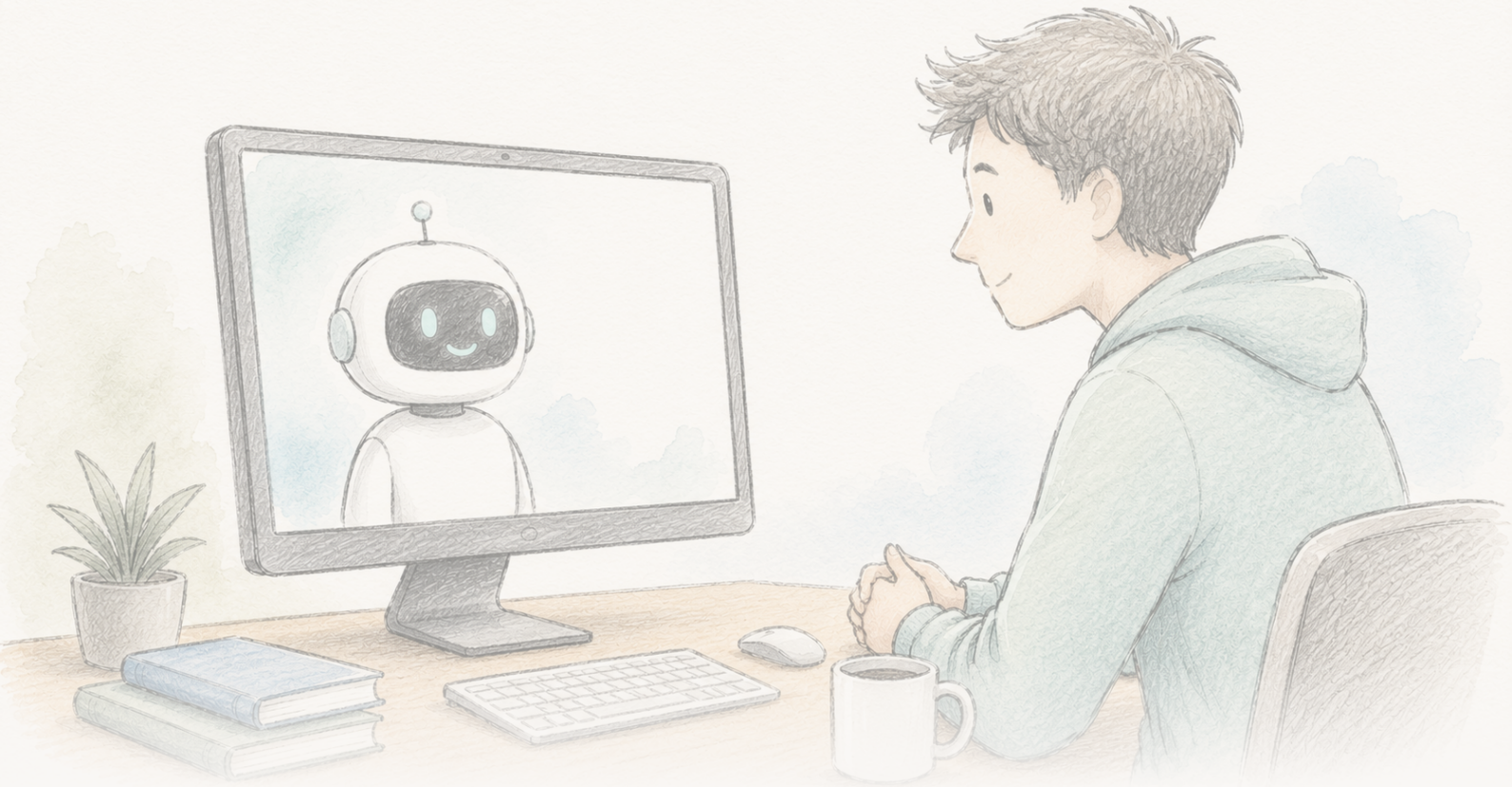


The Clean Paper

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

Chatbot zdawał się na wiele miesięcy osłabiać wiarę w teorie spiskowe

Badanie z 2024 roku wykazało, że trzyetapowa rozmowa z GPT-4 Turbo obniżała wiarę uczestników w wybraną teorię spiskową o około 20%; efekt trwał dwa miesiące, obejmował różne rodzaje spisków i opierał się na twierdzeniach AI ocenionych niemal całkowicie jako prawdziwe. W czerwcu 2026 roku pracę opatrzono redakcyjną notą wyrażającą zaniepokojenie z powodu problemów z przetwarzaniem danych i odtwarzalnością; autorzy twierdzą, że poprawiona analiza zachowuje kierunek, istotność i wielkość wyniku, a Science nadal go ocenia. To naprawdę interesujący, starannie zaprojektowany wynik — obecnie pod kontrolą, ani ustalony, ani obalony.

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 opatrzyło tę pracę redakcyjną notą wyrażającą zaniepokojenie na czas oceny pytań dotyczących przetwarzania danych i odtwarzalności. Nie jest to wycofanie pracy ani stwierdzenie nierzetelności; oznacza, że raportowany wynik należy traktować jako tymczasowy do zakończenia kontroli.

Editorial Expression of Concern →

Fakty jednak działają — i ostrzeżenie przy pracy

