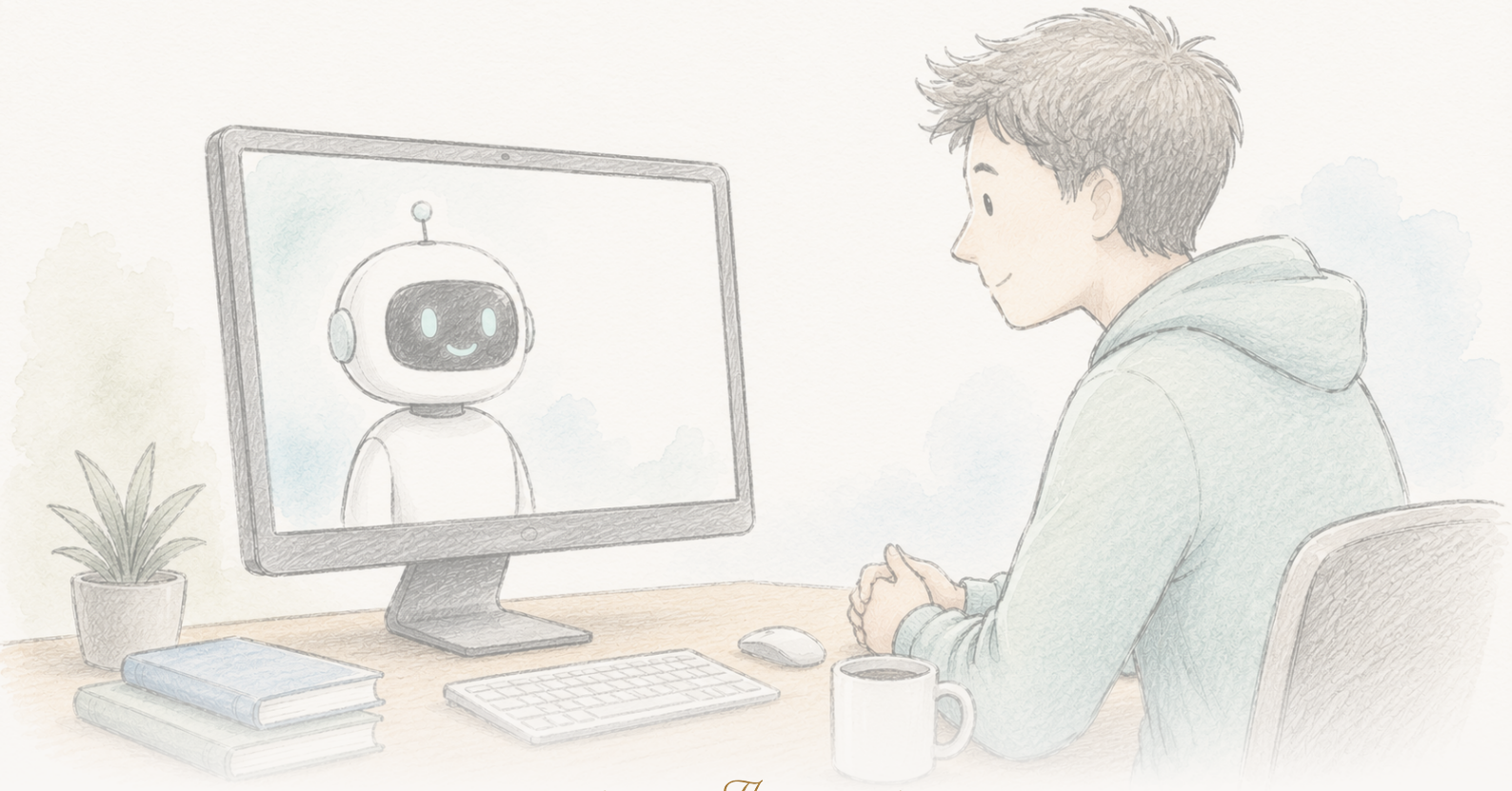


The Clean Paper

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

Чатбот изглежда намалява вярата в конспиративни теории за месеци

Проучване от 2024 г. установява, че разговор в три кръга с GPT-4 Turbo намалява с около 20% вярата на хората в конспиративна теория, която те приемат, като ефектът се запазва два месеца, наблюдава се при различни видове конспирации и се основава на твърдения на AI, оценени като преобладаващо точни. През юни 2026 г. към статията е добавено Editorial Expression of Concern заради проблеми с обработката на данни и възпроизводимостта; авторите казват, че коригиран анализ запазва посоката, статистическата значимост и размера на резултата, а Science все още го оценява. Наистина интересен и внимателно построен резултат — в момента под преглед, нито окончателно установен, нито опроверган.

