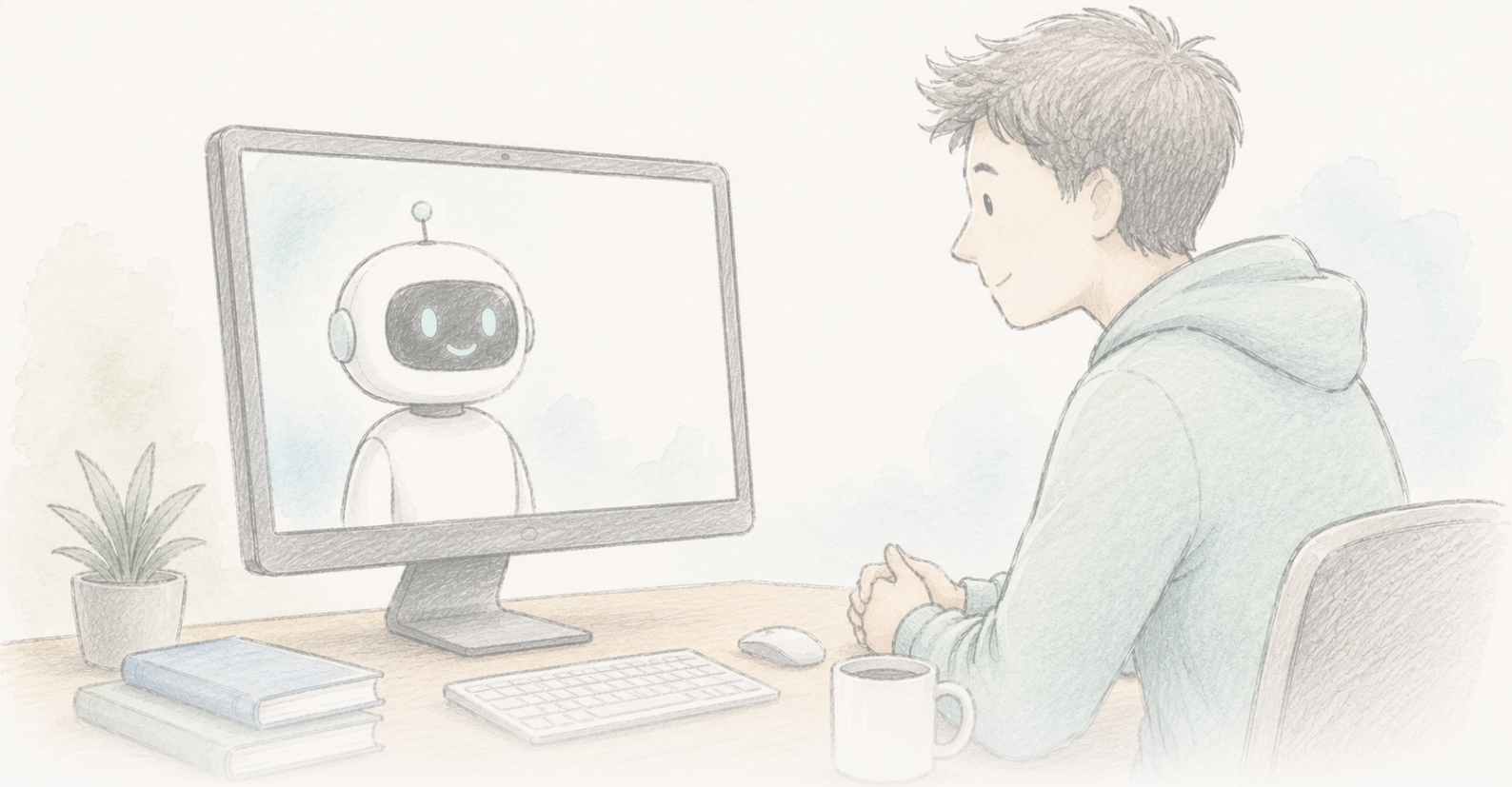


The Clean Paper

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

Chatbot tokko amantii conspiracy ji'ootaaf hir'isu fakkaate

*Qorannoon 2024 GPT-4 Turbo waliin haasawa rounds sadii amantii conspiracy namni tokko qabu naannoo 20% hir'ise jedhe; effect ji'a lama ture, conspiracy types adda adda irratti generalize godhe, claims AI ammoo baay'inaan accurate jedhamee madaalame. Waxabajjii 2026 paperichi data-handling fi reproducibility problems irratti Editorial Expression of Concern jala gale; authors corrected analysis direction, significance fi size finding eega jedhu, *Science* ammallee evaluate godhaa jira. Bu'aa dhugumaan nama hawwatu, of-eeggannoon ijaarame — amma review jala, settle hin taane, debunk illee hin taane.*

