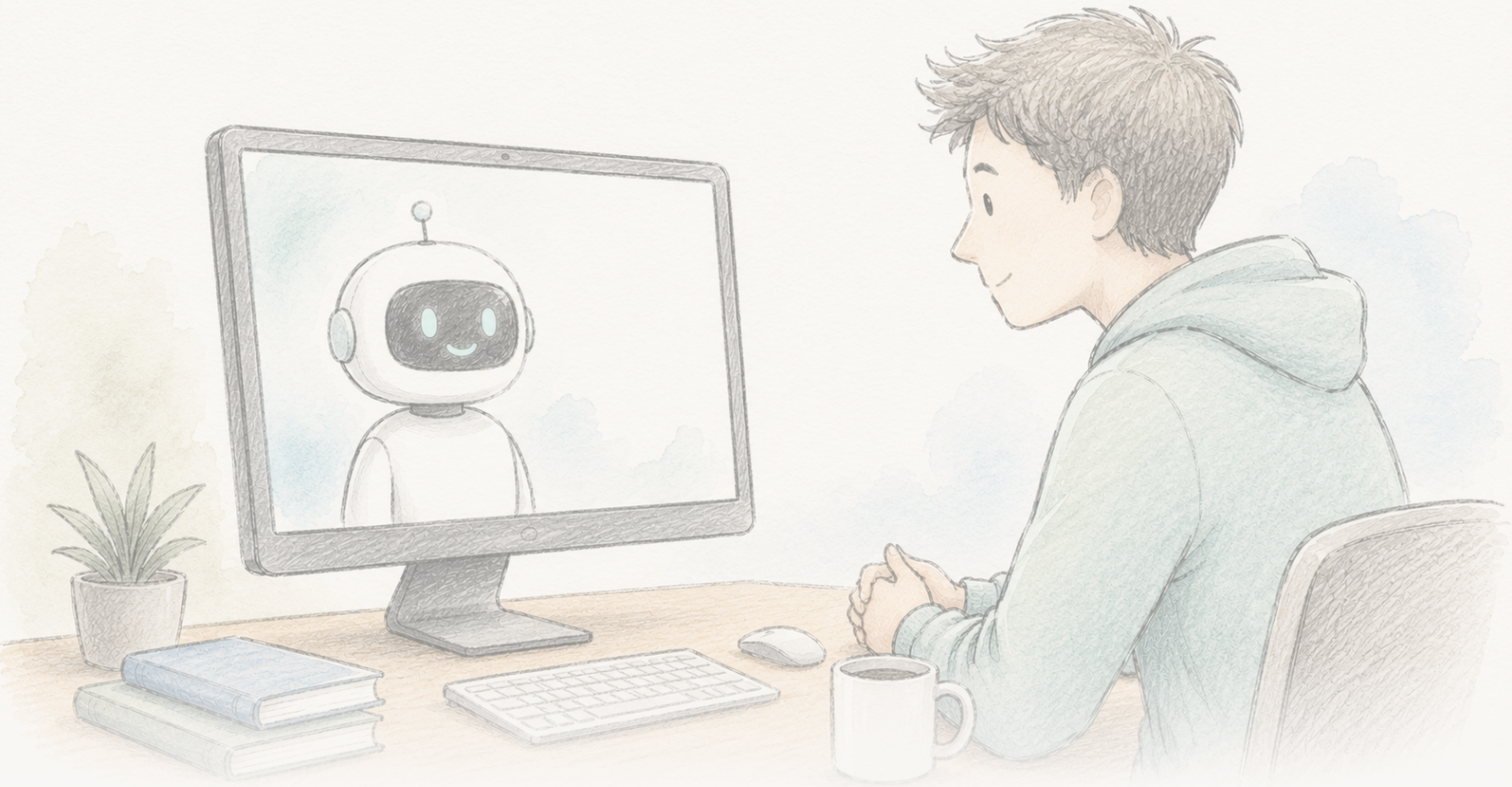


The Clean Paper

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

Схоже, чатбот на місяці послаблював віру в теорії змови

Дослідження 2024 року показало, що трираундова розмова з GPT-4 Turbo зменшувала віру людей у теорію змови, якої вони дотримувалися, приблизно на 20%; ефект тривав два місяці, спостерігався для різних типів змов і спирався на твердження ШІ, які фактчекер оцінив як майже повністю точні. У червні 2026 року на статті з'явилося Editorial Expression of Concern через проблеми з обробкою даних і відтворюваністю; автори стверджують, що виправлений аналіз зберігає напрямок, значущість і величину результату, а Science усе ще проводить оцінку. Справді цікаве й ретельно побудоване дослідження, яке зараз переглядають: воно ще не остаточно підтверджене, але й не спростоване.