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 е добавил Editorial Expression of Concern към тази статия, докато оценява въпроси относно обработката на данните и възпроизводимостта. Това не е оттегляне на статията и не е установяване на нарушение; означава, че докладваният резултат трябва да се разглежда като предварителен, докато проверката приключи.

Editorial Expression of Concern →

В крайна сметка — факти, но и предупреждение върху статията

През 2024 г. едно изследване съобщи нещо, което върви срещу удобно разпространено убеждение. Дайте на човек кратък разговор в три кръга с AI — GPT-4 Turbo, инструктиран да отговори конкретно на доказателствата, които човекът посочва за конспиративна теория, в която лично вярва — и средно вярата му в тази теория спада с около 20%. Спадът все още се вижда два месеца по-късно. Ефектът се наблюдава както при класически конспирации (убийството на John F. Kennedy, извънземни, илюминатите), така и при актуални (COVID-19, изборите в САЩ през 2020 г.), дори сред хора, чието убеждение е силно и свързано с идентичността им. Когато професионален проверител на факти оценява 128 твърдения на AI, 127 са верни, едно е подвеждащо, а нито едно не е невярно.

Заглавието, което се разпространи, беше, че *можете* да изведете хората от конспиративната „заешка дупка“ с разговор — че вярващите в конспирации не са недостижими за доказателствата, просто се нуждаят от правилните доказателства, поднесени достатъчно конкретно. Това е действително интересна теза и изследването е построено по-внимателно от повечето работи върху убеждаването.

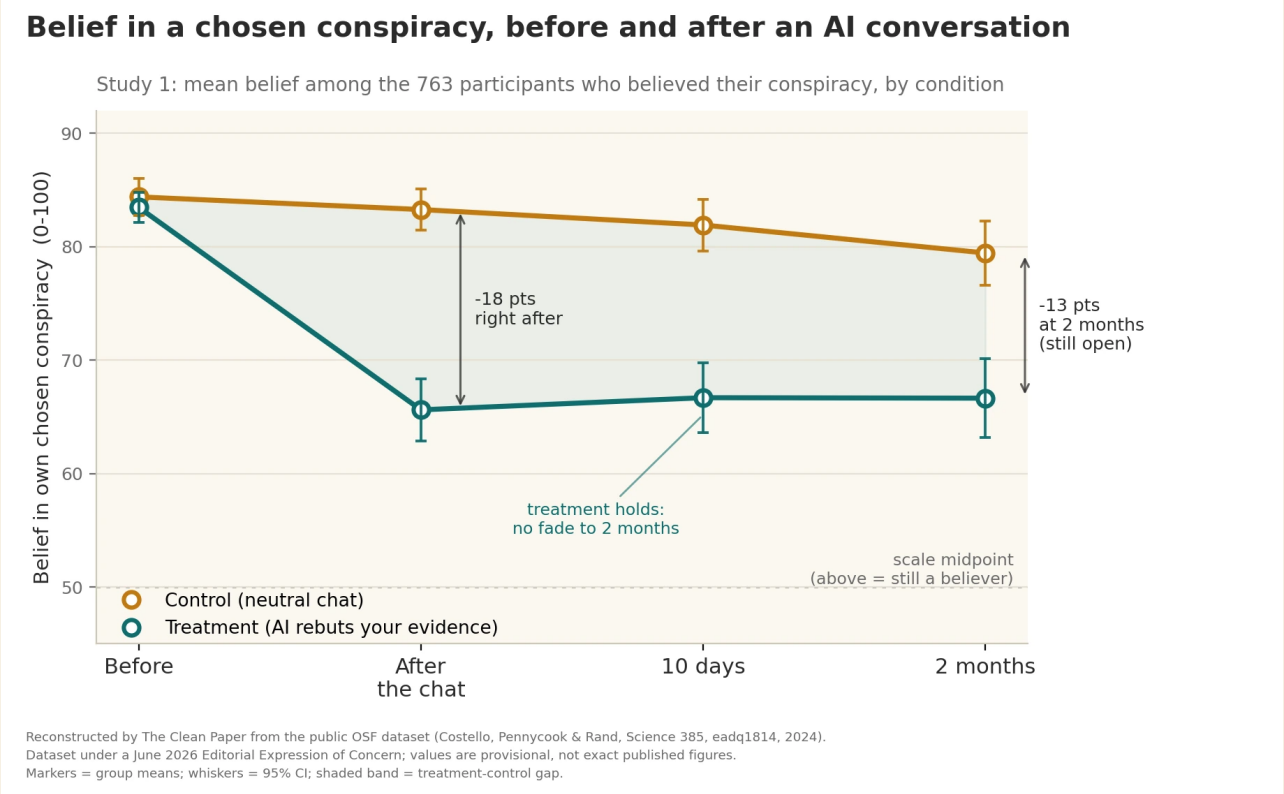
После, през юни 2026 г., *Science* публикува към статията **Editorial Expression of Concern**. Именно тази част повечето преразкази вероятно ще пропуснат, а точно тя определя колко тежест може да носи резултатът в момента.