W 2024 roku opublikowano badanie, które przeczy wygodnemu elementowi obiegowej wiedzy. Jeśli ktoś odbędzie krótką, trzyetapową rozmowę ze sztuczną inteligencją — GPT-4 Turbo, poinstruowanym, by odpowiedzieć na konkretne dowody przytaczane przez tę osobę na poparcie teorii spiskowej, w którą wierzy — jego wiara w tę teorię spada przeciętnie o około 20%. Spadek był nadal widoczny dwa miesiące później. Efekt występował w przypadku klasycznych spisków (zamach na Johna F. Kennedy'ego, kosmici, iluminaci) i aktualnych (COVID-19, wybory w USA w 2020 roku), nawet u osób, których przekonanie było silne i związane z tożsamością. Gdy zawodowy weryfikator faktów ocenił 128 twierdzeń przedstawionych przez AI, 127 było prawdziwych, jedno wprowadzające w błąd, a żadne fałszywe.

Nagłówek, który obiegł świat, głosił, że *da się* wyprowadzić ludzi z króliczej nory rozmową — zwolennicy teorii spiskowych nie są poza zasięgiem dowodów, tylko potrzebują właściwych dowodów podanych dostatecznie konkretnie. To naprawdę interesujące twierdzenie, a badanie zaprojektowano staranniej niż większość badań nad perswazją.

Następnie w czerwcu 2026 roku czasopismo *Science* opatrzyło pracę **redakcyjną notą wyrażającą zaniepokojenie** (Editorial Expression of Concern). Większość późniejszych opowieści pominie tę część, choć to właśnie ona określa, jaki ciężar wynik może obecnie udźwignąć.

