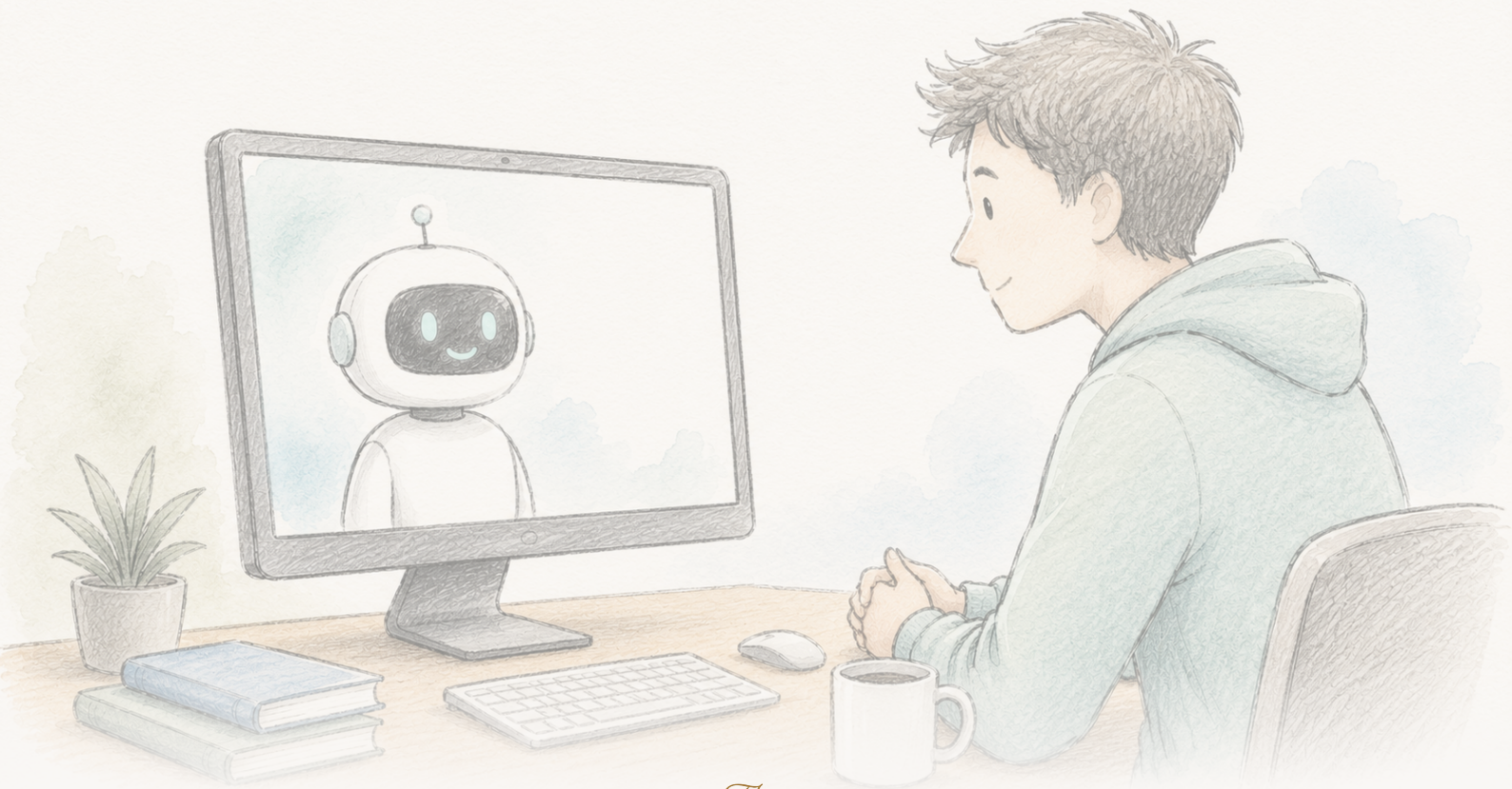




The CLEAN PAPER

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

Падобна, размова з чатботам на месяцы аслабляла веру ў тэорыі змовы

Даследаванне 2024 года паказала, што трохраўндавая размова з GPT-4 Turbo памяншала веру людзей у тэорыю змовы, якой яны прытрымліваліся, прыкладна на 20%; эфект захоўваўся два месяцы, назіраўся для розных тыпаў змоў і абаніраўся на сцвярджэнні III, якія фактчэкер ацаніў як амаль цалкам дакладныя. У чэрвені 2026 года да артыкула было дададзена Editorial Expression of Concern праз праблемы з апрацоўкай даных і ўзнаўляльнасцю; аўтары сцвярджаюць, што выпраўлены аналіз захоўвае кірунак, значнасць і велічыню выніку, а Science усё яшчэ праводзіць ацэнку. Сапраўды цікавае і старанна пабудаванае даследаванне, якое цяпер пераглядаюць: яно яшчэ не канчаткова пацверджана, але і не абвергнута.

