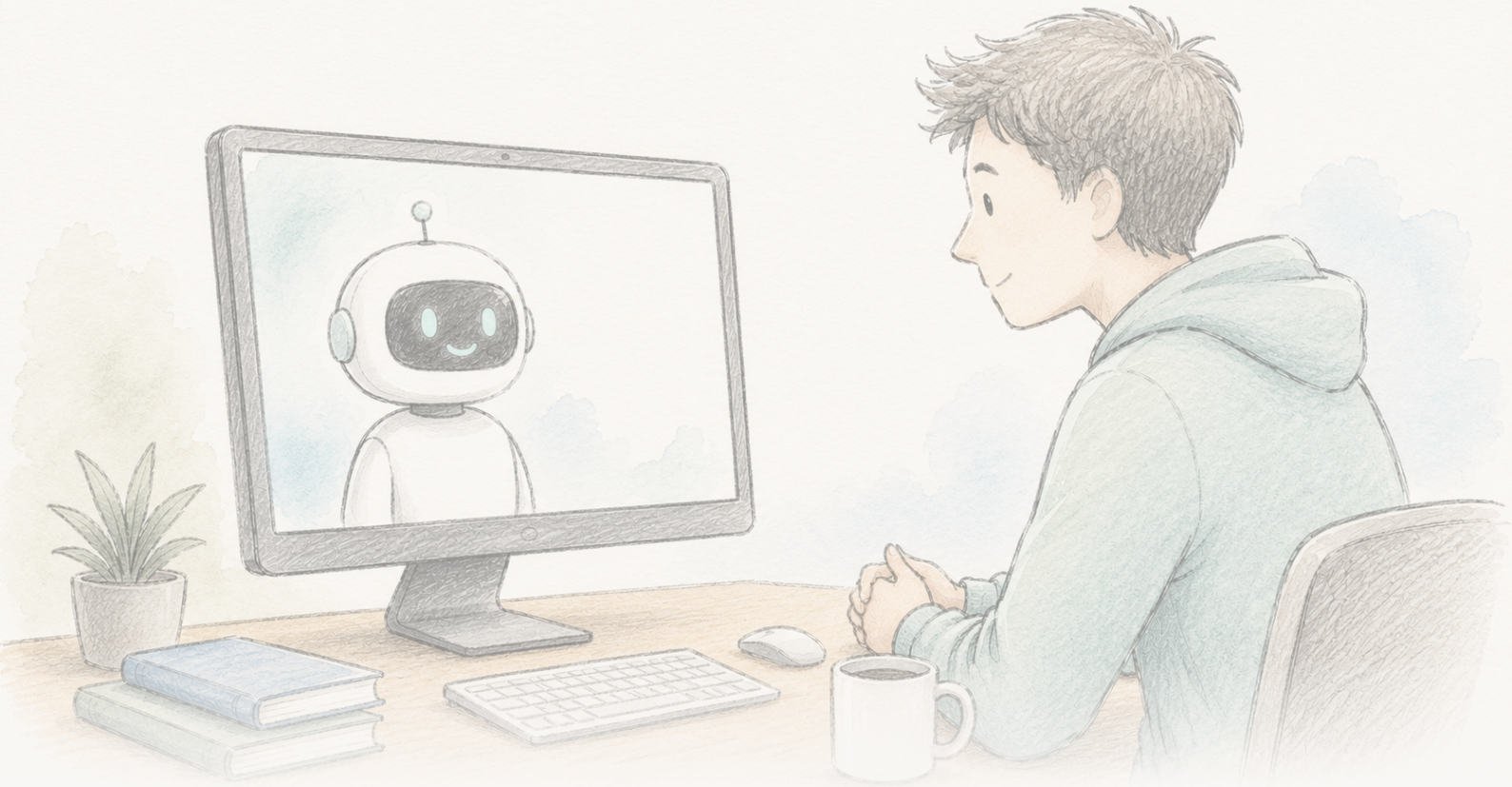


The Clean Paper

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

Чатбот изгледа го намалил верувањето во теории на заговор со месеци

Студија од 2024 година објави дека разговор во три рунди со GPT-4 Turbo го намалил верувањето на луѓето во теорија на заговор што ја поддржувале за околу 20%, ефект што траел два месеца и се појавувал кај различни типови заговори, при што тврдењата на AI биле оценети како речиси целосно точни. Во јуни 2026 година трудот добил Editorial Expression of Concern поради проблеми со обработката на податоците и репродуктивноста; авторите велат дека исправената анализа го задржува наодот по насока, значајност и големина, а Science сè уште оценува. Навистина интересен и внимателно изграден резултат — моментално под проверка, ниту конечен ниту побиен.

