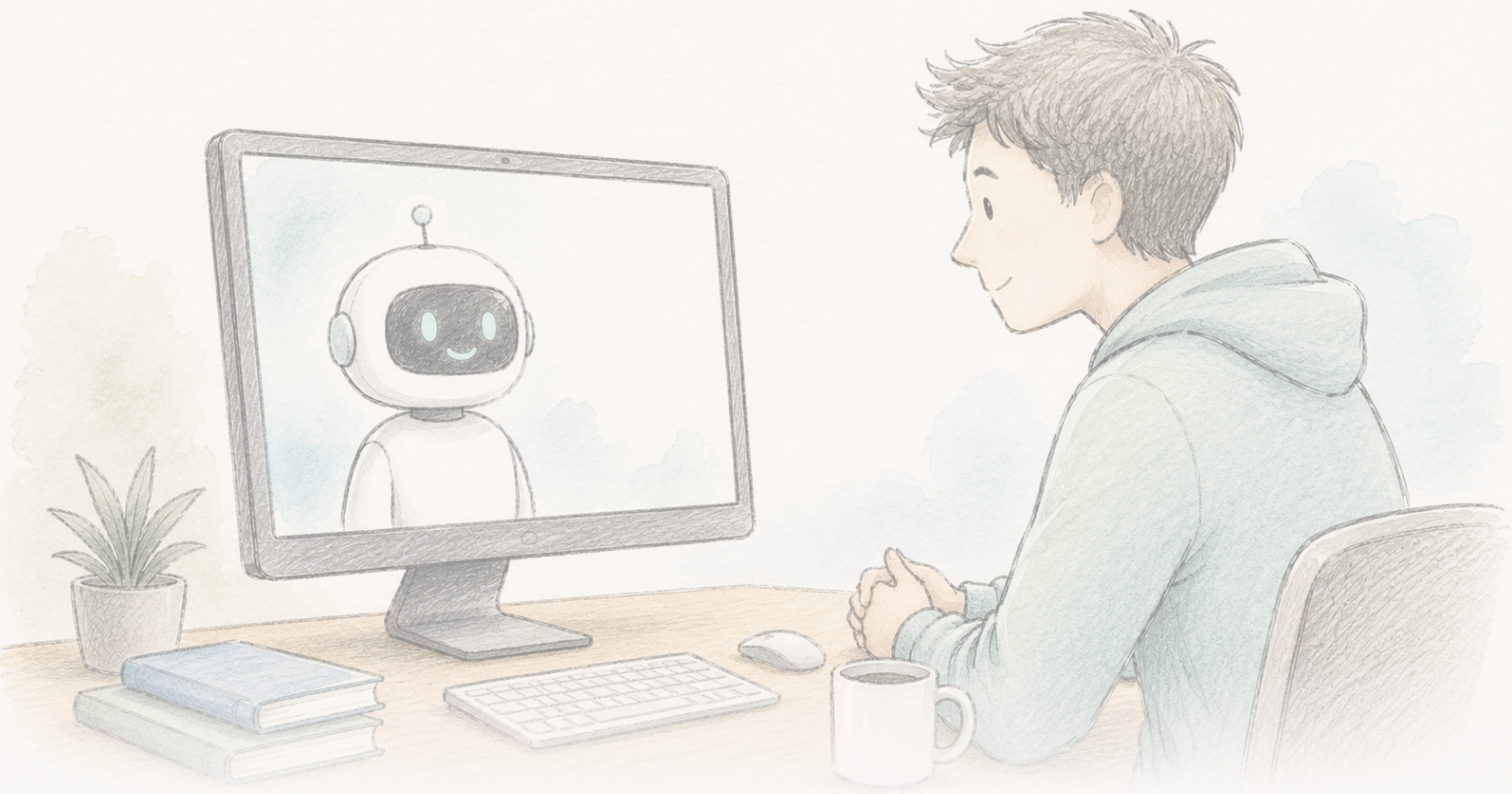




The CLEAN PAPER

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

Chatbot kan dà bí ẹnì pé ó dín conspiracy beliefs kù fún oṣù

Study 2024 kan rí pé three-round conversation pèlú GPT-4 Turbo dín belief nínú conspiracy theory tí participants gbà kù ní ayíká 20%, effect tó pé oṣù méjì, tí ó generalize kọjá conspiracy types, pèlú AI claims tí a rated highly accurate. Ní June 2026, paper nàà gba Editorial Expression of Concern fún data-handling àti reproducibility problems; authors sọ corrected analysis pa direction, significance àti size mọ, Science sì sì n evaluate. Result tó fanimọra, tí a kọ pèlú iṣọra — sùgbọ́n tí review sì n 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 yí nígbà tí ó ní evaluate data-handling àti reproducibility questions. Kì í ẹ̀ retraction, kò sì jẹ́ finding of misconduct; ó túmọ̀ sí pé reported result yẹ́ ká ka gégẹ́ bí provisional tí review yóò fi parí.

Editorial Expression of Concern →

Nígbà míù, facts lẹ̀ ẹ̀ṣẹ́ — ẹ̀gbẹ̀n ìwé ìwádìí nàà ní flag kan

