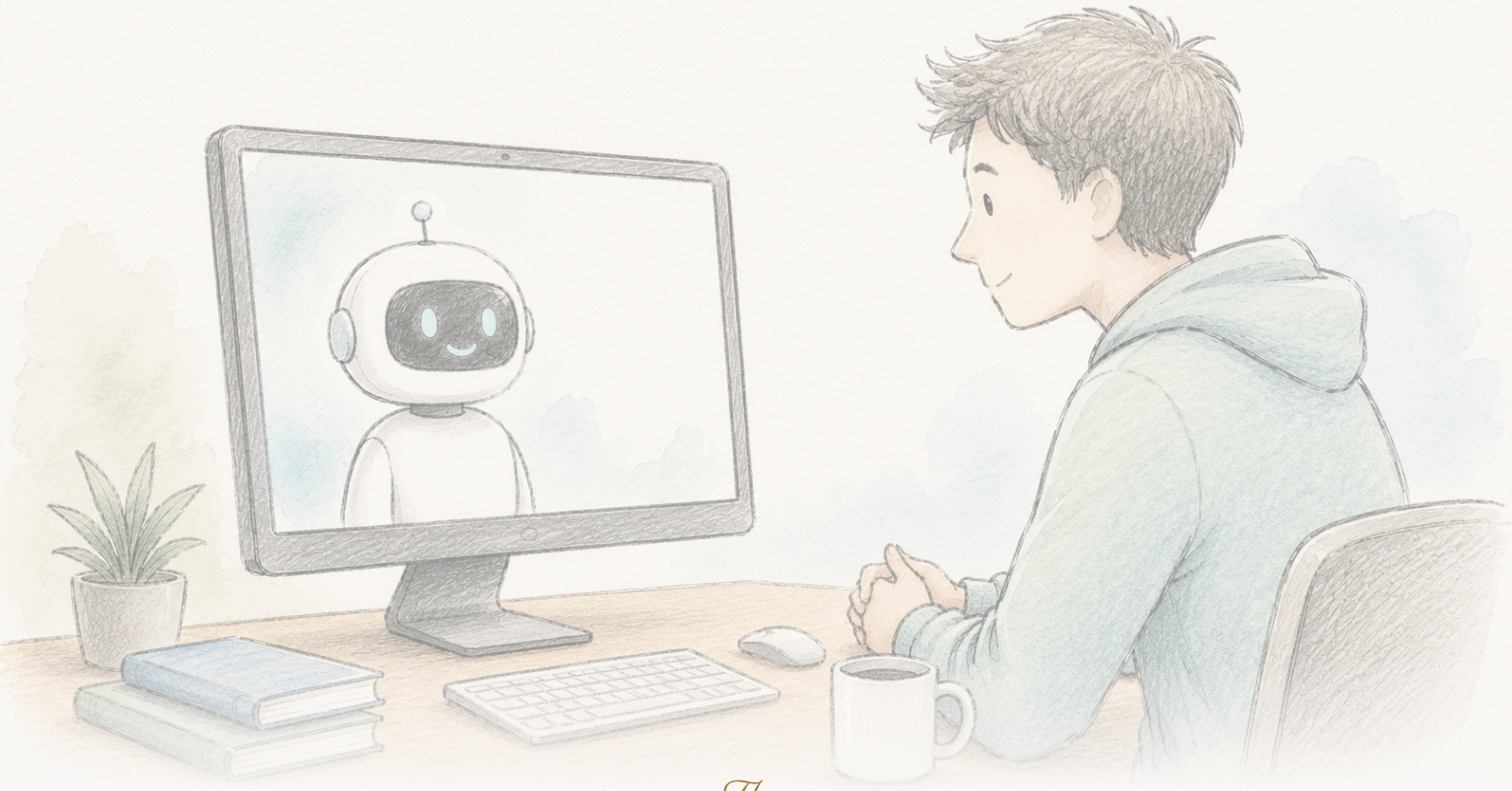


The Clean Paper

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

Похоже, разговор с чатботом на месяцы ослаблял веру в теории заговора

Исследование 2024 года показало, что трёхраундовый разговор с GPT-4 Turbo уменьшал веру людей в теорию заговора, которой они придерживались, примерно на 20%; эффект сохранялся два месяца, наблюдался для разных типов заговоров и опирался на утверждения ИИ, которые фактчекер оценил как почти полностью точные. В июне 2026 года к статье было добавлено Editorial Expression of Concern из-за проблем с обработкой данных и воспроизводимостью; авторы утверждают, что исправленный анализ сохраняет направление, значимость и величину результата, а Science всё ещё проводит оценку. Действительно интересное и тщательно построенное исследование, которое сейчас пересматривают: оно ещё не окончательно подтверждено, но и не опровергнуто.

