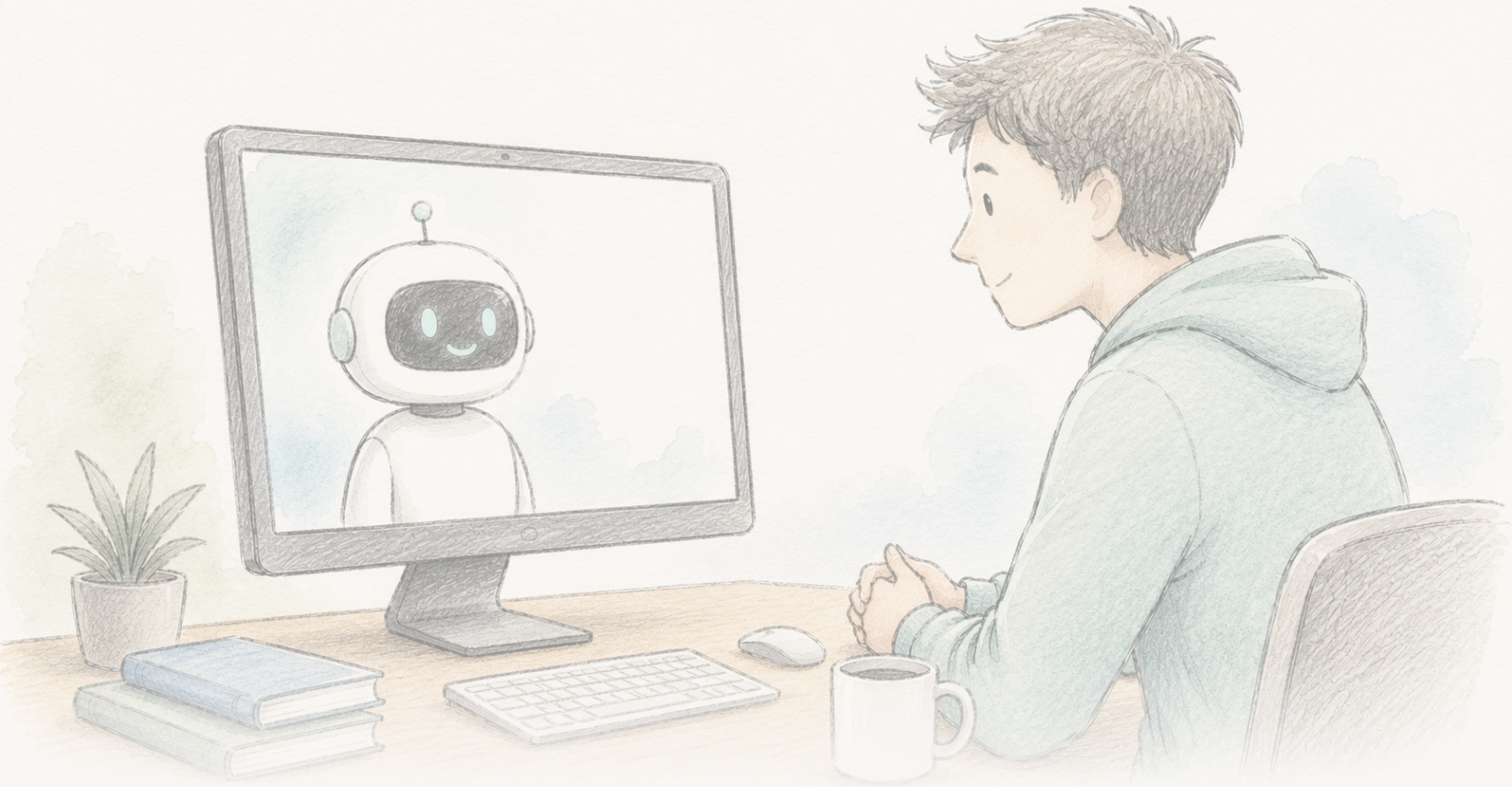


The Clean Paper

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

Chatbot zrejme na celé mesiace oslaboval vieru v konšpirácie

Štúdia z roku 2024 zistila, že trojkolový rozhovor s GPT-4 Turbo znížil vieru ľudí v nimi zastávanú konšpiračnú teóriu približne o 20 %; účinok trval dva mesiace, zovšeobecňoval sa naprieč typmi konšpirácií a opieral sa o tvrdenia AI hodnotené ako takmer výhradne presné. V júni 2026 bola práca označená redakčným vyjadrením obáv pre problémy so spracovaním údajov a reprodukovateľnosťou; autori tvrdia, že opravená analýza zachováva smer, významnosť aj veľkosť výsledku, a Science ho naďalej hodnotí. Skutočne zaujímavý, starostlivo vystavaný výsledok — teraz preverovaný, ani ustálený, ani vyvrátený.