Какво е Expression of Concern — и какво не е

Editorial Expression of Concern е официален предупредителен знак, който списание прикрепя към статия, за да уведоми читателите, че са повдигнати въпроси, докато те все още се изясняват. Това **не е** оттегляне (статията не е изтеглена) и само по себе си не е установяване на научно нарушение. Означава: четете с тази уговорка, защото научният запис може да се промени. Тук, след като са били уведомени за проблемите, авторите са извършили проверка, докладвали са подробностите на Science и са предоставили коригиран анализ.

Конкретно е отбелязано следното. Авторите са уведомени за несъответствия в начина, по който критериите за скрининг на участниците са приложени между ръкописа и публикувания код за анализ, както и че публичният набор от данни съдържа излишни редове, вмъкнати поради грешка при сливане на код — проблеми, които правят някои от докладваните числа трудни за възпроизвеждане от публикуваните материали. Авторите предоставят на *Science* коригиран анализ и набор обновени резултати, които според тях съвпадат с оригиналните **по посока, статистическа значимост и съществен размер**. *Science* ги оценява. Докато тази оценка не приключи, точните числа остават под въпросителна, която само списанието може да премахне.

Така тук има две истории едновременно: интригуващ резултат за това дали фактите могат да променят вкоренени убеждения и жив пример как научният запис се поправя открито. Целта на този текст е да не ги смесва.



В изследването е съобщено, че разговор в три кръга с AI, който опровергава собствените посочени от участника доказателства, намалява средно с около 20% вярата в избраната от него конспирация; за разлика от контролен разговор по несвързана тема, спадът все още се измерва след 10 дни и след 2 месеца. Докладваният ефект е траен, но частичен: повечето участници продължават да вярват в конспирацията след това — намаление, не „лечение“. В момента статията е под Editorial Expression of Concern (*Science*, юни 2026) заради несъответствия в докладването и грешка в публичния набор от данни; авторите казват, че коригиран анализ запазва посоката, значимостта и размера на ефекта, докато *Science* продължава оценката. Тази графика е реконструирана от The Clean Paper от публичния набор от данни, затова показаните стойности са предварителни, а не окончателните числа на статията.

Original figure — The Clean Paper CC BY 4.0

Какво са направили авторите

- Провеждат два експеримента с 2190 участници, всеки от които назовава — със свои думи — конспиративна теория, в която вярва, и доказателствата, които според него я подкрепят.
- Всеки участник води разговор в три кръга с GPT-4 Turbo. В **експерименталното** условие AI е инструктиран да опровергае конкретните доказателства на участника; в **контролното** условие обсъжда несвързана тема.
- Измерват вярата в избраната конспирация преди и веднага след разговора, след което отново се свързват с участниците след **10 дни и 2 месеца**.
- Професионален проверител на факти оценява точността на всички 128 фактически твърдения, които AI прави в извадка от разговорите.
- Проверяват дали ефектът се пренася към други, несвързани конспирации — и като тест за специфичност дали намалява и вярата в конспирации, които се оказват **верни**.

Какво са установили

- **Средно намаление с приблизително 20%** във вярата в избраната конспирация в експерименталната група спрямо контролната (изследване 1: 95% доверителен интервал [13.8, 19.7] пункта по 100-точкова скала, $P < 0.001$).
- **Ефектът се запазва.** В експерименталната група спадът не избледнява: след 10 дни и два месеца убеждението остава близо до нивото веднага след разговора, вместо да се връща нагоре, докато контролата остава по-висока през целия период — статията описва ефекта върху избраната конспирация като запазен без отслабване поне два месеца, а не моментно колебание.
- **Ефектът се обобщава** върху различните конспирации, които хората назовават, и се запазва дори при участници с първоначално силно убеждение.
- **Твърденията на AI са точни** в тази среда: 127 от 128 (99,2%) са оценени като верни, 1 (0,8%) като подвеждащо, нито едно като невярно.
- **Ефектът е специфичен**, а не общо съмнение: интервенцията **не** намалява вярата във верни конспирации, което подсказва, че измества неподкрепени убеждения, вместо да прави хората цинични към всичко.
- **Има умерен пренос** — вярата в други, несвързани конспирации спада с около 12%, а участниците съобщават по-голямо намерение да се противопоставят на конспиративни твърдения. Основният ефект се запазва в изследване 2 и сочи в същата посока в по-малка извадка без контролна група (по-слабо доказателство, но съвместимо).