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 paper kana irratti Editorial Expression of Concern kaa'eera, data-handling fi reproducibility questions qorachaa jira. Kun retraction miti fi misconduct mirkaneessuu miti; review hanga xumuramutti result provisional ta'ee dubbifamuu qaba.

Editorial Expression of Concern →

Dhuma irratti, facts — garuu barruu irratti mallattoo of-eeggannoo jira

Bara 2024 qorannoon tokko yaada namaaf mijataa ta'e tokko faallessu gabaase. Nama tokko AI waliin haasawa gabaabaa, rounds sadii qabu keessa galchi — GPT-4 Turbo, ragaa inni conspiracy tiyoorii inni ofumaan amanuuf dhiyeessu irratti kallattiin deebii akka kennu ajaja godhamee — fi giddugaleessa irratti amantaan isaa tiyoorii sana irratti naannoo 20% gadi bu'a. Hir'inni sun ji'a lama booda illee ture. Conspiracies klassikii (ajjeechaa John F. Kennedy, aliens, Illuminati) irraa kaasee hanga topics yeroo ammaa (COVID-19, filannoo US 2020) irratti hojjete; namoota amantaan isaanii cimaa fi identity isaanii waliin hidhata qabu irratti illee. Ogeessi dhugaa-mirkaneessaa himannoowwan AI 128 keessaa 127 dhugaa, tokko misleading, fi tokko illee soba miti jedhee madaale.

mata-duree baay'ee faca'e: namoota "rabbit boolla" keessaa haasawaan baasuun **ni danda'ama** — namoonni conspiracy amanan ragaa irraa ala miti; ragaa sirrii, isaanii irratti qajeelfame barbaadu qofa. Kun himannaa dhugumaan nama hawwatu dha, fi qorannoon persuasion research hedduu caalaa of-eeggannoon ijaarame.

Garuu Waxabajjii 2026 keessatti *Saayinsii* barruu sana irratti **Editorial Expression of Concern** maxxanse. Kutaan kun retellings hedduun dhiisan ta'uu danda'a, garuu amma bu'aa kana irratti ulfaatina hammam kaa'uu akka dandeenyu murteessa.