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 додаде Editorial Expression of Concern на овој труд додека ги оценува прашањата за обработката на податоците и репродуктивноста. Тоа не е повлекување на трудот ниту утврдување на недолично однесување; значи дека резултатот треба да се чита како привремен додека не заврши проценката.

Editorial Expression of Concern →

Фактите сепак имаат ефект — а трудот сега носи предупредување

Во 2024 година една студија објави резултат што се коси со удобен дел од вообичаената мудрост. Ако на човек му дадете краток разговор со AI во три рунди — GPT-4 Turbo, насочен да одговори токму на доказите што самиот човек ги наведува за теорија на заговор во која верува — во просек верувањето во таа теорија се намалува за околу 20%. Намалувањето било присутно и два месеца подоцна. Ефектот се појавил кај класични теории на заговор (атентатот врз Џон Ф. Кенеди, вонземјани, Илуминати) и кај понови теми (COVID-19, изборите во САД во 2020 година), дури и кај луѓе чие верување било силно и поврзано со нивниот идентитет. Кога професионален проверувач на факти оценил 128 тврдења што ги дал AI, 127 биле точни, едно било заведувачко, а ниту едно не било лажно.

Насловот што најмногу се прошири беше дека *може* да се разговара со луѓето така што ќе излезат од „зајачката дупка“ — дека луѓето што веруваат во заговори не се недостижни за докази, туку можеби им требаат вистинските докази, доволно конкретно насочени. Тоа е навистина интересно тврдење, а студијата била изградена повнимателно од голем дел од истражувањата за убедување.

Потоа, во јуни 2026 година, *Science* објави **Editorial Expression of Concern** за трудот. Токму овој дел најлесно исчезнува во прераскажувањата, а во моментов токму тој одредува колкава тежина може да носи резултатот.

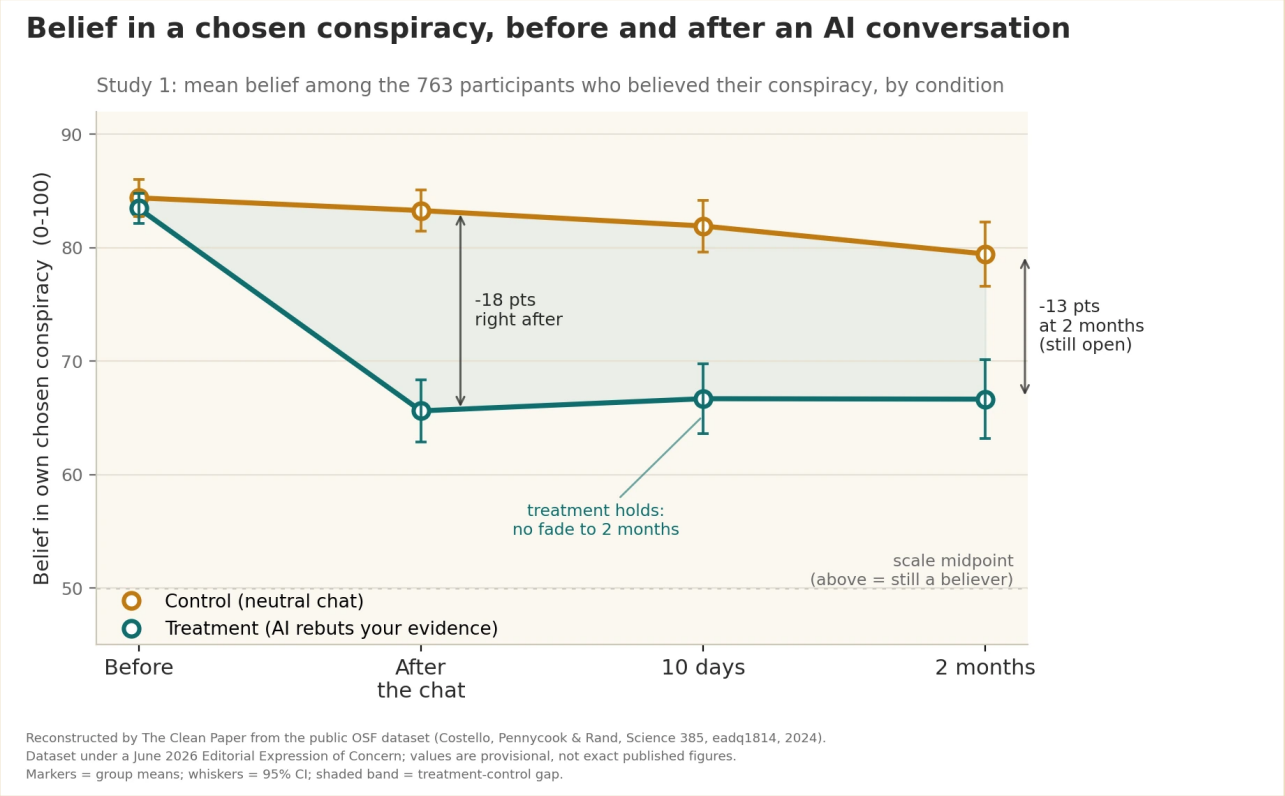
Што е Expression of Concern — и што не е