Ní 2024, ìwádìí kan rọ̀yìn ohun tó tako conventional wisdom kan tí ó rọ̀rùn láti gbà. Fún ènìyàn ní short, three-round conversation pèlú AI — GPT-4 Turbo, tí a prompt láti dáhùn ẹ̀rì pàtó tí ẹ̀nì nàà ní ló fún conspiracy àbá-ẹ̀kọ́ tí ó gbàgbọ́ — ní apapọ̀, ìgbàgbọ́ ẹ̀nì nàà nínú àbá-ẹ̀kọ́ nàà dín kù ní ayíká 20%. Ìdínkù nàà sì hàn oṣù méjì lẹ́yìn nàà. Ó ẹ̀ṣẹ́ kọ́já classic conspiracies (assassination John F. Kennedy, aliens, Illuminati) àti topical ones (COVID-19, 2020 US election), paápáà lára àwọn ènìyàn tí ìgbàgbọ́ wọn lágbara tí ó sì ní ìbáṣepọ̀ pèlú identity. Nígbà tí professional fact-checker kan grade claims 128 tí AI ẹ̀, 127 jẹ́ true, ọ̀kan misleading, kò sí false.

àkólé tó tàn ká ni pé o lẹ̀ bá ènìyàn sọ̀rọ̀ kúrò nínú rabbit hole — pé conspiracy believers kò kọ́já ibi tí ẹ̀rì lẹ̀ dé, wọn kan nílò ẹ̀rì tó tọ́, tí a fi hàn ní ọ̀nà tó specific. Claim yí fànimọ́ra gidi, àpẹ̀rẹ́ ìwádìí nàà sì sọ́ra ju ọ̀pọ̀ persuasion research lọ.

Lẹ́yìn nàà, ní June 2026, *Science* fi **Editorial Expression of Concern** sí ìwé ìwádìí nàà. Èyí ni apá tí ọ̀pọ̀ retellings lẹ̀ fo kọ́já, ẹ̀gbẹ̀n apá yí gan-an ló pinnu iye weight tí èsì lẹ̀ ru ní bá yí.