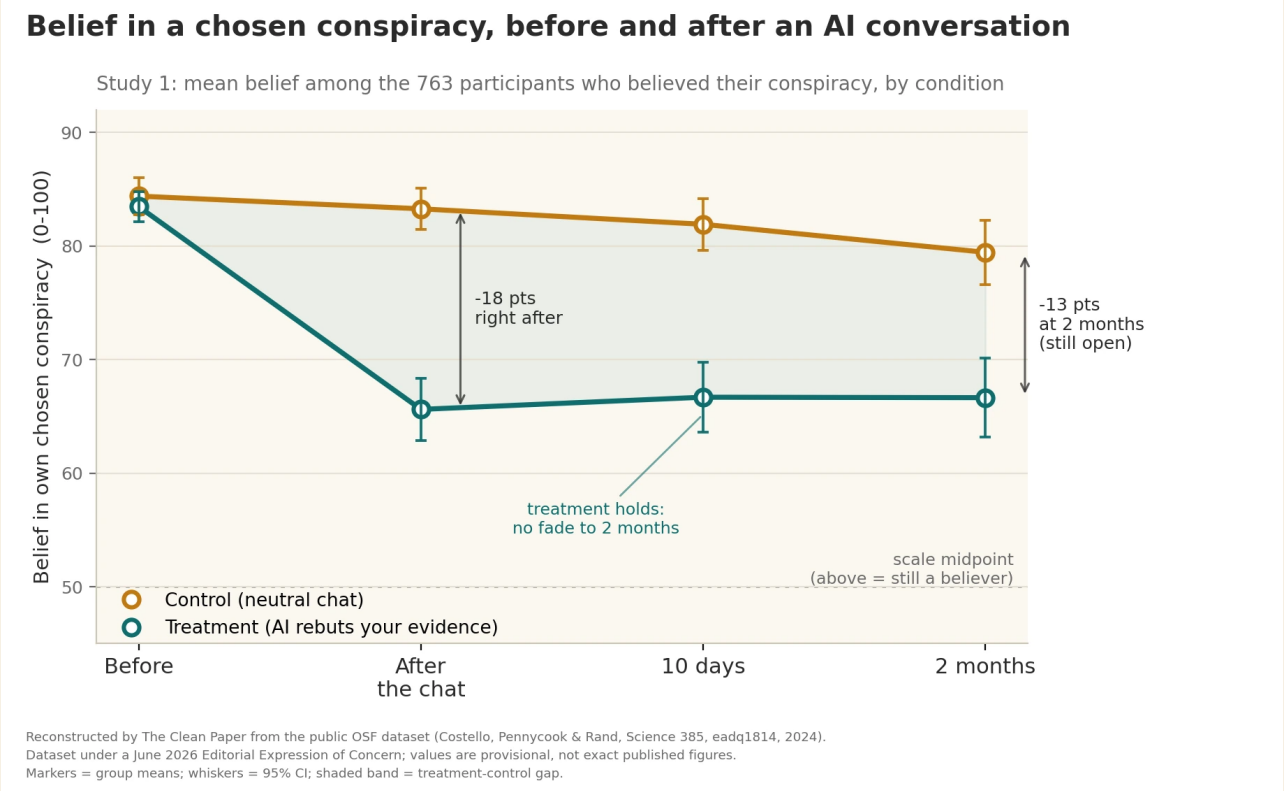
Czym jest — i czym nie jest — redakcyjna nota wyrażająca zaniepokojenie

Redakcyjna nota wyrażająca zaniepokojenie to formalne ostrzeżenie dołączane przez czasopismo do pracy, aby poinformować czytelników o zgłoszonych pytaniach, gdy ich wyjaśnianie nadal trwa. **Nie** jest to wycofanie publikacji ani samo w sobie stwierdzenie nierzetelności. Oznacza: czytaj z uwzględnieniem tego zastrzeżenia, ponieważ zapis naukowy może się zmienić. W tym przypadku autorzy po uzyskaniu informacji o problemach przeprowadzili dochodzenie, zgłosili szczegóły czasopismu *Science* i dostarczyli poprawioną analizę.

Zastrzeżenia są konkretne. Autorów poinformowano o niespójnościach między tekstem pracy a opublikowanym kodem analitycznym w stosowaniu kryteriów kwalifikacji uczestników oraz o tym, że publiczny zbiór danych zawierał zbędne wiersze włączone wskutek błędu przy łączeniu kodu — problemy utrudniające odtworzenie części raportowanych liczb na podstawie udostępnionych materiałów. Autorzy przekazali *Science* poprawioną procedurę analizy i zestaw zaktualizowanych wyników, które

według nich odpowiadają pierwotnym pod względem **kierunku, istotności statystycznej i wielkości merytorycznej**. *Science* je ocenia. Dopóki ocena się nie zakończy, nad dokładnymi liczbami pozostaje znak zapytania, który może usunąć tylko czasopismo.

To zatem dwie historie jednocześnie: intrygujący wynik dotyczący wpływu faktów na głęboko zakorzenione przekonania oraz żywy przykład publicznego samokorygowania się zapisu naukowego. Celem tego artykułu jest ich nie pomylić.



