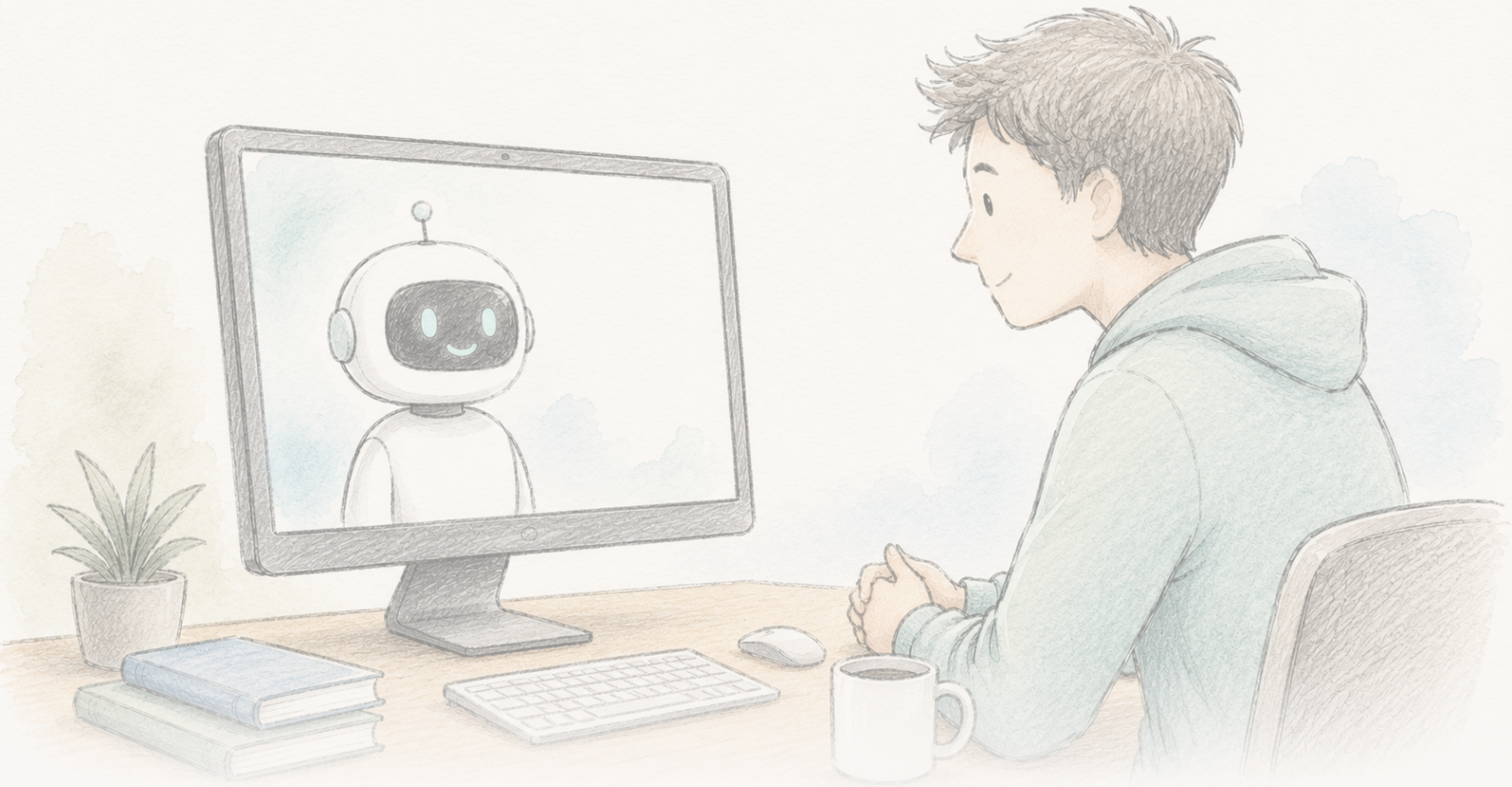


The Clean Paper

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

O AI a redus credințele conspiraționiste ale oamenilor timp de luni — iar studiul este acum sub Expression of Concern

Un studiu din 2024 a găsit că o conversație în trei runde cu GPT-4 Turbo reducea credința oamenilor într-o teorie a conspirației pe care o susțineau cu aproximativ 20%, un efect care a durat două luni, s-a generalizat peste tipuri de conspirații și s-a sprijinit pe afirmații ale AI evaluate covârșitor ca exacte. În iunie 2026, articolul a fost plasat sub Editorial Expression of Concern pentru probleme de gestionare a datelor și reproducibilitate; autorii spun că o analiză corectată păstrează rezultatul în direcție, semnificație și mărime, iar Science încă evaluează. Un rezultat cu adevărat interesant, construit cu grijă — în prezent sub revizuire, nici stabilit, nici demontat.