Kí ni Expression of Concern — àti ohun tí kò jẹ́

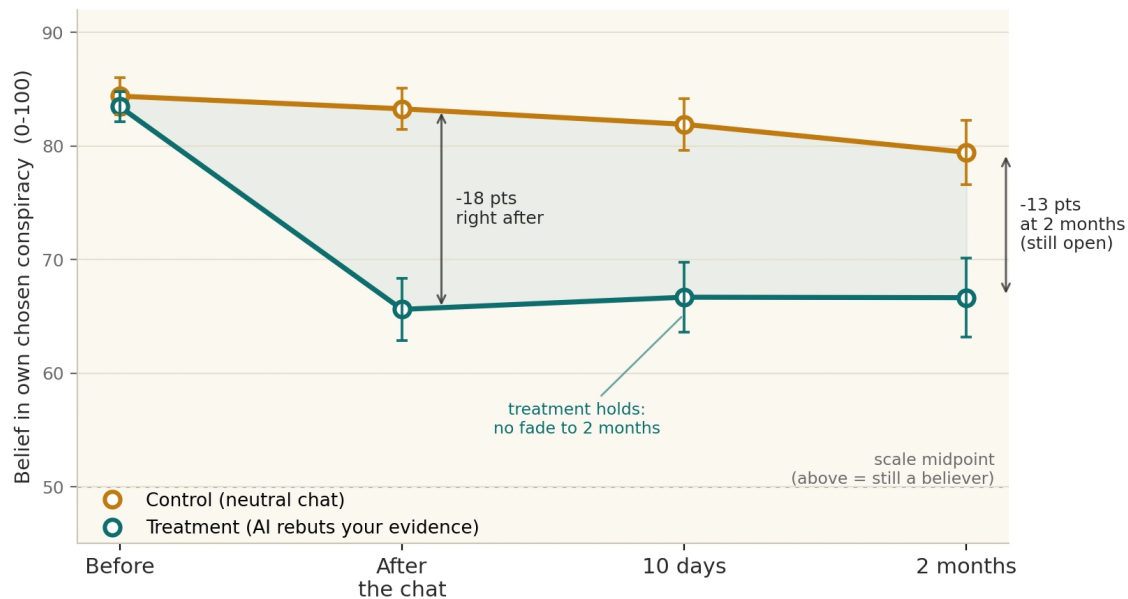
Editorial Expression of Concern jẹ́ formal flag tí journal máa ní fi sí ìwé ìwádìí láti kílò fún readers pé ìbèèrè tí dide, nígbà tí a sì ní ẹ̀ṣẹ́ lórí wọn. **Kì í ẹ̀ retraction** — a kò yọ ìwé ìwádìí kúrò — kò sì túmọ̀ sí àwárí of misconduct fúnra rẹ̀. Ó túmọ̀ sí: ka ìwé ìwádìí pèlú caveat yí, nítorí scientific àkọ̀sílẹ̀ lẹ̀ yí. Níhìn-ín, lẹ́yìn tí òhńkọ̀wé mọ́ issues nàà, wọn ẹ̀ investigation, wọn rọ̀yìn specifics sí *Science*, wọn sì fi corrected itúpalẹ̀ ránṣẹ́.

Ohun tí a flag jẹ́ specific. A sọ́ fún òhńkọ̀wé pé inconsistencies wà nínú bí participant-screening criteria ẹ̀ apply láàárín manuscript àti published itúpalẹ̀ code, àti pé public dataset ní extraneous rows tí code-merging àṣìṣe fi splice wọ́ inú rẹ̀ — ìṣòro tó mú kí díẹ̀ ninu reported numbers nira láti reproduce láti released ohun èlò. Òhńkọ̀wé fún *Science* ní corrected itúpalẹ̀ pipeline àti updated èsì, tí wọn sọ́ pé ó bá original mu **ní ìtọ̀sọ̀nà, statistical significance àti substantive ìwọ̀n**. *Science* ní evaluate wọn. Títí evaluation yẹn yóò parí, exact figures sì wà lábẹ́ ìbèèrè mark tí journal nikan lẹ̀ yọ.

Nítorí nàà, itan méjì ni èyí ní àkókò kan: intriguing èsì nípa bóyá facts lẹ̀ yí entrenched beliefs, àti live example tí scientific àkọ̀sílẹ̀ tó ní ẹ̀ correction ara rẹ̀ ní gbangba. Kókó article yí ni láti má ẹ̀ da àwọn itan méjì nàà 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 iwádíí nàà, three-round AI conversation tó rebut ẹrí tí olukopa funra rẹ sọ ni a ròyìn pé ó dín belief nínú conspiracy tí ẹni nàà yan kù ní ayíká 20% ní apapọ; unlike control chat lórí unrelated topic, drop nàà ẹ̀ measurable ní 10 days àti 2 months. Effect nàà durable sùgbón partial: ọpọ olukopa ẹ̀ gbà conspiracy nàà lẹyìn nàà — reduction, kì í ẹ́ cure. iwé iwádíí nàà wà lábẹ Editorial Expression of Concern (*Science*, June 2026) lórí reporting inconsistencies àti àṣìṣe nínú public dataset; òhòkòwé sọ corrected itupalẹ̀ pa itọ̀sọ̀nà, significance àti iwòn mọ, nígbà tí *Science* ẹ̀ n evaluate. The Clean iwé iwádíí tún chart yíi kọ látí public dataset nàà, nítorí nàà values yíi provisional ní, kì í ẹ́ final figures iwé iwádíí.