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 дадаў да гэтага артыкула Editorial Expression of Concern, пакуль ацэньвае пытанні, звязаныя з апрацоўкай даных і ўзнаўляльнасцю. Гэта не рэтракцыя і не выснова пра парушэнне; гэта азначае, што апублікаваны вынік варта лічыць папярэднім, пакуль праверка не завершана.

Editorial Expression of Concern →

Урэшце факты — і папераджальная адзнака на артыкуле

У 2024 годзе даследаванне паведаміла вынік, які супярэчыць зручнай частцы агульнапрынятай мудрасці. Дайце чалавеку кароткую размову з ШІ ў тры раўнды — GPT-4 Turbo, настроеным адказваць менавіта на тые доказы, якія чалавек прыводзіць у падтрымку тэорыі змовы, у якую сам верыць, — і ў сярэднім яго вера ў гэтую тэорыю памяншаецца прыкладна на 20%. Зніжэнне ўсё яшчэ назіралася праз два месяцы. Эфект працаваў і для класічных змоў (забойства John F. Kennedy, іншапланецяне, ілюмінаты), і для актуальных (COVID-19, выбары ў ЗША 2020 года), нават сярод людзей з моцнай верай, звязанай з іх ідэнтычнасцю. Калі прафесійны фактчэкер ацаніў 128 сцвярджэнняў, зробленых ШІ, 127 аказаліся праўдзівымі, адно — такім, што ўводзіць у зман, і ніводнае — ілжывым.

Загалолак, які разляцеўся па свеце, быў такі: людзей *можна* выцягнуць з «трусінай нары» размовай — прыхільнікі тэорыі змовы не недасягальныя для доказаў, ім проста патрэбныя правільныя доказы, пададзеныя дастаткова канкрэтна. Гэта сапраўды цікавае сцвярджэнне, і само даследаванне было пабудавана старанней, чым большасць прац пра перакананні.

А затым, у чэрвені 2026 года, *Science* апублікаваў да артыкула **Editorial Expression of Concern** — рэдакцыйнае паведамленне пра ўзніклыя пытанні. Менавіта гэтую частку большасць пераказаў прапусціць, хоць цяпер менавіта яна вызначае, якую вагу можа несці вынік.