Editorial Expression of Concern е формално предупредување што списанието го додава кон труд за да им каже на читателите дека се поставени прашања што сè уште се истражуваат. Тоа **не** е повлекување на трудот и само по себе не е утврдување на недолично научно однесување. Значи: читајте го резултатот со оваа резерва, затоа што научниот запис може да се промени. Во овој случај, откако биле известени за проблемите, авторите ги истражиле, му ги пријавиле деталите на *Science* и доставиле исправена анализа.

Проблемите што биле означени се конкретни. Авторите биле известени за недоследности меѓу ракописот и објавениот аналитички код во начинот на кој биле применети

критериумите за селекција на учесниците, како и дека јавната база содржела дополнителни редови што биле вметнати по грешка при спојување код — проблеми поради кои дел од пријавените бројки тешко се репродуцирале од објавените материјали. Авторите му доставиле на *Science* исправен аналитички процес и ажурирани резултати, за кои велат дека се совпаѓаат со првичните **по насока, статистичка значајност и суштинска големина на ефектот**. *Science* сè уште ги оценува. Додека не заврши таа проценка, точните бројки остануваат привремени.

Затоа тука има две приказни одеднаш: интригантен резултат за тоа дали фактите можат да придвижат цврсто вкоренети верувања, и жив пример како научниот запис јавно се коригира. Целта е двете да не се измешаат.



Во студијата, AI-разговор во три рунди што директно ги оспорувал доказите наведени од самиот учесник бил пријавен како интервенција што во просек го намалила верувањето во избраната теорија на заговор за околу 20%; за разлика од контролната дискусија на неповрзана тема, намалувањето било мерливо и по 10 дена и по 2 месеца. Пријавениот ефект бил траен, но делумен: повеќето учесници и потоа верувале во заговорот — станува збор за намалување, не за „лек“. Трудот моментално е под Editorial Expression of Concern (*Science*, јуни 2026) поради недоследности во известувањето и грешка во јавната база; авторите велат дека исправената анализа ја задржува насоката, значајноста и големината на ефектот, додека *Science* продолжува со оценувањето. Овој график е реконструиран од The Clean Paper од јавната база, па прикажаните вредности се привремени, не конечни бројки на трудот.

Original figure — The Clean Paper CC BY 4.0

Што направиле авторите

- Спровеле два експеримента со 2190 учесници, при што секој со свои зборови именувал теорија на заговор во која верува и доказите за кои сметал дека ја поддржуваат.
- Секој учесник водел разговор во три рунди со GPT-4 Turbo. Во **третманската** група AI бил насочен да ги оспорува конкретните докази на учесникот; во **контролната** група разговарал за неповрзана тема.
- Го измериле верувањето во избраната теорија пред разговорот и веднаш по него, а потоа повторно ги контактирале учесниците по **10 дена и по 2 месеца**.
- Професионален проверувач на факти ја оценил точноста на сите 128 фактички тврдења што AI ги дал во избран примерок.
- Провериле дали ефектот се пренесува и на други, неповрзани теории на заговор — и, како тест за специфичност, дали ја намалува вербата и во заговори што навистина се **вистинити**.

Што откриле

- **Просечно намалување од приближно 20%** на верувањето во избраната теорија во третманската група во однос на контролата (студија 1: 95% интервал на доверба [13,8; 19,7] поени на скала од 100 поени, $P < 0{,}001$).
- **Ефектот траел.** Во третманската група падот не исчезнал: верувањето по 10 дена и два месеца останало близу до нивото веднаш по разговорот, наместо постепено да се враќа, додека контролата останувала повисока. Трудот го опишува ефектот врз избраната теорија како непроменето присутен најмалку два месеца, не како краткотрајно колебање.
- **Се појавувал кај различни теории** што учесниците ги наведувале, вклучително и кај луѓе со првично силно верување.
- **Тврдењата на AI биле точни** во оваа поставеност: 127 од 128 (99,2%) биле оценети како вистинити, 1 (0,8%) како заведувачко, ниту едно како лажно.
- **Ефектот бил специфичен**, а не општа недоверба: интервенцијата **не** го намалила верувањето во вистинити заговори, што укажува дека се придвижувале неподдржаните верувања, а не дека луѓето станале цинични кон сè.
- **Имало умерено прелевање** — верувањето во други, неповрзани заговори паднало за околу 12%, а учесниците пријавиле поголема намера да се спротивставуваат на тврдења за заговор. Главниот ефект се повторил во студијата 2 и се движел во иста насока во помал примерок без контролна група (послаб доказ, но согласен со останатото).