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

Editorial Expression of Concern

Science pripojil k práci redakčné vyjadrenie obáv, kým hodnotí otázky spracovania údajov a reprodukovateľnosti. Nejde o stiahnutie článku ani zistenie pochybenia; znamená to, že uvedený výsledok treba do uzavretia hodnotenia považovať za predbežný.

Editorial Expression of Concern →

Fakty predsa len — a varovná značka pri práci

V roku 2024 vyšla štúdia, ktorá odporuje pohodlnej súčasti bežného názoru. Nechajte človeka absolvovať krátky trojkolový rozhovor s umelou inteligenciou — GPT-4 Turbo, ktoré dostalo pokyn odpovedať na konkrétne dôkazy, ktorými dotyčný podporuje konšpiračnú teóriu, ktorej sám verí — a jeho viera v túto teóriu klesne v priemere približne o 20 %. Pokles bol viditeľný aj o dva mesiace neskôr. Fungovalo to pri klasických konšpiráciách (atentát na Johna F. Kennedyho, mimozemšťania, illumináti) aj súčasných (COVID-19, americké voľby v roku 2020), dokonca pri ľuďoch so silnou vierou spojenou s vlastnou identitou. Keď profesionálny overovateľ faktov ohodnotil 128 tvrdení AI, 127 bolo pravdivých, jedno zavádzajúce a žiadne nepravdivé.

Rozšíril sa titulok, že ľudí *možno* z králičej nory vyviesť rozhovorom — že zástancovia konšpirácií nie sú mimo dosahu dôkazov, iba potrebujú správne dôkazy podané dostatočne konkrétne. Je to skutočne zaujímavé tvrdenie a štúdia bola navrhnutá starostlivejšie než väčšina výskumu presviedčania.

V júni 2026 však časopis *Science* k práci pripojil **redakčné vyjadrenie obáv** (Editorial Expression of Concern). Práve túto časť väčšina prerozprávaní vynechá, hoci rozhoduje o tom, akú váhu dnes výsledok unesie.