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 додав до цієї статті Editorial Expression of Concern, поки оцінює питання щодо обробки даних і відтворюваності. Це не ретракція й не висновок про порушення; це означає, що опублікований результат слід вважати попереднім, доки перевірку не завершено.

Editorial Expression of Concern →

Зрештою, факти — і попереджувальна позначка на статті

У 2024 році дослідження повідомило результат, який суперечить зручній частині загальноприйнятої мудрості. Дайте людині коротку розмову з ШІ у три раунди — GPT-4 Turbo, налаштованим відповідати саме на ті докази, які вона наводить на підтримку теорії змови, у яку сама вірить, — і в середньому її віра в цю теорію зменшується приблизно на 20%. Зниження все ще спостерігалося через два місяці. Ефект працював і для класичних змов (убивство John F. Kennedy, прибульці, ілюмінати), і для актуальних (COVID-19, вибори у США 2020 року), навіть серед людей із сильною вірою, пов'язаною з їхньою ідентичністю. Коли професійний фактчекер оцінив 128 тверджень, зроблених ШІ, 127 виявилися правдивими, одне — оманливим, і жодне — хибним.

Заголовок, який розлетівся світом, був таким: людей *можна* витягнути з «кролячої нори» розмовою — прихильники теорій змови не недосяжні для доказів, їм просто потрібні правильні докази, подані достатньо конкретно. Це справді цікаве твердження, і саме дослідження було побудоване ретельніше, ніж більшість робіт про переконання.

А потім, у червні 2026 року, *Science* опублікував до статті **Editorial Expression of Concern** — редакційне повідомлення про занепокоєння. Саме цю частину більшість переказів пропустить, хоча зараз саме вона визначає, яку вагу може нести результат.

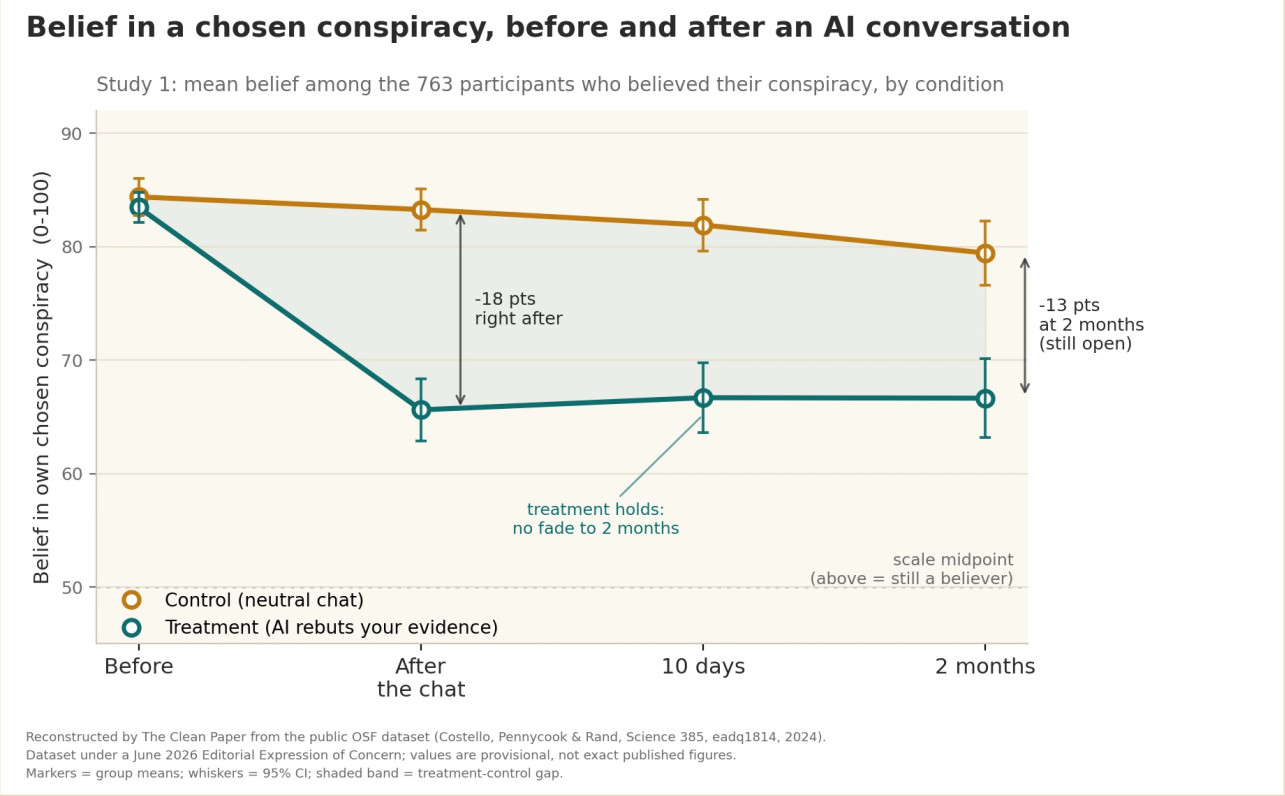
Що таке Expression of Concern — і чим воно не є