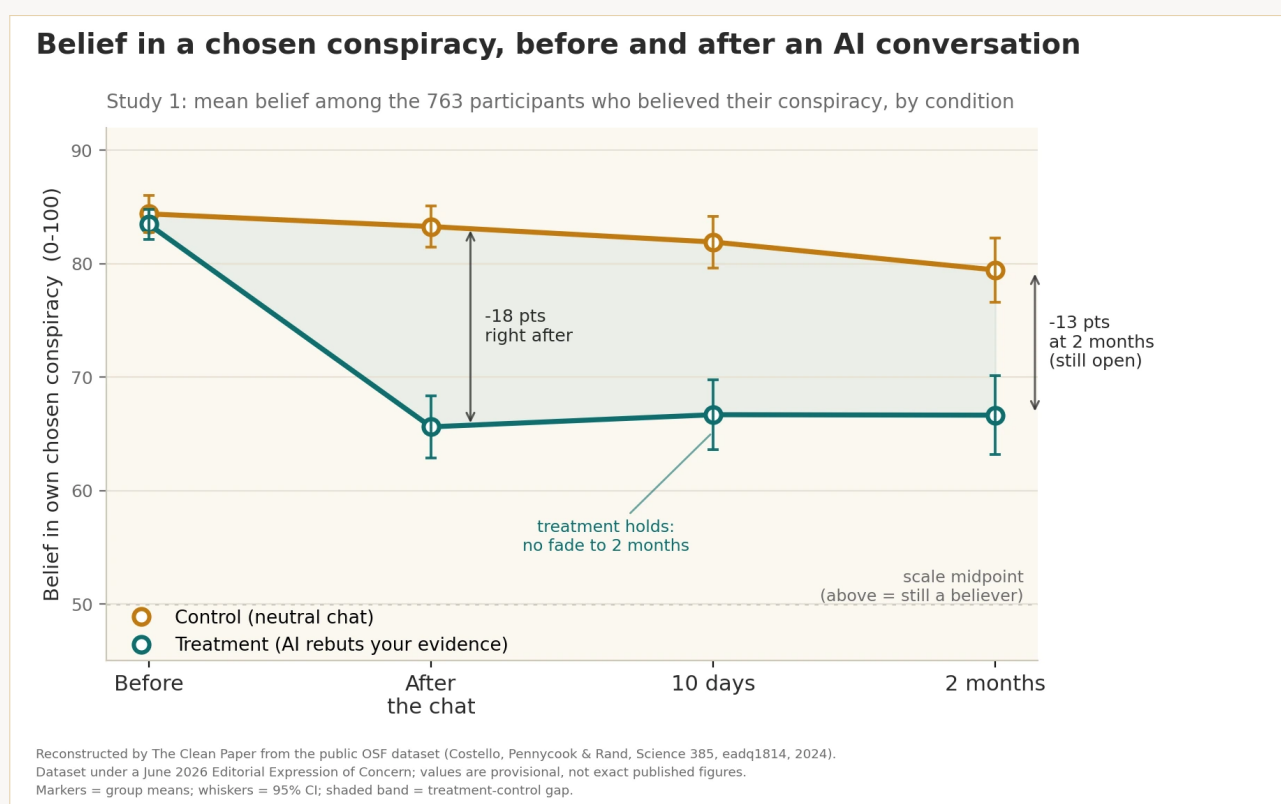
Што такое Expression of Concern — і чым яно не з'яўляецца

Editorial Expression of Concern — афіцыйная адзнака, якую часопіс дадае да артыкула, каб папярэдзіць чытачоў: да працы ўзніклі пытанні, і іх усё яшчэ высвятляюць. Гэта **не** рэтракцыя — артыкул не адклікалі — і само па сабе гэта не выснова пра парушэнне навуковай этыкі. Яно азначае: чытайце з агаворкай, бо навуковы запіс можа змяніцца. У гэтым выпадку пасля таго, як аўтараў апавясцілі пра праблемы, яны правялі праверку, далі *Science* канкрэтныя дэталі і перадалі выпраўлены аналіз.

Адзначаныя праблемы канкрэтныя. Аўтараў праінфармавалі пра неадпаведнасці

паміж крытэрыямі адбору ўдзельнікаў, апісанымі ў рукапісе, і тым, як гэтыя крытэрыі былі рэалізаваныя ў апублікаваным кодзе аналізу; акрамя таго, у адкрыты набор даных праз памылку пры зліцці кода трапілі лішнія радкі. Праз гэта частку апублікаваных лічбаў было цяжка ўзнавіць з даступных матэрыялаў. Аўтары перадалі *Science* выпраўлены аналітычны канвеер і абноўленыя вынікі, якія, паводле іх слоў, супадаюць з зыходнымі **па кірунку, статыстычнай значнасці і змястоўнай велічыні эфекту**. *Science* іх ацэньвае. Пакуль гэтая ацэнка не завершана, дакладныя лічбы застаюцца пад пытаннем, якое можа зняць толькі часопіс.

Такім чынам, тут адначасова дзве гісторыі: інтрыгуючы вынік пра тое, ці могуць факты змяняць укаранёныя перакананні, і жывы прыклад таго, як навуковы запіс адкрыта выпраўляе сам сябе. Задача гэтага тэксту — не змешваць іх.



У даследаванні паведамлялася, што трохраўндвая размова з ШІ, які абвяргаў уласныя аргументы ўдзельніка ў падтрымку выбранай ім тэорыі змовы, у сярэднім памяншала веру ў яе прыкладна на 20%; у адрозненне ад кантрольнай размовы на староннюю тэму, зніжэнне ўсё яшчэ вымяралася праз 10 дзён і 2 месяцы. Эфект быў працяглым, але частковым: большасць удзельнікаў пасля гэтага ўсё роўна верылі ў змову — іх вера аслабла, а не знікла. Цяпер на артыкуле стаіць Editorial Expression of Concern (*Science*, чэрвень 2026 года) праз неадпаведнасці ў справаздачнасці і памылку ў адкрытым наборе даных; аўтары сцвярджаюць, што выпраўлены аналіз захоўвае кірунак, значнасць і велічыню эфекту, пакуль *Science* працягвае праверку. Гэты графік The Clean Paper рэканструяваў з таго ж адкрытага набору даных, таму паказаныя значэнні з'яўляюцца папярэднімі, а не канчатковымі лічбамі артыкула.