W badaniu trzyetapowa rozmowa z AI, która odpierała dowody wskazane przez uczestnika, miała zmniejszać jego wiarę w wybraną teorię spiskową średnio o około 20%; w przeciwieństwie do rozmowy kontrolnej na niezwiązany temat spadek pozostawał mierzalny po 10 dniach i 2 miesiącach. Raportowany efekt był trwały, lecz częściowy: większość uczestników nadal wierzyła później w spisek — była to redukcja, nie lekarstwo. Praca jest obecnie opatrzona redakcyjną notą wyrażającą zaniepokojenie (*Science*, czerwiec 2026) z powodu niespójności raportowania i błędu w publicznym zbiorze danych; autorzy twierdzą, że poprawiona analiza zachowuje kierunek, istotność i wielkość efektu, a *Science* nadal ją ocenia. Wykres odtworzył The Clean Paper na podstawie publicznego zbioru danych, dlatego przedstawione wartości są tymczasowe i nie stanowią ostatecznych liczb z pracy.

Co zrobili autorzy

- Przeprowadzili dwa eksperymenty z udziałem 2190 osób, z których każda własnymi słowami wskazała teorię spiskową, w którą wierzy, oraz dowody, które uważa za jej poparcie.

- Każdy uczestnik odbył trzyetapową rozmowę z GPT-4 Turbo. W warunku **eksperymentalnym** AI miała odpierać konkretne dowody uczestnika; w warunku **kontrolnym** omawiała niezwiązany temat.
- Zmierzono wiarę w wybrany spisek przed rozmową i bezpośrednio po niej, a następnie ponownie skontaktowano się z uczestnikami po **10 dniach i 2 miesiącach**.
- Zawodowy weryfikator faktów ocenił prawdziwość wszystkich 128 twierdzeń faktycznych przedstawionych przez AI w próbie.
- Sprawdzono, czy efekt przenosił się na inne, niezwiązane spiski — oraz, jako test swoistości, czy obniżał również wiarę w spiski, które rzeczywiście były **prawdziwe**.

Co odkryli

- **Średni spadek o około 20%** wiary w wybrany spisek w grupie eksperymentalnej względem kontrolnej (badanie 1: 95-procentowy przedział ufności [13,8; 19,7] punktu na skali 100-punktowej, $P < 0{,}001$).
- **Efekt się utrzymał.** W grupie eksperymentalnej spadek nie zanikał: po 10 dniach i dwóch miesiącach wiara pozostawała blisko poziomu tuż po rozmowie, zamiast stopniowo wracać, podczas gdy grupa kontrolna przez cały czas pozostawała wyżej. Praca opisuje wpływ na wybrany spisek jako utrzymujący się bez osłabienia przez co najmniej dwa miesiące, a nie jako chwilowe zachwianie.
- **Efekt uogólniał się** na różne spiski wskazywane przez uczestników i występował nawet u osób początkowo głęboko przekonanych.
- **Twierdzenia AI były w tym kontekście trafne:** 127 ze 128 (99,2%) uznano za prawdziwe, 1 (0,8%) za wprowadzające w błąd, a żadnego za fałszywe.
- **Efekt był swoisty**, a nie oparty na ogólnym zwątpieniu: interwencja **nie** obniżała wiary w prawdziwe spiski, co sugeruje, że zmieniała nieuzasadnione przekonania, zamiast wywoływać cynizm wobec wszystkiego.
- **Efekt przenosił się umiarkowanie** — wiara w inne, niezwiązane spiski spadła o około 12%, a uczestnicy deklarowali większą gotowość do sprzeciwiania się twierdzeniom spiskowym. Główny efekt utrzymał się w badaniu 2 i wskazywał ten sam kierunek w mniejszej próbie bez grupy kontrolnej (dowód słabszy, ale zgodny).

Czego to nie dowodzi

- Praca **nie** pozostaje obecnie niekwestionowana. Jest opatrzona redakcyjną notą wyrażającą zaniepokojenie z powodu problemów z przetwarzaniem danych i odtwarzalnością, a *Science* nadal ocenia poprawione liczby. Konkretne wartości należy traktować jako tymczasowe do zakończenia tej oceny.
- Wynik **nie** pokazuje, że AI „leczy” wiarę w spiski. Średni spadek o około 20% jest znaczącą zmianą, nie usunięciem przekonania — większość uczestników nadal wierzyła później w teorię, tylko słabiej.
- To **nie** jest wynik rzeczywistego wdrożenia. Uczestnicy dobrowolnie wzięli udział w badaniu i

rozmawiali z AI w dobrej wierze; poza laboratorium osoby najbardziej przywiązane do spisku najrzadziej będą szukać chatbota, który podejmie z nimi spór.
- Trafność 99,2% jest **właściwa dla tego zadania, modelu i starannego promptu** — nie gwarantuje ogólnie, że modele językowe mówią prawdę. Autorzy wyraźnie zaznaczają, że ta sama spersonalizowana siła perswazji mogłaby wzmacniać *fałszywe* przekonania, gdyby została skierowana w tym celu.
- „Trwały" oznacza tu **dwa miesiące** — imponujący czas po pojedynczej rozmowie, ale nie to samo co trwałość permanentna.