Editorial Expression of Concern — це офіційна позначка, яку журнал додає до статті, щоб попередити читачів: щодо роботи виникли питання, і їх усе ще з'ясовують. Це **не** ретракція — статтю не відкликано — і саме по собі це не висновок про порушення наукової етики. Воно означає: читайте з урахуванням застереження, бо науковий запис може змінитися. У цьому випадку після того, як авторів повідомили про проблеми, вони провели перевірку, надали *Science* конкретні деталі й передали виправлений аналіз.

Позначені проблеми конкретні. Авторів поінформували про невідповідності між критеріями відбору учасників, описаними в рукописі, і тим, як ці критерії були реалізовані в опублікованому коді аналізу; крім того, у відкритий набір даних через помилку під час об'єднання коду потрапили зайві рядки. Через це частину

опублікованих чисел було важко відтворити з доступних матеріалів. Автори передали *Science* виправлений аналітичний конвеєр та оновлені результати, які, за їхніми словами, збігаються з початковими **за напрямком, статистичною значущістю та змістовною величиною ефекту**. *Science* їх оцінює. Поки ця оцінка не завершиться, точні числа залишаються під знаком питання, зняти який може лише журнал.

Отже, тут одночасно дві історії: інтригуючий результат про те, чи можуть факти змінювати вкорінені переконання, і живий приклад того, як науковий запис відкрито виправляє сам себе. Завдання цього тексту — не змішувати їх.



У дослідженні повідомлялося, що трираундова розмова з ШІ, який спростовував власні аргументи учасника на підтримку обраної ним теорії змови, у середньому зменшувала віру в неї приблизно на 20%; на відміну від контрольної розмови на сторонню тему, зниження все ще вимірювалося через 10 днів і 2 місяці. Ефект був тривалим, але частковим: більшість учасників після цього все одно вірили в змову — їхня віра ослабла, а не зникла. Наразі на статті стоїть Editorial Expression of Concern (*Science*, червень 2026 року) через невідповідності у звітності та помилку у відкритому наборі даних; автори стверджують, що виправлений аналіз зберігає напрямок, значущість і величину ефекту, поки *Science* продовжує перевірку. Цю діаграму The Clean Paper реконструював із того самого відкритого набору даних, тому показані значення є попередніми, а не остаточними числами статті.

Original figure — The Clean Paper CC BY 4.0

Що зробили автори

- Провели два експерименти з 2190 учасниками, кожен із яких своїми словами назвав теорію змови, у яку вірив, та докази, які вважав її підтвердженням.
- Кожен учасник провів трираундову розмову з GPT-4 Turbo. В **експериментальній** умові ШІ отримував інструкцію спростовувати конкретні докази учасника; в **контрольній** — обговорював сторонню тему.
- Виміряли віру в обрану теорію змови до розмови й одразу після неї, а потім повторно зв'язалися з учасниками через **10 днів і 2 місяці**.
- Попросили професійного фактчекера оцінити точність усіх 128 фактичних тверджень, зроблених ШІ в підвибірці розмов.
- Перевірили, чи переноситься ефект на інші, не пов'язані теорії змови, а як тест специфічності — чи знижує він також віру в змови, які насправді виявилися **правдивими**.

Що вони виявили

- **У середньому приблизно 20% зниження** віри в обрану теорію змови в експериментальній групі порівняно з контролем (study 1: 95% довірчий інтервал [13,8; 19,7] пункту за 100-бальною шкалою, P < 0,001).
- **Ефект зберігався.** В експериментальній групі падіння не зникало: через 10 днів і два місяці віра залишалася приблизно на рівні одразу після розмови, тоді як контрольна група весь час була вище. Стаття описує ефект для обраної теорії як такий, що зберігався без помітного ослаблення щонайменше два місяці, а не як короткий момент сумніву.
- **Він узагальнювався** на широкий спектр змов, які називали люди, і зберігався навіть у тих, чия початкова віра була сильною.
- **Твердження ШІ були точними** в цих умовах: 127 із 128 (99,2%) оцінено як правдиві, 1 (0,8%) — як оманливе, жодного — як хибне.
- **Ефект був специфічним**, а не просто породжував загальний сумнів: втручання **не** зменшило віру в правдиві змови, що вказує на послаблення саме непідтверджених переконань, а не на перетворення людей на циніків щодо всього.
- **Був і помірний перенос**: віра в інші, не пов'язані теорії змови знизилася приблизно на 12%, а учасники повідомляли про сильніший намір заперечувати конспірологічні твердження. Основний ефект зберігся в study 2 і мав той самий напрямок у меншій вибірці без контрольної групи — слабший тип доказу, але узгоджений.