Original figure — The Clean Paper CC BY 4.0

Ohun tí àwọn olùkòwé ẹ

- Wọn ẹ̀ idánwò méjì pẹ̀lú olukopa 2190, tí ọkọọkan sọ — ní ọrò tirẹ — conspiracy àbá-ẹ̀kọ́ tí ó gbàgbọ̀ àti ẹ̀rí tí ó rò pé ó ẹ̀ atilẹ́yìn rẹ.
- Participant ọkọkan ní three-round conversation pẹ̀lú GPT-4 Turbo. Nínú **itọju** condition, AI ní prompt látí rebut participant's specific ẹ̀rí; nínú **control** condition, ó jíròrò unrelated topic.
- Wọn wọn belief nínú chosen conspiracy sáájú àti lẹ́sẹ́kẹ́sẹ̀ lẹ́yìn conversation, wọn sì tún contact olukopa ní **10 days àti 2 months**.
- Professional fact-checker kan evaluate ipéye gbogbo 128 factual claims tí AI ẹ̀ nínú àpẹrẹ kan.
- Wọn sàyèwò bóyá ipa nàà spill over sí other unrelated conspiracies — àti, gégé bí specificity idánwò, bóyá ó tún dín belief nínú conspiracies tí ó jẹ́ **true** kù.

Ohun tí wọn rí

- **Ayíká 20% average reduction** nínú belief nínú chosen conspiracy fún ìtoju ẹgbẹ relative sí control (ìwádíí 1: 95% ìgbẹkẹlẹ interval [13.8, 19.7] kókó lórí 100-kókó scale, $P < 0.001$).
- **Ó pẹ.** Nínú ìtoju ẹgbẹ, drop náà kò fade: belief dúró nítòsì just-after-chat ìpele ní 10 days àti 2 months dípò kó padà sókè, control sì dúró higher. ìwé ìwádíí náà sọ ipa lórí chosen conspiracy tèsíwájú láì dín kù fún o kere tán oṣù méjì, kì í ẹ momentary wobble.
- **Ó generalize** kojá ààlà ti conspiracies tí people named, ó sì ẹ ẹ paápáá fún olukopa tí initial belief wọn lágbára.
- **AI claims jẹ accurate** nínú setting yí: 127 nínú 128 (99.2%) ni rated true, 1 (0.8%) misleading, kò sí false.
- **Ó specific**, kì í ẹ blanket doubt: intervention kò dín belief nínú true conspiracies kù, tó n dàba pé ó yí unsupported beliefs padà dípò kó mú people cynical nípa ohun gbogbo.
- **Spillover diẹ wà** — belief nínú other unrelated conspiracies dín kù ní ayíká 12%, olukopa sì rọ̀yìn greater intention láti push back lórí conspiracy claims. Main ipa dúró nínú ìwádíí 2, ó sì tọ sí ìtọ̀sọ̀nà kan náà nínú smaller àpẹrẹ tí kò ní control ẹgbẹ (weaker ẹrì, ẹ̀gbọ̀n consistent).

Ohun tí ẹyí kò fi ẹrì múlẹ̀

- Ní bá yí ìwé ìwádíí náà kò dúró **láìsí ibéèrè**. Ó wà lábẹ Editorial Expression of Concern fún data-handling àti reproducibility ẹ̀sòrò, corrected numbers sì ẹ ní jẹ evaluated nipasẹ *Science*. Ka exact values bí provisional tí review yẹn yòò parí.
- Kò fi hàn pé AI “**cures**” conspiracy belief. ~20% average reduction jẹ meaningful shift, kì í ẹ erasure — ọ̀pọ̀ olukopa sì gbà àbá-ẹ̀kọ̀ náà lẹyìn, wọn kan gbà á kéré sí.
- Kì í ẹ **ayé gidi deployment èsì**. Participants yan láti wọ ìwádíí, wọn sì engage pẹ̀lú AI ní good faith; ní gidi world, àwọn tó committed jù sí conspiracy lè jẹ àwọn tí kò fẹ bá chatbot kan sọ̀rọ̀ tí yòò tako wọn.
- 99.2% ìpéye jẹ **specific sí ẹ̀sẹ-ẹ̀sẹ yí, àwòṣe yí àti careful prompting** — kì í ẹ gbogbogbò guarantee pé èdè àwòṣe máa ní sọ truth. ònḱòwé sọ kedere pé kan náà personalized persuasive power lè push *false* beliefs tí a bá lo o sí i.
- “Durable” nìhìn-ín tùmọ̀ sí **oṣù méjì**, ohun tó impressive fún conversation kan, ẹ̀gbọ̀n kì í ẹ permanent.