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 добавил к этой статье Editorial Expression of Concern, пока оценивает вопросы, связанные с обработкой данных и воспроизводимостью. Это не ретракция и не вывод о нарушении; это означает, что опубликованный результат следует считать предварительным, пока проверка не завершена.

Editorial Expression of Concern →

В конце концов факты — и предупреждающая отметка на статье

В 2024 году исследование сообщило результат, который противоречит удобной части общепринятой мудрости. Дайте человеку короткий разговор с ИИ в три раунда — GPT-4 Turbo, настроенным отвечать именно на те доказательства, которые человек приводит в поддержку теории заговора, в которую сам верит, — и в среднем его вера в эту теорию уменьшается примерно на 20%. Снижение всё ещё наблюдалось через два месяца. Эффект работал и для классических заговоров (убийство John F. Kennedy, пришельцы, иллюминаты), и для актуальных (COVID-19, выборы в США 2020 года), даже среди людей с сильной верой, связанной с их идентичностью. Когда профессиональный фактчекер оценил 128 утверждений, сделанных ИИ, 127 оказались истинными, одно — вводящим в заблуждение, и ни одно — ложным.

Заголовок, разлетевшийся по миру, был таким: людей *можно* вытащить из «кроличьей норы» разговором — сторонники теорий заговора не недосягаемы для доказательств, им просто нужны правильные доказательства, поданные достаточно конкретно. Это действительно интересное утверждение, и само исследование было построено тщательнее, чем большинство работ об убеждениях.

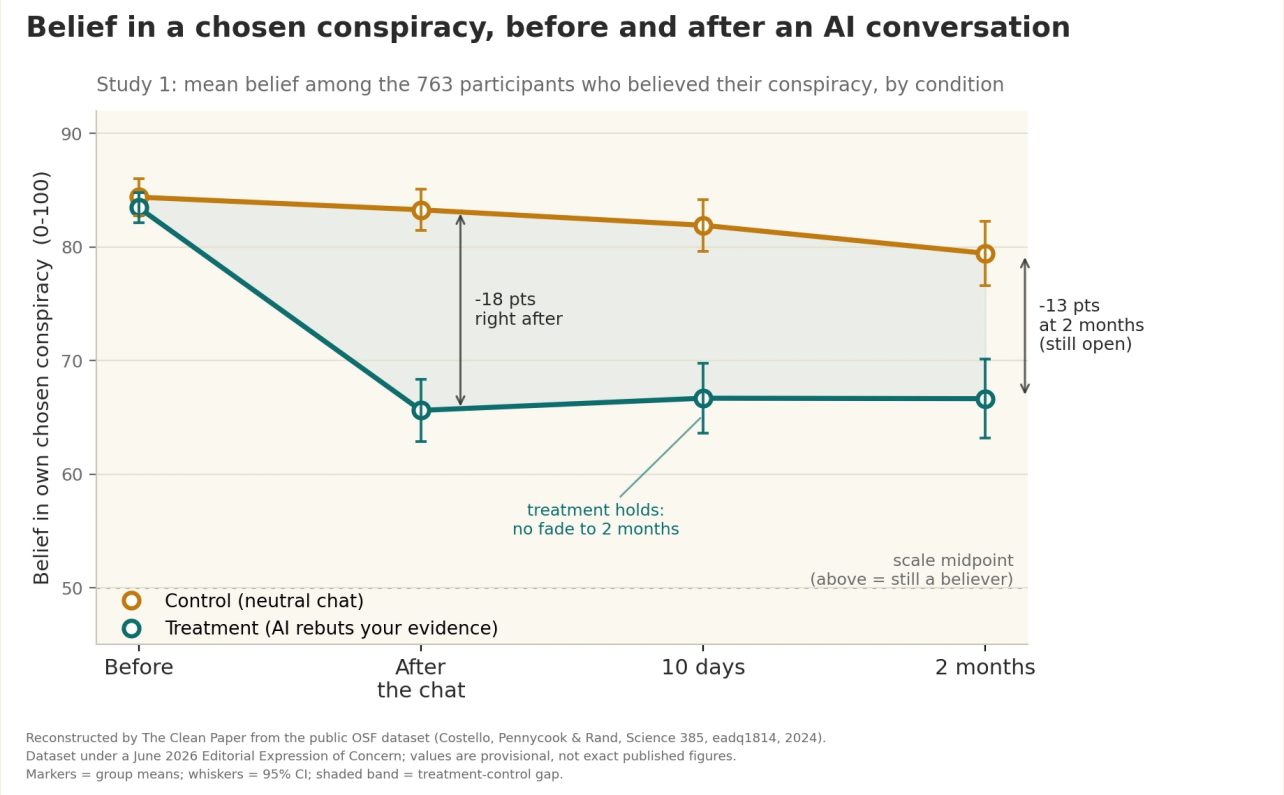
А затем, в июне 2026 года, *Science* опубликовал к статье **Editorial Expression of Concern** — редакционное уведомление о возникших вопросах. Именно эту часть большинство пересказов пропустит, хотя сейчас именно она определяет, какой вес может нести результат.