Чого це не доводить

- Наразі результат **не можна вважати беззаперечним**. На статті стоїть Editorial Expression of Concern через проблеми з обробкою даних і відтворюваністю, а виправлені числа все ще оцінює *Science*. Конкретні значення слід вважати попередніми, доки перевірка не завершиться.
- Це **не** показує, що ШІ «лікує» віру в теорії змови. Середнє зниження приблизно на 20% — змістовний зсув, але не стирання переконання; більшість учасників після експерименту все ще вірили в теорію, просто слабше.
- Це **не** результат реального масового застосування. Учасники добровільно вступили в дослідження й добросовісно взаємодіяли з ШІ; у реальному світі люди, найглибше переконані в теорії змови, можуть якраз найменше хотіти звертатися до чатбота, який із ними сперечатиметься.
- Точність 99,2% **стосується саме цього завдання, моделі та ретельного промптингу** — це не загальна гарантія правдивості мовних моделей. Автори прямо зазначають, що та сама персоналізована сила переконання могла б штовхати людей і до *хибних* переконань, якби її спрямувати в той бік.
- «Тривалий» тут означає **два місяці** — це вражає для однієї розмови, але не означає назавжди.

Наскільки сильні докази

- **Дизайн — справжня сильна сторона.** Контрольовані експерименти з treatment і control, чистий плацебоподібний контроль, повторення у двох дослідженнях та незалежній вибірці, відкладені вимірювання тривалості ефекту й окрема перевірка точності тверджень ШІ — це ретельніше за більшість досліджень переконання, а напрямок ефекту був узгодженим усюди, де його тестували.
- **Але Expression of Concern — не примітка дрібним шрифтом.** Можливість відтворити опубліковані числа з відкритих даних і коду — частина довіри до результату, і саме тут сталася поломка: невідповідності в критеріях скринінгу та дубльовані рядки через помилку під час об’єднання коду. Заява авторів, що виправлений конвеєр зберігає результат, обнадіює й виглядає правдоподібно, але це їхня власна оцінка, а не незалежний остаточний висновок. Вирішальною буде оцінка *Science*.
- **Чесний статус — «позначено для перевірки», а не «спростовано» чи «підтверджено».** Це добре побудоване, яскраве дослідження, точні числа якого зараз формально переглядають. Незручне місце, щоб залишити гарну історію, але саме воно наразі точне.

Чому це важливо

Роками впливовою була думка, що прихильниками теорій змови керують психологічні потреби та ідентичність, тому фактами їх переконати неможливо. Тривале зниження віри на основі доказів — *якщо результат витримає перевірку* — є важливою противагою такому песимізму. Воно припускає, що частина проблеми полягала в тому, що загальні спростування ніколи не працювали саме з *конкретними* доказами, на які посилається конкретний прихильник змови; LLM здатна робити це масштабовано.

Це також важливий приклад того, як має працювати наука. Урок Expression of Concern не в тому, що гучні результати нічого не варті, а в тому, що помилки спливають, дані переглядаються, а твердження повторно перевіряються публічно. Робота цього механізму — ознака системи, а не скандал; саме тому правильна позиція щодо цього результату зараз — терпіння, а не аплодисменти чи відмахування.

І щодо ШІ ця історія має два боки. Та сама здатність послабити хибне переконання персоналізованим аргументом може так само ефективно посилити інше хибне переконання. Техніка нейтральна; мета — ні.

Короткий підсумок

Дослідження 2024 року показало, що трираундова розмова з GPT-4 Turbo зменшувала віру людей у теорію змови, якої вони дотримувалися, приблизно на 20%; ефект зберігався два місяці, спостерігався для різних типів змов, а фактичні твердження ШІ були оцінені як майже повністю точні. У червні 2026 року на статті з'явилося Editorial Expression of Concern через проблеми з обробкою даних і відтворюваністю; автори повідомляють, що виправлений аналіз зберігає напрямок, значущість і величину результату, а *Science* усе ще проводить оцінку. Читати це слід як справді цікаве й ретельно побудоване дослідження, яке зараз переглядають, — не як установлену істину й не як спростовану роботу. Найчесніше речення про нього зараз пише сам науковий процес: перевірити ще раз.

Джерела

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