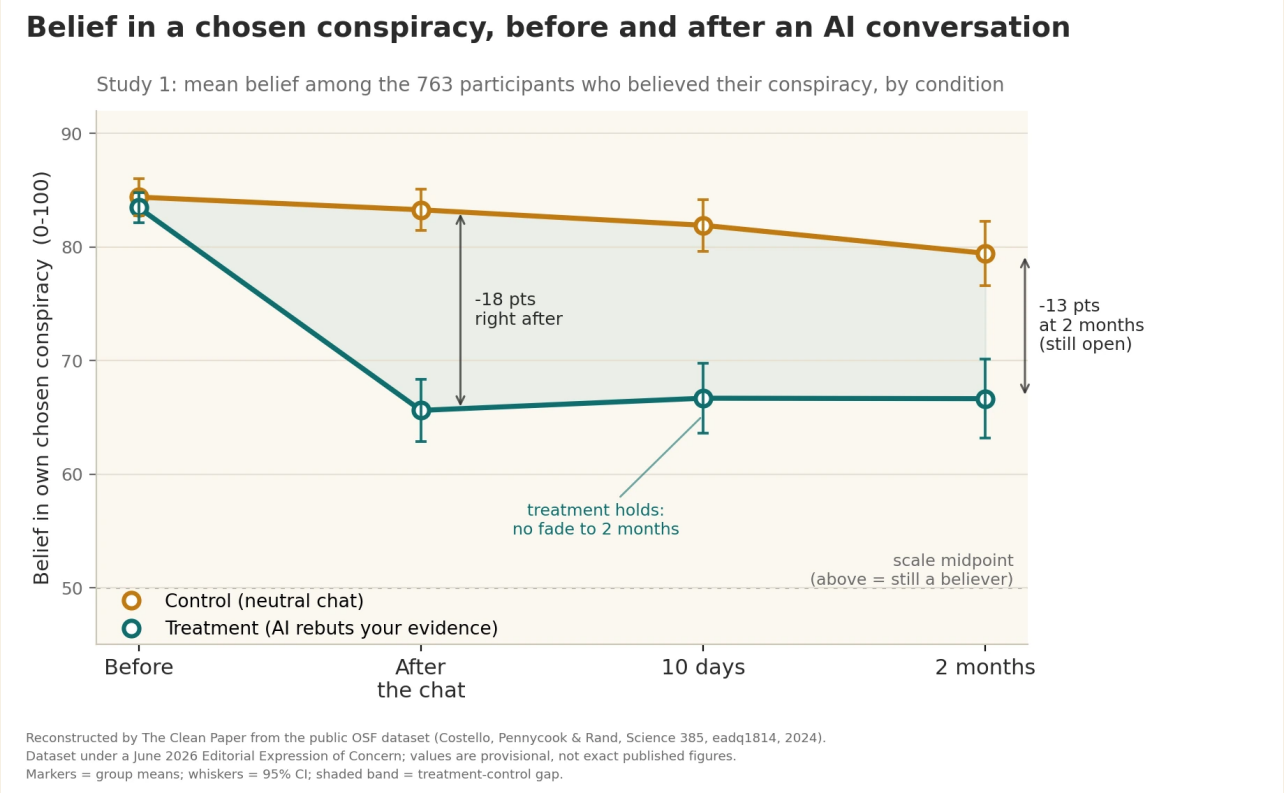
Expression of Concern maal dha — maal miti

Editorial Expression of Concern jechuun journal tokko readers akka of-eeggatan barruu irratti mallattoo idilee kaa'u dha, yeroo gaaffileen ka'an ammallee qoramaa jiran. Kun **retraction miti** — barruun hin baafamne — fi ofumaan misconduct mirkaneessuu miti. Hiikni isaa: caveat kana qalbeeffadhu; galmee jijjiiramuu danda'a. As, barreessitoonni rakkoo akka jiru erga beekanii booda qoratan, details *Saayinsii*tti gabaasan, fi xiinxala corrected dhiyeessan.

Wanti flagged ta'e addaa dha. Participant-screening ulaagaalee manuscript fi maxxanfame xiinxala koodii gidduutti akkaataa itti applied ta'e irratti inconsistencies turan; ummataaf banaa kuusaa daataa keessattis rows dabalataa code-merging dogoggora irraa spliced in ta'an jiru — kun gabaafame lakkoofsota keessaa muraasa gadhiifame maateriyaalota irraa deebisee uumu rakkise. Barreessitoonni *Saayinsii*tti corrected xiinxala pipeline fi updated bu'aawwan kennan; isaan jalqabaa bu'aa **direction, istaatistikii barbaachisummaa istaatistikii fi substantive hammamtaa** keessatti

wal fakkaata jedhu. *Saayinsii* qorachaa jira. Hanga gamaaggama sun xumuramutti, sirrii fakkiiwwan provisional dha.

Kanaaf seenaa lama yeroo tokkootti qabna: facts amantii cimaa sochoosuu danda'aa jedhu irratti bu'aa nama hawwatu, fi saayinsawaa galmee ofumaan banaa ta'een sirreeffamaa jiru fakkeenya live. Kaayyoon barruu kanaa lamaan walitti hin makne eeguu dha.



Qorannoo keessatti AI waliin rounds sadii haasawuun ragaa hirmaataa ofumaan dhiyeesse rebut godhu, amantii conspiracy inni filate irratti giddugaleessa naannoo 20% hir'ise jedhamee gabaafame; to'annoo chat topic adda ta'e waliin wal-bira qabamee, hir'inni 10 days fi 2 months booda illee safaramuu danda'e. Dhiibbaa durable garuu partial ture: hirmaattota baay'een booda illee conspiracy sana amanu turan — hir'ina, "cure" miti. barruun amma gabaasuu inconsistencies fi ummataaf banaa kuusaa daataa dogoggora irratti Editorial Expression of Concern (*Saayinsii*, June 2026) jala jira; barreessitoonni corrected xiinxala direction, barbaachisummaa istaatistikii fi hammamtaa dhiibbaa eega jedhu, *Saayinsii* ammoo qorachaa jira. Chaartiin kun The Clean Paper'n kuusaa daataa ummataaf banaa sana irraa deebi'ee ijaarame; gatiiwwan yeroo ammaa kanneen jalqabaa dha, fakkiiwwan dhumaa barruu keessaa miti.