Jak silne są dowody

- **Projekt badania jest prawdziwą zaletą.** Kontrolowane eksperymenty porównujące interwencję z kontrolą, dobrze zdefiniowany warunek kontrolny przypominający placebo, powtórzenie w dwóch badaniach i niezależnej próbie, pomiary trwałości oraz jawna kontrola trafności twierdzeń AI — to staranniejsze podejście niż w większości badań nad perswazją, a kierunek efektu jest zgodny wszędzie, gdzie go testowano.
- **Redakcyjna nota wyrażająca zaniepokojenie nie jest jednak przypisem.** Możliwość odtworzenia raportowanych liczb z udostępnionych danych i kodu należy do podstaw zaufania do wyniku, a właśnie ona zawiodła — przez niespójności kryteriów kwalifikacji i powielone wiersze wynikające z błędu łączenia kodu. Twierdzenie autorów, że poprawiona procedura zachowuje wynik, jest zachęcające i wiarygodne, ale pozostaje ich własnym opisem, nie niezależnym werdyktem. Trzeba poczekać na ocenę *Science*.
- **Najuczciwszy status to „opatrzone ostrzeżeniem", a nie „obalone" ani „potwierdzone".** To dobrze zaprojektowane, uderzające badanie, którego dokładne liczby są formalnie sprawdzane. Pozostawienie dobrej historii w tak niewygodnym miejscu jest właśnie trafne.

Dlaczego to ważne

Przez lata wpływowy pogląd głosił, że zwolennikami teorii spiskowych kierują potrzeby psychologiczne i tożsamość, dlatego fakty nie potrafią ich przekonać. Trwały spadek oparty na dowodach — jeśli przetrwa kontrolę — stanowi znaczącą przeciwwagę dla tego pesymizmu: sugeruje, że problem częściowo polegał na tym, iż ogólne obalanie spisków nie odpowiadało na *konkretne* dowody przytaczane przez daną osobę, a model językowy może robić to na dużą skalę.

Wynik ma też znaczenie jako studium właściwego działania nauki. Lekcja z redakcyjnej noty wyrażającej zaniepokojenie nie brzmi: głośne wyniki są bezwartościowe. Błędy wychodzą na jaw, dane są ponownie badane, a twierdzenia sprawdzane publicznie. Działanie tego mechanizmu jest zaletą, nie skandalem — i dlatego właściwą reakcją na wynik jest cierpliwość, a nie natychmiastowy zachwyt ani odrzucenie.

W przypadku AI ostrze działa w obie strony. Ta sama zdolność osłabiania fałszywego przekonania za

pomocą dobranego argumentu może równie skutecznie wzmocnić inne. Technika jest neutralna; cel już nie.

Zwięzłe podsumowanie

Badanie z 2024 roku wykazało, że trzyetapowa rozmowa z GPT-4 Turbo zmniejszała wiarę uczestników w wybraną teorię spiskową o około 20%, a efekt utrzymywał się przez dwa miesiące i uogólniał na różne rodzaje spisków; twierdzenia faktyczne AI oceniono jako niemal całkowicie prawdziwe. W czerwcu 2026 roku pracę opatrzono redakcyjną notą wyrażającą zaniepokojenie z powodu problemów z przetwarzaniem danych i odtwarzalnością; autorzy informują, że poprawiona analiza zachowuje kierunek, istotność i wielkość wyniku, a *Science* nadal go ocenia. Należy to czytać jako naprawdę interesujący, starannie zaprojektowany wynik obecnie poddawany kontroli — ani ustalony, ani obalony. Najuczciwsze zdanie pisze sam proces: sprawdźmy jeszcze raz.

Źródła

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`