Што зрабілі аўтары

- Правялі два эксперыменты з 2190 удзельнікамі, кожны з якіх сваімі словамі называў тэорыю змовы, у якую верыў, і доказы, якія лічыў яе пацвярджэннем.
- Кожны ўдзельнік правёў трохраўндавую размову з GPT-4 Turbo. У **эксперыментальнай** умове ІІІ атрымліваў інструкцыю абвяргаць канкрэтныя доказы ўдзельніка; у **кантрольнай** — абмяркоўваў староннюю тэму.
- Вымералі веру ў выбраную тэорыю змовы да размовы і адразу пасля яе, а затым паўторна звязаліся з удзельнікамі праз **10 дзён і 2 месяцы**.
- Папрасілі прафесійнага фактчэкара ацаніць дакладнасць усіх 128 фактычных сцвярджэнняў, зробленых ІІІ ў падвыбарцы размоў.
- Праверылі, ці пераносіцца эффект на іншыя, не звязаныя тэорыі змовы, а як тэст спецыфічнасці — ці зніжае ён таксама веру ў змовы, якія ў рэчаіснасці аказаліся **праўдзівымі**.

Што яны выявілі

- У **сярэдным прыкладна 20%-ае зніжэнне** веры ў выбраную тэорыю змовы ў эксперыментальнай групе ў параўнанні з кантролем (даследаванне 1: 95%-ы даверны інтэрвал [13.8; 19.7] пункта па 100-бальнай шкале, $P < 0.001$).
- **Эфект захоўваўся**. У эксперыментальнай групе падзенне не знікала: праз 10 дзён і два месяцы вера заставалася прыкладна на ўзроўні адразу пасля размовы, тады як кантрольная група ўвесь час была вышэй. Артыкул апісвае эффект для выбранай тэорыі як такі, што захоўваўся без прыкметнага аслаблення як мінімум два месяцы, а не як кароткі момант сумневу.
- **Ён абагульняўся** на шырокі спектр змоў, якія называлі людзі, і захоўваўся нават у тых, чыя зыходная вера была моцнай.
- **Сцвярджэнні ІІІ былі дакладнымі** ў гэтых умовах: 127 з 128 (99.2%) ацэненыя як праўдзівыя, 1 (0.8%) — як такі, што ўводзіць у зман, ніводнага — як ілжывага.
- **Эфект быў спецыфічным**, а не проста спараджаў агульны скепсіс: умяшанне **не** паменшыла веру ў праўдзівыя змовы, што паказвае на аслабленне менавіта непацверджаных перакананняў, а не на ператварэнне людзей у цынікаў адносна ўсяго.
- **Быў і ўмераны перанос**: вера ў іншыя, не звязаныя тэорыі змовы знізілася прыкладна на 12%, а ўдзельнікі паведамлялі пра больш моцны намер прырэчыць канспіралагічным сцвярджэнням. Асноўны эффект захаваўся ў даследаванні 2 і меў той жа кірунак у меншай выбарцы без кантрольнай групы — слабейшы тып доказу, але ўзгоднены з астатнімі данымі.

Чаго гэта не даказвае

- Цяпер вынік **нельга лічыць бяспрэчным**. На артыкуле стаіць Editorial Expression of Concern праз праблемы з апрацоўкай даных і ўзнаўляльнасцю, а выпраўленыя лічбы ўсё яшчэ ацэньвае *Science*. Канкрэтныя значэнні варта лічыць папярэднімі, пакуль праверка не завершана.
- Гэта **не** паказвае, што ІІІ «лечыць» веру ў тэорыі змовы. Сярэдняе зніжэнне прыкладна на 20% — змястоўны зрух, але не сціранне пераканання; большасць удзельнікаў пасля эксперыменту ўсё яшчэ верылі ў тэорыю, проста слабей.
- Гэта **не** вынік рэальнага масавага прымянення. Удзельнікі добраахвотна ўступілі ў даследаванне і добрасумленна ўзаемадзейнічалі з ІІІ; у рэальным свеце людзі, найбольш глыбока перакананыя ў тэорыі змовы, могуць якраз найменш хацець звяртацца да чатбота, які будзе з імі спрачацца.
- Дакладнасць 99.2% **адносіцца менавіта да гэтай задачы, мадэлі і стараннага прамптынг**у — гэта не агульная гарантыя праўдзівасці моўных мадэляў. Аўтары найпрост адзначаюць, што тая ж персаналізаваная сіла пераканання магла б штурхаць людзей і да *ілжывых* перакананняў, калі накіраваць яе ў той бок.
- «Працяглы» тут азначае **два месяцы** — гэта ўражвае для адной размовы, але не азначае назаўжды.