Original figure — The Clean Paper CC BY 4.0

Barreessitoonni maal godhan

- Yaaliiwwan lama hirmaattota **2190** waliin raawwatan. Namni tokko tokko conspiracy tiyoorii inni amanu fi ragaa isa deeggaru jedhee yaadu jecha ofiitiin ibse.
- Hirmaataa tokko tokko GPT-4 Turbo waliin haasawa rounds sadii qaba. **Yaala** keessatti AI ragaa

addaa hirmaataa sanaa rebut gochuuf ajaja ta'e; **to'annoo** keessatti topic wal hin qabne irratti haasa'e.
- Amantii conspiracy filatame haasawa dura fi battalumatti booda safaran; hirmaattota **guyyaa 10 fi ji'a 2** booda deebi'anii qunnamti.
- Ogeessi dhugaa-mirkaneessaa himannoowwan dhugaa irratti hundaa'e AI saamudaa keessa dhiyeesse 128 hunda sirrummaa irratti madaale.
- Dhiibbaa conspiracies biroo, wal hin qabne irratti gara amantiiwwan biraatti darbuun isaa qoran; specificity qormaata ta'ee conspiracy dhugaa ta'an irratti amantii illee gadi buusaa ilaale.

Maal argatan

- garee yaalaa keessatti to'annoo waliin wal-bira qabamee, conspiracy filatame irratti **giddugaleessa gara 20% hir'ina** (qorannoo 1: 95% daangaa amanamummaa [13.8, 19.7] qabxiilee safartuu 100 irratti, $P < 0.001$).
- **Ture.** garee yaalaa keessatti hir'inni hin badne: guyyaa 10 fi ji'a lama booda belief sadarkaa haasawa booda jiru naannoo ture; to'annoo caalaatti ol ture. barruun dhiibbaa conspiracy filatame irratti ji'a lamaaf osoo hin dadhabin hafe jechuun ibsa.
- **haalawwan biraattis hojii irra oolchu**, conspiracies adda adda hirmaattota waaman irratti; amantaan jalqabaa cimaa ta'es dhiibbaa ture.
- **Himannoowwan AI setting kana keessatti sirrii turan:** 127/128 (99.2%) true, 1 (0.8%) misleading, none false.
- **Addaa ture**, shakkina waliigalaa miti: intervention conspiracy dhugaa ta'an irratti amantii hin hir'isne. Kun unsupported beliefs sochoosuu malee namoota waan hunda shakku hin goone jedhu deeggaru.
- **Spillover xiqqoo** ture — conspiracies biroo irratti belief naannoo 12% hir'ate, hirmaattota conspiracy himannoowwan irratti deebii kennuuf intention guddaa gabaasan. Ijoo dhiibbaa qorannoo 2 keessatti ture; saamudaa xixiqqoo to'annoo hin qabne keessattis kallattii wal fakkaataa agarsiise (ragaa dadhabaa, garuu wal-simu).

Kun maal hin mirkaneessu

- Amma **gaaffii irraa bilisa miti**. barruun data-handling fi reproducibility rakkoowwan irratti Editorial Expression of Concern jala jira; corrected lakkoofsota *Saayinsii* dhaan ammallee evaluate ta'aa jiru. Sirrii gatiiwwan akka provisionalitti ilaali.
- AI conspiracy belief "fayyisa" jechuu **hin agarsiisu**. giddugaleessa ~20% hir'inni hiika qabu dha, garuu belief balleessuu miti; hirmaattota baay'een booda illee amanu turan, qofa xiqqeessan.
- Kun **haala addunyaa dhugaa hojiirra oolmaa bu'aa miti**. Hirmaattota qorannoo keessaa ofumaan seenan fi AI waliin gaarii faith keessatti haasa'an. Outside lab, namoonni conspiracy irratti baay'ee committed ta'an chatbot isaaniin falmu barbaaduu keessaa isa gadi-aanaa ta'uu danda'u.
- sirrummaa 99.2% **hojii, moodeela fi prompting of-eeggannoo kanaaf addaa** dha; afaan moodeelota yeroo hunda dhugaa jedhuuf guarantee miti. Barreessitoonni persuasive humna personal-

ized sun false beliefs illee kaasuuf fayyadamuu danda'a jedhu.
- "Durable" asitti **ji'a lama** jechuu dha; conversation tokkoof impressive, garuu permanent miti.