Written by Lucio Vaglio · Cross-checked by Cinzia Vaglio and Vera Rubin · 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

Faptele, până la urmă — și un semnal pe articol

În 2024, un studiu a raportat ceva care taie împotriva unei bucăți comode de înțelepciune convențională. Dă cuiva o conversație scurtă, în trei runde, cu o AI - GPT-4 Turbo, promptată să răspundă la dovezile specifice pe care persoana le invocă pentru o teorie a conspirației în care crede personal - și, în medie, credința acelei persoane în teorie scade cu aproximativ 20%. Scăderea era încă acolo două luni mai târziu. A funcționat peste conspirații clasice (asasinarea lui John F. Kennedy, extratereștri, Illuminati) și teme actuale (COVID-19, alegerile din SUA din 2020), și chiar printre oameni a căror credință era puternică și legată de identitate. Când un fact-checker profesionist a evaluat 128 dintre afirmațiile făcute de AI, 127 au fost adevărate, una a fost înșelătoare și niciuna falsă.

Titlul care a circulat a fost că *poți* vorbi oamenii afară din vizuina iepurelui: că cei care cred în conspirații nu sunt dincolo de raza dovezilor, ci au nevoie de dovezile potrivite, livrate suficient de specific. Este o afirmație cu adevărat interesantă, iar studiul a fost construit mai atent decât majoritatea cercetării despre persuasiune.

Apoi, în iunie 2026, *Science* a publicat o **Editorial Expression of Concern** asupra articolului. Aceasta este partea pe care multe repovestiri o vor sări, și este exact partea care decide câtă greutate poate purta rezultatul acum.

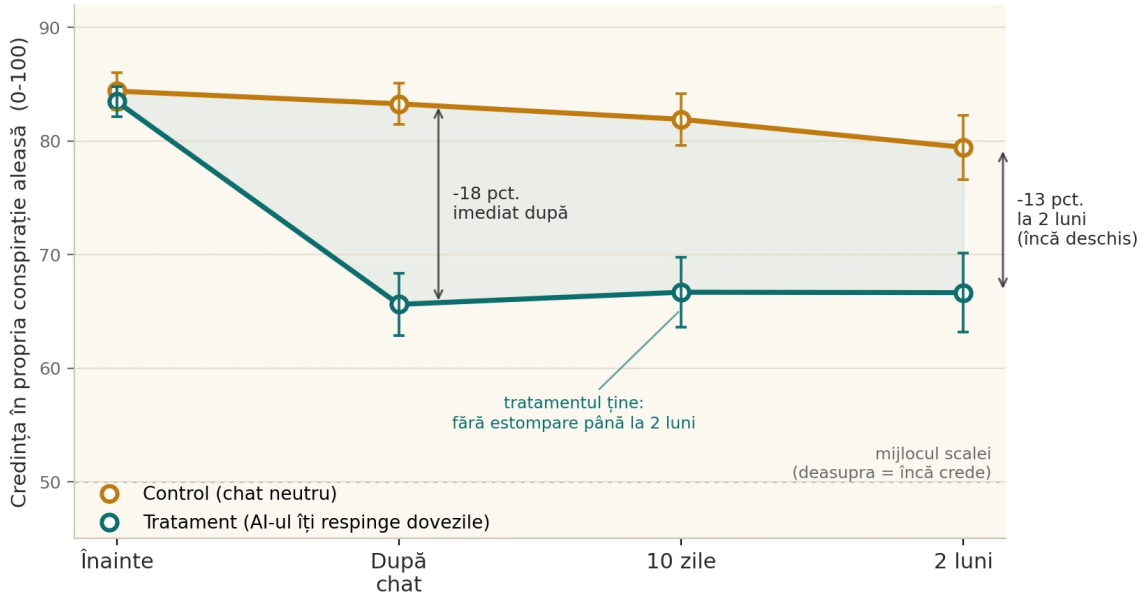
> **Ce este o Expression of Concern — și ce nu este**
>
> O Editorial Expression of Concern este un semnal formal pe care o revistă îl atașează unui articol pentru a alerta cititorii că au fost ridicate întrebări, în timp ce acele întrebări sunt încă investigate. **Nu** este o retragere (articolul nu este retras), și nu este prin ea însăși o constatare de conduită greșită. Înseamnă: citește cu acest caveat în minte, pentru că recordul se poate schimba. Aici, după ce au fost informați despre probleme, autorii au investigat, au raportat detaliile către Science și au furnizat o analiză corectată.