Наколькі моцныя доказы

- **Дызайн — сапраўдны моцны бок**. Кантраляваныя эксперыменты з treatment і control, чысты плацэбападобны кантроль, паўтарэнне ў двух даследаваннях і незалежнай выбарцы, адкладзеныя вымярэнні працягласці эфекту і асобная праверка дакладнасці сцвярджэнняў ІІІ — усё гэта старанней за большасць даследаванняў пераканання, а кірунак эфекту быў узгоднены ўсюды, дзе яго правяралі.
- **Але Expression of Concern — не заўвага дробным шрыфтам**. Магчымасць узнавіць апублікаваныя лічбы з адкрытых даных і кода — частка даверу да выніку, і менавіта тут адбылася паломка: неадпаведнасці ў крытэрыях скрынінгу і дубляваныя радкі праз памылку пры зліцці кода. Заява аўтараў, што выпраўлены канвеер захоўвае вынік, абнадзейвае і выглядае праўдападобна, але гэта іх уласная ацэнка, а не незалежная канчатковая выснова. Вырашальнае значэнне будзе мець ацэнка *Science*.
- **Сумленны статус — «пазначана для праверкі», а не «абвергнута» або «пацверджана»**. Гэта добра пабудаванае, яркае даследаванне, дакладныя лічбы якога цяпер фармальна пераглядаюць. Нязручнае месца, каб пакінуць прыгожую гісторыю, але менавіта яно цяпер дакладнае.

Чаму гэта важна

Гадамі ўплывовай была ідэя, што прыхільнікамі тэорый змовы кіруюць псіхалагічныя патрэбы і ідэнтычнасць, таму фактамі іх пераканаць немагчыма. Працяглае зніжэнне веры на аснове доказаў — *калі вынік вытрымае праверку* — з'яўляецца важнай процівагай такому песімізму. Яно мяркуе, што частка праблемы была ў тым, што агульныя абвяржэнні ніколі не працавалі менавіта з *канкрэтнымі* доказамі, на якія спасылаецца канкрэтны прыхільнік змовы; LLM здольная рабіць гэта маштабавана.

Гэта таксама важны прыклад таго, як павінна працаваць навука. Урок Expression of Concern не ў тым, што гучныя вынікі нічога не вартыя, а ў тым, што памылкі ўсплываюць, даныя пераглядаюцца, а сцвярджэнні публічна пераправяраюцца. Праца гэтага механізму — прыкмета сістэмы, а не скандал; менавіта таму правільная пазіцыя адносна гэтага выніку цяпер — цяргненне, а не апладысменты або адмахванне.

І для III гэтага гісторыя мае два бакі. Тая ж здольнасць аслабіць ілжывае перакананне персаналізаваным аргументам можа гэтак жа эфектыўна ўзмацніць іншае ілжывае перакананне. Тэхніка нейтральная; мэта — не.

Кароткі вынік

Даследаванне 2024 года паказала, што трохраўндая размова з GPT-4 Turbo памяншала веру людзей у тэорыю змовы, якой яны прытрымліваліся, прыкладна на 20%; эфект захоўваўся два месяцы, назіраўся для розных тыпаў змоў, а фактычныя сцвярджэнні III былі ацэненыя як амаль цалкам дакладныя. У чэрвені 2026 года да артыкула было дададзена Editorial Expression of Concern праз праблемы з апрацоўкай даных і ўзнаўляльнасцю; аўтары паведамляюць, што выпраўлены аналіз захоўвае кірунак, значнасць і велічыню выніку, а *Science* усё яшчэ праводзіць ацэнку. Чытаць гэта трэба як сапраўды цікавае і старанна пабудаванае даследаванне, якое цяпер пераглядаюць, — не як устаноўленую ісціну і не як абвергнутую працу. Самы сумленны сказ пра яго цяпер піша сам навуковы працэс: праверыць яшчэ раз.

Крыніцы

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>