Какво не доказва това

- В момента резултатът **не стои без въпросителни**. Статията е под Editorial Expression of Concern заради проблеми с обработката на данните и възпроизводимостта, а коригираните числа все още се оценяват от *Science*. Конкретните стойности трябва да се разглеждат като предварителни, докато този преглед приключи.
- То **не** показва, че AI „лекува“ вярата в конспирации. Средно намаление от ~20% е смислено изместване, не заличаване — повечето участници продължават да вярват в теорията след това, просто по-малко.
- Това **не е резултат от внедряване в реалния свят**. Участниците доброволно са се включили в изследване и са общували с AI добросъвестно; извън лабораторията най-убедените в конспирация вероятно са и най-малко склонни да потърсят чатбот, който ще спори с тях.
- Точността от 99,2% е **специфична за тази задача, този модел и внимателното формулиране на подканите (prompting)** — не обща гаранция, че езиковите модели казват истината. Авторите изрично отбелязват, че същата персонализирана убедителна сила може да насърчава *неверни* убеждения, ако бъде насочена натам.
- „Траен“ тук означава **два месеца**, което е впечатляващо за един разговор, но не е същото като постоянен ефект.

Колко силни са доказателствата

- **Дизайнът е истинска силна страна.** Контролирани експерименти лечение-срещу-контрола, чиста плацебоподобна контролна задача, репликация в две изследвания и независима извадка, проследяване на трайността и изрична проверка на точността на собствените твърдения на AI — това е по-внимателно от повечето изследвания на убеждаването, а посоката на ефекта е последователна навсякъде, където е проверена.
- **Но Expression of Concern не е бележка под линия.** Възможността да се възпроизведат докладваните числа от публикуваните данни и код е част от това, което прави един резултат надежден, и точно там има проблем — несъответствия в критериите за скрининг и дублирани редове от грешка при сливане на код. Твърдението на авторите, че коригираният анализ запазва резултата, е окуражаващо и правдоподобно, но е тяхна собствена оценка, не независима присъда. Оценката на *Science* е това, което трябва да се изчака.
- **Честният статус е „отбелязано за проверка“, не „опровергано“ или „потвърдено“.** Добре построено, впечатляващо изследване, чиито точни числа са под официален преглед. Това е неудобно място, на което да оставим добра история, и точно затова е вярното място.

Защо е важно

В продължение на години влиятелна гледна точка твърди, че вярващите в конспирации са движени от психологически нужди и идентичност и затова не могат да бъдат разубедени с факти. Трайно, основано на доказателства намаление — ако издържи проверката — е значим противовес на този песимизъм: то подсказва, че част от проблема е, че общото опровергаване никога не се е занимавало със *специфичните* доказателства, които конкретният човек цитира, а LLM може да прави точно това в голям мащаб.

Резултатът е важен и като пример за това как би трябвало да работи науката. Урокът от Expression of Concern не е, че известните резултати са без стойност; а че грешките излизат на повърхността, данните се преглеждат отново и твърденията се проверяват повторно публично. Работата на този механизъм е особеност, не скандал — и затова правилната позиция към този резултат е търпение, а не нито аплодисменти, нито отхвърляне.

И за AI това реже в двете посоки. Същата способност да отслаби невярно убеждение с персонализиран аргумент може също толкова ефективно да го усили. Техниката е неутрална; целта не е.

Кратко обобщение

Проучване от 2024 г. установява, че разговор в три кръга с GPT-4 Turbo намалява с около 20% вярата на хората в конспиративна теория, която те приемат, като ефектът се запазва два месеца и се наблюдава при различни видове конспирации, а фактическите твърдения на AI са оценени като преобладаващо точни. През юни 2026 г. статията е поставена под Editorial Expression of Concern заради проблеми с обработката на данните и възпроизводимостта; авторите съобщават, че коригиран анализ запазва посоката, значимостта и размера на резултата, а *Science* все още оценява въпроса. Четете го като действително интересен и внимателно построен резултат, който в момента е под проверка — нито окончателно установен, нито опроверган. Най-честното изречение за него е това, което самият процес пише: проверете отново.

Източници

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