Что такое Expression of Concern — и чем оно не является

Editorial Expression of Concern — официальная отметка, которую журнал добавляет к статье, чтобы предупредить читателей: к работе возникли вопросы, и их всё ещё выясняют. Это **не** ретракция — статью не отозвали — и само по себе это не вывод о нарушении научной этики. Оно означает: читайте с оговоркой, потому что научная запись может измениться. В данном случае после того, как авторов уведомили о проблемах, они провели проверку, предоставили *Science* конкретные детали и передали исправленный анализ.

Отмеченные проблемы конкретны. Авторов проинформировали о несоответствиях между критериями отбора участников, описанными в рукописи, и тем, как эти критерии были реализованы в опубликованном коде анализа; кроме того, в открытый набор данных из-за ошибки при слиянии кода попали лишние строки. Из-за этого часть опубликованных чисел было трудно воспроизвести из доступных материалов. Авторы передали *Science* исправленный аналитический конвейер и обновлённые результаты, которые, по их словам, совпадают с исходными **по направлению, статистической значимости и содержательной величине эффекта**. *Science* их оценивает. Пока эта оценка не завершена, точные числа остаются под знаком вопроса, снять который может только журнал.

Таким образом, здесь одновременно две истории: интригующий результат о том, могут ли факты менять укоренившиеся убеждения, и живой пример того, как научная запись открыто исправляет сама себя. Задача этого текста — не смешивать их.



В исследовании сообщалось, что трёхраундовый разговор с ИИ, который опровергал собственные аргументы участника в поддержку выбранной им теории заговора, в среднем уменьшал веру в неё примерно на 20%; в отличие от контрольного разговора на постороннюю тему, снижение всё ещё измерялось через 10 дней и 2 месяца. Эффект был длительным, но частичным: большинство участников после этого всё равно верили в заговор — их вера ослабла, а не исчезла. Сейчас на статье стоит Editorial Expression of Concern (*Science*, июнь 2026 года) из-за несоответствий в отчётности и ошибки в открытом наборе данных; авторы утверждают, что исправленный анализ сохраняет направление, значимость и величину эффекта, пока *Science* продолжает проверку. Этот график The Clean Paper реконструировал из того же открытого набора данных, поэтому показанные значения являются предварительными, а не окончательными числами статьи.

Original figure — The Clean Paper CC BY 4.0

Что сделали авторы

- Провели два эксперимента с 2190 участниками, каждый из которых своими словами называл теорию заговора, в которую верил, и доказательства, которые считал её подтверждением.
- Каждый участник провёл трёхраундовый разговор с GPT-4 Turbo. В **экспериментальном** условии ИИ получал инструкцию опровергать конкретные доказательства участника; в **контрольном** — обсуждал постороннюю тему.
- Измерили веру в выбранную теорию заговора до разговора и сразу после него, а затем повторно связались с участниками через **10 дней и 2 месяца**.
- Попросили профессионального фактчекера оценить точность всех 128 фактических утверждений, сделанных ИИ в подвыборке разговоров.
- Проверили, переносится ли эффект на другие, не связанные теории заговора, а как тест специфичности — снижает ли он также веру в заговоры, которые в действительности оказались **правдивыми**.

Что они обнаружили

- **В среднем примерно 20%-е снижение** веры в выбранную теорию заговора в экспериментальной группе по сравнению с контролем (исследование 1: 95%-й доверительный интервал [13.8; 19.7] пункта по 100-балльной шкале, $P < 0.001$).
- **Эффект сохранялся.** В экспериментальной группе падение не исчезало: через 10 дней и два месяца вера оставалась примерно на уровне сразу после разговора, тогда как контрольная группа всё время была выше. Статья описывает эффект для выбранной теории как сохранявшийся без заметного ослабления как минимум два месяца, а не как короткий момент сомнения.
- **Он обобщался** на широкий спектр заговоров, которые называли люди, и сохранялся даже у тех, чья исходная вера была сильной.
- **Утверждения ИИ были точными** в этих условиях: 127 из 128 (99.2%) оценены как истинные, 1 (0.8%) — как вводящее в заблуждение, ни одного — как ложное.
- **Эффект был специфичным**, а не просто порождал общий скепсис: вмешательство **не** уменьшило веру в правдивые заговоры, что указывает на ослабление именно неподтверждённых убеждений, а не на превращение людей в циников по отношению ко всему.
- **Был и умеренный перенос**: вера в другие, не связанные теории заговора снизилась примерно на 12%, а участники сообщали о более сильном намерении возражать против конспирологических утверждений. Основной эффект сохранился в исследование 2 и имел то же направление в меньшей выборке без контрольной группы — более слабый тип доказательства, но согласующийся с остальными данными.