Ce a fost semnalat este specific. Autorii au fost informați despre inconsistențe în felul în care criteriile de screening ale participanților au fost aplicate între manuscris și codul de analiză publicat, și despre faptul că datasetul public conținea rânduri străine inserate printr-o eroare de merge a codului - probleme care făceau unele dintre numerele raportate greu de reprodus din materialele publicate. Autorii au dat către *Science* o pipeline de analiză corectată și un set de rezultate actualizate, despre care spun că se potrivesc cu originalul **în direcție, semnificație statistică și mărime substanțială**. *Science* le evaluează. Până când acea evaluare se termină, cifrele exacte stau sub un semn de întrebare pe care doar revista îl poate ridica.

Așadar sunt două povești deodată: un rezultat intrigant despre dacă faptele pot mișca credințe înrădăcinate, și un exemplu viu al recordului științific corectându-se în public. Ideea acestui text este să le țină separate.

Credința în conspirația aleasă, înainte și după o conversație cu o AI

Studiul 1: credința medie printre cei 763 participanți care credeau în conspirația lor, pe condiții



Reconstruit de The Clean Paper din datasetul public OSF (Costello, Pennycook & Rand, Science 385, eadq1814, 2024).
Dataset sub Editorial Expression of Concern din iunie 2026; valori provizorii, nu cifre publicate exacte.
Markeri = medii de grup; whiskere = IC 95%; bandă umbrită = gap tratament-control.

În studiu, o conversație AI în trei runde care răspundea la dovezile proprii declarate de participant a fost raportată ca reducând cu aproximativ 20% în medie credința în conspirația aleasă de acea persoană; spre deosebire de un chat de control pe un subiect nelegat, scăderea era încă măsurabilă la 10 zile și la 2 luni. Efectul raportat a fost durabil, dar parțial: majoritatea participanților încă mai credeau în conspirație după aceea - o reducere, nu o vindecare. Articolul este în prezent sub Editorial Expression of Concern (*Science*, iunie 2026) pentru inconsistențe de raportare și o eroare în datasetul public; autorii spun că o analiză corectată păstrează direcția, semnificația și mărimea efectului, în timp ce *Science* continuă evaluarea. Acest grafic este reconstruit de The Clean Paper din acel dataset public, deci valorile arătate sunt provizorii, nu cifrele finale ale articolului.

Ce au făcut autorii

- Au rulat două experimente cu 2190 de participanți, fiecare numind - în propriile cuvinte - o teorie a conspirației în care credea și dovezile pe care credea că se sprijină.
- Fiecare participant a avut o conversație în trei runde cu GPT-4 Turbo. În condiția de **tratament**, AI-ul a fost promptat să respingă dovezile specifice ale participantului; în condiția de **control**, a discutat un subiect nelegat.
- Au măsurat credința în conspirația aleasă înainte și imediat după conversație, apoi au recontactat participanții la **10 zile și 2 luni**.
- Au cerut unui fact-checker profesionist să evalueze acuratețea tuturor celor 128 de afirmații factuale făcute de AI într-un eșantion.
- Au verificat dacă efectul se răsfrângea asupra altor conspirații nelegate și, ca test de specificitate, dacă reducea și credința în conspirații care se întâmplă să fie **adevărate**.

Ce au găsit

- **O reducere medie de aproximativ 20%** în credința în conspirația aleasă în grupul de tratament față de control (studiul 1: interval de încredere de 95% [13,8, 19,7] puncte pe o scară de 100 de puncte, $P < 0,001$).
- **A durat.** În grupul de tratament, scăderea nu s-a estompat: credința a rămas aproape de nivelul imediat după chat la 10 zile și două luni, în loc să urce înapoi, în timp ce controlul a rămas mai sus pe tot parcursul; articolul descrie efectul asupra conspirației alese ca persistând nediminuat cel puțin două luni, nu ca o oscilație de moment.
- **S-a generalizat** în gama de conspirații numite de oameni și a ținut chiar și pentru participanții a căror credință era inițial puternică.
- **Afirmațiile AI-ului au fost exacte** în acest cadru: 127 din 128 (99,2%) evaluate adevărate, 1 (0,8%) înșelătoare, niciuna falsă.
- **A fost specific**, nu scepticism generalizat: intervenția **nu** a redus credința în conspirații adevărate, sugerând că a mișcat credințe nesusținute, nu i-a făcut pe oameni cinici despre orice.
- **S-a răsfrânt modest**: credința în alte conspirații nelegate a scăzut cu aproximativ 12%, iar participanții au raportat intenție mai mare de a respinge afirmații conspiraționiste. Efectul principal a ținut în studiul 2 și a arătat în aceeași direcție într-un eșantion mai mic fără grup de control (evidență mai slabă, dar consistentă).