Ragaan hammam cimaa dha

- **Qophii strength dhugaa dha.** To'atame yaala vs to'annoo, placebo-style to'annoo qulqulluu, replication qorannoowwan lama fi saamudaa walaba, follow-ups durability, fi explicit sirrummaa sakatta'a himannoowwan AI irratti — persuasion research hedduu caalaa of-eeggannoo qaba; direction dhiibbaa bakka qoratame hunda keessatti wal-simu ture.
- **Garuu Expression of Concern footnote miti.** Gadhiifame daataa fi koodii irraa gabaafame lakkoofsota deebi'anii uumuun trustworthiness keessaa kutaa dha; rakkoon sirriitti achitti mudate — screening-criteria inconsistencies fi duplicated rows code-merging dogoggora irraa. Barreessitoonni corrected pipeline bu'aa eega jedhu encouraging fi amansiisaa dha, garuu barreessitoonni account qofa, walaba verdict miti. *Saayinsii* madaallii eeguu qabna.
- **Status amanamaa "flagged" dha; "debunked" yookaan "confirmed" miti.** Qorannoo sirriitti ijaarame, nama hawwatu, sirrii lakkoofsota gamaaggama idilee jala jiran. Seenaan gaariin bakka kana irratti hafuun namaaf hin mijatu, garuu sirrii dha.

Maaliif barbaachisaa dha

Waggoota hedduuf yaadni dhiibbaa guddaa qabu namoonni conspiracy amanan psychological needs fi identity irraa ka'anii, facts'n irraa haasawuun hin danda'amu jedhu ture. Yoo bu'aan kun gamaaggama booda dhaabbate, hir'ina durable, evidence-based kun pessimism sanaaf counterweight hiika qabu dha: waliigalaa debunking ragaa **addaa** nama amanu sanaa irratti hin qajeelne waan ta'eef rakkoo keessaa kutaa ta'uu danda'a; LLM ammoo safartuu irratti akkas gochuu danda'a.

Saayinsii akkamitti hojjechuu qabu irratti haala qorannoo ta'uunis barbaachisaa dha. Expression of Concern lesson isaa "celebrated bu'aawwan faayidaa hin qabu" miti; dogoggorawwan ifa bahu, daataa irra deebi'ee qoratamu, himannoowwan ummataaf banaa keessatti sakatta'a godhamu dha. Machinery kun hojjechuun amaloota dha, scandal miti — kanaaf posture sirrii bu'aa kanaaf patience dha, applause yookaan dismissal ariifachiisaa miti.

AI irratti lamaan kallattii qaba. dandeettii false belief tokko tailored argument'n gadi buusu, akkuma sana false belief tokko ol kaasuu danda'a. Technique neutral dha; kaayyoon neutral miti.

Cuunfaa gabaabaa

Qorannoon 2024 tokko GPT-4 Turbo waliin haasawa rounds sadii godhuun amantii conspiracy tiyoorii namni tokko qabu naannoo 20% hir'ise jedhe; dhiibbaa ji'a lama ture, conspiracy gosoota adda adda irratti haalawwan biraattis hojii irra oolchu, himannoowwan dhugaa irratti hundaa'e AI ammoo baay'inaan sirrii jedhaman. Waxabajjii 2026 barruun data-handling fi reproducibility rakkoowwan irratti Editorial Expression of Concern jala gale; barreessitoonni corrected xiinxala direction, barbaachisummaa istaatistikii fi hammamtaa argannoo eega jedhu, *Saayinsii* ammallee evaluate godhaa jira. Kana bu'aa dhugumaan nama hawwatu, of-eeggannoon ijaarame, amma gamaaggama jala jiru akka ta'etti dubbisi — settle hin taane, debunk illee hin taane. Hima caalaatti amanamaa: irra deebi'ii sakatta'a godhi.

Maddoota

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