Чего это не доказывает

- Сейчас результат **нельзя считать бесспорным**. На статье стоит Editorial Expression of Concern из-за проблем с обработкой данных и воспроизводимостью, а исправленные числа всё ещё оценивает *Science*. Конкретные значения следует считать предварительными, пока проверка не завершена.
- Это **не** показывает, что ИИ «лечит» веру в теории заговора. Среднее снижение примерно на 20% — содержательный сдвиг, но не стирание убеждения; большинство участников после эксперимента всё ещё верили в теорию, просто слабее.
- Это **не** результат реального массового применения. Участники добровольно вступили в исследование и добросовестно взаимодействовали с ИИ; в реальном мире люди, наиболее глубоко убеждённые в теории заговора, могут как раз меньше всего хотеть обращаться к чатботу, который будет с ними спорить.
- Точность 99.2% **относится именно к этой задаче, модели и тщательному промптингу** — это не общая гарантия правдивости языковых моделей. Авторы прямо отмечают, что та же персонализированная сила убеждения могла бы толкать людей и к *ложным* убеждениям, если направить её в ту сторону.
- «Длительный» здесь означает **два месяца** — это впечатляет для одного разговора, но не означает навсегда.

Насколько сильны доказательства

- **Дизайн — настоящая сильная сторона.** Контролируемые эксперименты с treatment и control, чистый плацебоподобный контроль, повторение в двух исследованиях и независимой выборке, отложенные измерения длительности эффекта и отдельная проверка точности утверждений ИИ — всё это тщательнее большинства исследований убеждения, а направление эффекта было согласованным везде, где его проверяли.
- **Но Expression of Concern — не примечание мелким шрифтом.** Возможность воспроизвести опубликованные числа из открытых данных и кода — часть доверия к результату, и именно здесь произошла поломка: несоответствия в критериях скрининга и дублированные строки из-за ошибки при слиянии кода. Заявление авторов, что исправленный конвейер сохраняет результат, обнадёживает и выглядит правдоподобно, но это их собственная оценка, а не независимый окончательный вывод. Решающее значение будет иметь оценка *Science*.
- **Честный статус — «помечено для проверки», а не «опровергнуто» или «подтверждено».** Это хорошо построенное, яркое исследование, точные числа которого сейчас формально пересматривают. Неудобное место, чтобы оставить красивую историю, но именно оно сейчас точное.

Почему это важно

Годами влиятельной была идея, что сторонниками теорий заговора движут психологические потребности и идентичность, поэтому фактами их переубедить невозможно. Длительное снижение веры на основе доказательств — *если результат выдержит проверку* — является важным противовесом такому пессимизму. Оно предполагает, что часть проблемы была в том, что общие опровержения никогда не работали именно с *конкретными* доказательствами, на которые ссылается конкретный сторонник заговора; LLM способна делать это масштабируемо.

Это также важный пример того, как должна работать наука. Урок Expression of Concern не в том, что громкие результаты ничего не стоят, а в том, что ошибки всплывают, данные пересматриваются, а утверждения публично перепроверяются. Работа этого механизма — признак системы, а не скандал; именно поэтому правильная позиция по отношению к этому результату сейчас — терпение, а не аплодисменты или отмахивание.

И для ИИ эта история имеет две стороны. Та же способность ослабить ложное убеждение персонализированным аргументом может столь же эффективно усилить другое ложное убеждение. Техника нейтральна; цель — нет.

Краткий итог

Исследование 2024 года показало, что трёхраундовый разговор с GPT-4 Turbo уменьшал веру людей в теорию заговора, которой они придерживались, примерно на 20%; эффект сохранялся два месяца, наблюдался для разных типов заговоров, а фактические утверждения ИИ были оценены как почти полностью точные. В июне 2026 года к статье было добавлено Editorial Expression of Concern из-за проблем с обработкой данных и воспроизводимостью; авторы сообщают, что исправленный анализ сохраняет направление, значимость и величину результата, а *Science* всё ещё проводит оценку. Читать это следует как действительно интересное и тщательно построенное исследование, которое сейчас пересматривают, — не как установленную истину и не как опровергнутую работу. Самое честное предложение о нём сейчас пишет сам научный процесс: проверить ещё раз.

Источники

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