Ce nu dovedește

- **Nu** stă neîntrebat acum. Articolul este sub Editorial Expression of Concern pentru probleme de gestionare a datelor și reproducibilitate, iar numerele corectate sunt încă evaluate de *Science*. Tratează valorile specifice ca provizorii până când acea revizuire ajunge.
- **Nu** arată că AI „vindecă” credința în conspirații. O reducere medie de ~20% este o schimbare semnificativă, nu o ștergere: majoritatea participanților încă mai credeau teoria după aceea, doar mai puțin.
- **Nu** este un rezultat de deployment în lumea reală. Participanții au optat pentru un studiu și au interacționat cu AI-ul cu bună-credință; în sălbăticie, oamenii cei mai dedicați unei conspirații sunt cei mai puțin probabil să caute un chatbot care să se certe cu ei.
- Acuratețea de 99,2% este **specifică acestei sarcini, acestui model și promptingului atent** - nu o garanție generală că modelele lingvistice spun lucruri adevărate. Autorii sunt expliciți că aceeași putere persuasivă personalizată ar putea împinge credințe *false* dacă ar fi țintită astfel.
- „Durabil” aici înseamnă **două luni**, ceea ce este impresionant pentru o singură conversație, dar nu este același lucru cu permanent.

Cât de puternică este evidența

- **Designul este un punct forte real.** Experimente controlate tratament-versus-control, un control curat de tip placebo, replicare în două studii și un eșantion independent, follow-up-uri de durabilitate și un control explicit al acurateții afirmațiilor AI-ului - este mai atent decât majoritatea cercetării despre persuasiune, iar direcția efectului este consistentă peste tot unde a fost testată.
- **Dar Expression of Concern nu este o notă de subsol.** Capacitatea de a regenera numerele raportate din datele și codul publicate face parte din ceea ce face un rezultat de încredere, iar exact asta s-a rupt: inconsistențe în criteriile de screening și rânduri duplicate dintr-o eroare de merge a codului. Declarația autorilor că o pipeline corectată păstrează rezultatul este încurajatoare și plauzibilă, dar este relatarea autorilor, nu încă un verdict independent. Evaluarea *Science* este lucrul de așteptat.
- **Statutul onest este „semnalat", nu „demontat" sau „confirmat".** Un studiu bine construit și izbitor, ale cărui numere exacte sunt sub revizuire formală. Este un loc inconfortabil în care să lași o poveste bună, și este cel corect.

De ce contează

Ani de zile, o viziune influentă a susținut că cei care cred în conspirații sunt conduși de nevoi psihologice și identitate, și deci nu pot fi convinși cu fapte. O reducere durabilă, bazată pe dovezi - dacă rezistă - este o contrapondere semnificativă la acel pesimism: sugerează că problema era parțial că debunkingul generic nu angaja niciodată *dovezile specifice* pe care credinciosul le citează de fapt, ceva ce un LLM poate face la scară.

Contează și ca studiu de caz despre cum ar trebui să funcționeze știința. Lecția Expression of Concern nu este că rezultatele celebrate sunt inutile; este că erorile ies la suprafață, datele sunt reexaminate și afirmațiile sunt reverificate public. Această mașinărie în funcțiune este o trăsătură, nu un scandal - și este motivul pentru care postura corectă față de acest rezultat este răbdarea, nu aplauzele sau respingerea.

Și taie în ambele sensuri despre AI. Aceeași capacitate de a dezumfla o credință falsă cu un argument personalizat ar putea umfla una la fel de eficient. Tehnica este neutră; numai scopul nu este.

Sinteză curată

Un studiu din 2024 a găsit că o conversație în trei runde cu GPT-4 Turbo reducea credința oamenilor într-o teorie a conspirației pe care o susțineau cu aproximativ 20%, un efect care persista două luni și se generaliza peste tipuri de conspirații, cu afirmațiile factuale ale AI-ului evaluate covârșitor ca exacte. În iunie 2026, articolul a fost plasat sub Editorial Expression of Concern pentru probleme de gestionare a datelor și reproducibilitate; autorii raportează că o analiză corectată păstrează rezultatul în direcție, semnificație și mărime, iar *Science* încă evaluează. Citește-l ca pe un rezultat cu adevărat

interesant, construit cu grijă, aflat în prezent sub revizuire - nu stabilit, nu demontat. Cea mai onestă propoziție despre el este cea pe care procesul însuși o scrie: verifică-l din nou.

Sources

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) — prepended to the article PDF
https://doi.org/10.1126/science.adq1814