Báwo ni ẹrì ẹe lágbára tó?

- **Design jẹ gidi strength.** Controlled treatment-versus-control idánwò, clean placebo-style control, replication kojá ìwádíí méjì àti independent àpẹrẹ, durability follow-ups, explicit ìpéye check lórí AI claims — ẹyí sọra ju ọ̀pọ̀ persuasion research lọ, ìtọ̀sọ̀nà ti ipa sì consistent ní gbogbo ibi tí a idánwò.
- **Ẹ̀gbọ̀n Expression of Concern kì í ẹ footnote.** Ability láti regenerate reported numbers láti released data àti code jẹ apá trustworthiness, ìyẹn gan-an ni ohun tí ó fọ — screening-criteria inconsistencies àti duplicated rows láti code-merging àṣiṣe. Statement ònḱòwé pé corrected pipeline pa èsì mó jẹ encouraging, ó sì ẹé gbà, ẹ̀gbọ̀n ó sì jẹ account ònḱòwé, kì í ẹ independent verdict. Evaluation *Science* ni ohun tó yẹ ká dúró de.
- **Status tó tọ ni “flagged,” kì í ẹ “debunked” tàbí “confirmed.”** ìwádíí tó kọ dáadára, striking, ẹ̀gbọ̀n exact numbers rẹ wà lábẹ formal review. Ó lè jẹ uncomfortable láti fi itàn rere sí ipo bẹẹ, ẹ̀gbọ̀n ipo tó pe ni.

Kí nìdí tí ó fi ẹ̀ pàtàkì?

Fún ọ̀dún púpò, view tó influential ni pé conspiracy believers ni psychological needs àti identity ní darí, nítorí náà facts kò lè yí belief wọ̀n. Durable, evidence-based reduction — tí ó bá survive review — jẹ́ counterweight tó meaningful sí pessimism yẹn: ó daba pé apá ọ̀dòro ni pé generic debunking kò ní bá *specific* ẹ̀rí tí believer gan-an ní cite ẹ̀şẹ̀, ohun tí LLM lè ẹ̀ at scale.

Ó tún ẹ̀ pàtàkì gégé bí ọ̀ràn ìwádíí ní bí science yẹ kí ó ẹ̀şẹ̀. Èkọ́ Expression of Concern kì í ẹ̀ pé celebrated èsì kò ní iye; ó jẹ́ pé àşìşẹ̀ ní hàn, dátà ní jẹ́ re-examined, claims sì ní jẹ́ re-checked ní gbangba. Machinery yí tó ní ẹ̀şẹ̀ jẹ́ àbùdá, kì í ẹ̀ scandal — ìdí ni pé posture tó tò sí èsì yí ni patience, kì í ẹ̀ applause lẹ̀şẹ̀şẹ̀ tàbí dismissal.

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

Àkótán kedere

ìwádíí 2024 kan rí pé three-round conversation pèlú GPT-4 Turbo dín belief àwọ̀n ènìyàn nínú conspiracy àbá-ẹ̀kọ́ tí wọ̀n gbà kù ní ayíká 20%, ipa tó tẹ̀sìwájú fún oşù méjì, tó sì generalize kọ́já conspiracy types, pèlú AI factual claims tí a rated overwhelmingly accurate. Ní June 2026, ìwé ìwádíí náà wà lábẹ́ Editorial Expression of Concern fún data-handling àti reproducibility ọ̀dòro; òhńkòwé rọ̀yìn pé corrected itúpalẹ̀ pa itọ̀sọ̀nà, significance àti ìwọ̀n tí àwárí mọ́, *Science* sì şì ní evaluate. Ka a gégé bí genuinely interesting, carefully built èsì tí a şì ní review — kì í ẹ̀ settled, kì í ẹ̀ debunked. Gbolohun tó honest jù nípa rẹ̀ ni ohun tí ilà̀nà náà funra rẹ̀ ní kọ: şàyẹ̀wò lẹ̀şẹ̀kan sí i.

Àwọ̀n 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>