Што ова не докажува

- Трудот во моментов **не е без отворени прашања**. Тој е под Editorial Expression of Concern поради проблеми со обработката на податоците и репродуктивноста, а исправените бројки сè уште ги оценува *Science*. Конкретните вредности треба да се сметаат за привремени додека не заврши проценката.
- **Не** покажува дека AI ја „лекува“ вербата во заговори. Просечно намалување од ~20% е значајна промена, не исчезнување — повеќето учесници и понатаму верувале, само помалку.
- Ова **не е резултат од реална примена**. Учесниците доброволно се вклучиле во студија и разговарале со AI во добра вера; во реалниот свет луѓето најцврсто врзани за заговор најмалку веројатно сами ќе побараат чатбот што ќе им се спротивстави.
- Точноста од 99,2% е **специфична за оваа задача, овој модел и внимателното насочување** — не е општа гаранција дека јазичните модели кажуваат вистина. Авторите експлицитно предупредуваат дека истата персонализирана моќ на убедување може да турка и *лажни* верувања ако е насочена во таа насока.
- „Трајно“ тука значи **два месеца**, што е впечатливо за еден разговор, но не значи трајно засекогаш.

Колку се убедливи доказите?

- **Дизајнот е вистинска силна страна.** Контролирани експерименти третман-наспроти-контрола, чиста контрола слична на плацебо, повторување во две студии и независен примерок, последователни мерења на трајноста и експлицитна проверка на точноста на тврдењата на AI — тоа е повнимателно од многу истражувања за убедување, а насоката на ефектот била согласна каде и да била тестирана.
- **Но Expression of Concern не е фуснота.** Можноста повторно да се добијат пријавените бројки од објавените податоци и код е дел од тоа што го прави резултатот доверлив, а токму таму се појавил проблемот — недоследности во критериумите за селекција и дуплирани редови по грешка при спојување код. Изјавата на авторите дека исправениот процес го задржува резултатот е охрабрувачка и веродостојна, но засега е нивна изјава, не независна пресуда. Треба да се почека проценката на *Science*.
- **Најточниот статус е „означено за проверка“, не „побиено“ или „потврдено“.** Добро осмислена и впечатлива студија чии точни бројки се под формална ревизија. Непријатно е добра приказна да се остави во таква состојба, но тоа е точната состојба.

Зошто е важно

Со години влијателно гледиште тврдеше дека луѓето што веруваат во заговори се водени од психолошки потреби и идентитет и затоа не можат да бидат разубедени со факти. Трајно намалување базирано на докази — ако се потврди — е значаен противтежок на тој песимизам: укажува дека дел од проблемот можеби бил што општото побивање никогаш не се занимавало со *конкретните* докази што верникот лично ги наведува, нешто што LLM може да го прави во голем обем.

Важно е и како студија на случај за тоа како треба да работи науката. Поуката од Expression of Concern не е дека славните резултати се безвредни; туку дека грешките излегуваат на виделина, податоците повторно се проверуваат, а тврдењата јавно се преоценуваат. Тоа што овој механизам работи е предност, не скандал — и затоа правилниот став кон резултатот е трпение, а не ниту аплауз ниту отфрлање.

И има две острици кога станува збор за AI. Истата способност што со прилагоден аргумент може да ослаби лажно верување може исто толку ефикасно и да го засили. Техниката е неутрална; целта не е.

Кратко резиме

Студија од 2024 година објави дека разговор во три рунди со GPT-4 Turbo го намалил верувањето на луѓето во теорија на заговор што самите ја поддржувале за околу 20%, ефект што опстојувал два месеца и се појавувал кај различни типови заговори, при што фактичките тврдења на AI биле оценети како речиси целосно точни. Во јуни 2026 година трудот добил Editorial Expression of Concern поради проблеми со обработката на податоците и репродуктивноста; авторите велат дека исправената анализа го задржува резултатот по насока, значајност и големина, а *Science* сè уште оценува. Најточно е да се чита како навистина интересен и внимателно изграден резултат што моментално е под проверка — ниту конечен, ниту побиен. Најчесната реченица е онаа што самиот процес ја бара: проверете повторно.

Извори

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