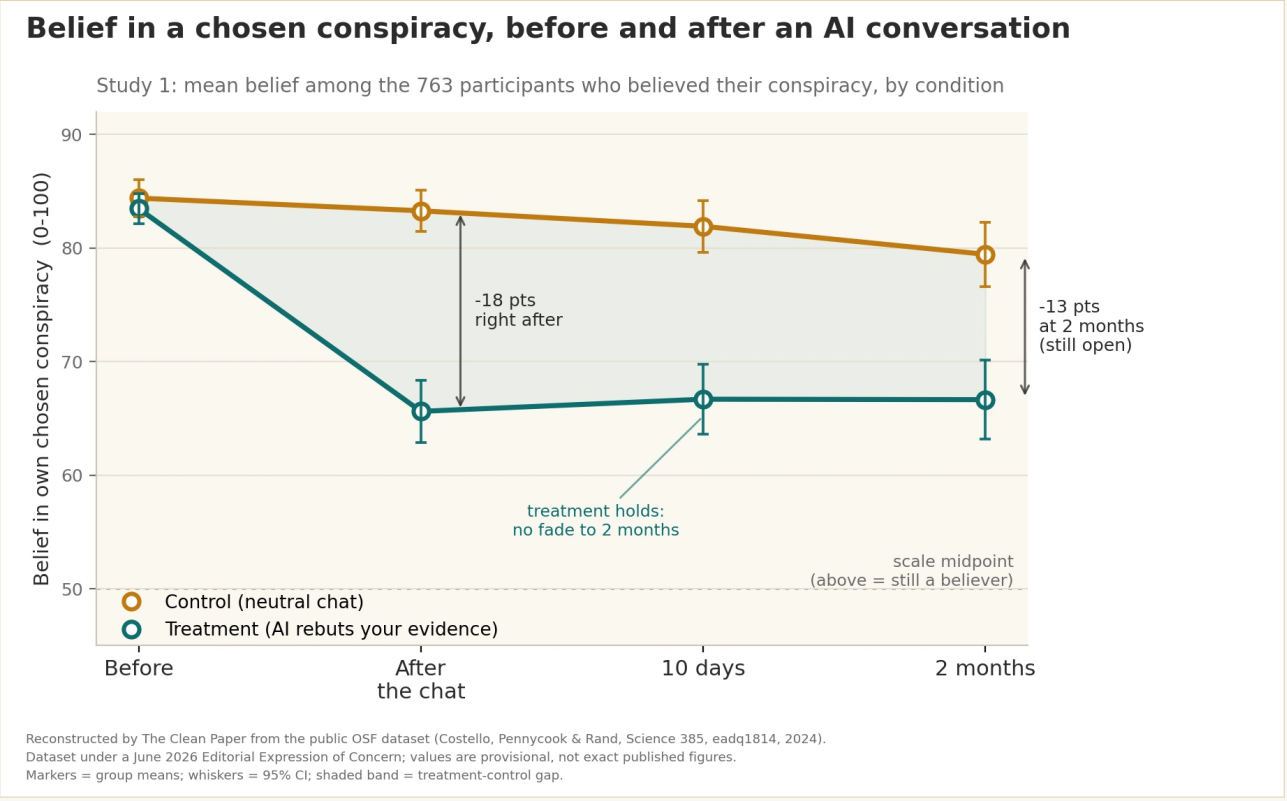
Čo je — a nie je — redakčné vyjadrenie obáv

Redakčné vyjadrenie obáv je formálne upozornenie, ktoré časopis pripojí k práci, aby čitateľov informoval o vznesených otázkach, kým sa tieto otázky riešia. **Nie je** to stiahnutie článku a samo osebe ani zistenie pochybenia. Znamená: čítajte s touto výhradou, pretože vedecký záznam sa môže zmeniť. V tomto prípade autori po upozornení na problémy vykonali šetrenie, oznámili podrobnosti časopisu *Science* a dodali opravenú analýzu.

Vznesené výhrady sú konkrétne. Autori boli upozornení na nezrovnalosti medzi rukopisom a zverejneným analytickým kódom v spôsobe použitia kritérií výberu účastníkov a na to, že verejný súbor údajov obsahoval nadbytočné riadky vložené chybou pri spájaní kódu — problémy, pre ktoré bolo ťažké reprodukovať niektoré uvedené čísla zo zverejnených materiálov. Autori poskytli *Science* opravený analytický postup a súbor aktualizovaných výsledkov, ktoré podľa nich zodpovedajú pôvodným **smerom, štatistickou významnosťou aj vecnou veľkosťou**. *Science* ich vyhodnocuje. Kým hodnotenie neskončí, visí nad presnými hodnotami otáznik, ktorý môže odstrániť iba časopis.

Sú to teda dva príbehy súčasne: zaujímavý výsledok o tom, či fakty dokážu pohnúť zakoreneným

presvedčením, a živý príklad toho, ako sa vedecký záznam verejne opravuje. Zmyslom tohto článku je oba príbehy nezameniť.



V štúdii mal trojkolový rozhovor, v ktorom AI vyvracala dôkazy uvedené samotným účastníkom, znižovať vieru v ním zvolenú konšpiráciu v priemere približne o 20 %; na rozdiel od kontrolného rozhovoru o nesúvisiacej téme bol pokles merateľný aj po 10 dňoch a 2 mesiacoch. Uvádzaný účinok bol trvalý, ale čiastočný: väčšina účastníkov konšpirácii verila aj potom — išlo o oslabenie, nie liečbu. Práca je teraz označená redakčným vyjadrením obáv (*Science*, jún 2026) pre nezrovnalosti vo vykazovaní a chybu vo verejnom súbore údajov; autori uvádzajú, že opravená analýza zachováva smer, významnosť aj veľkosť účinku, zatiaľ čo *Science* pokračuje v hodnotení. Graf vytvoril The Clean Paper z verejného súboru údajov, takže zobrazené hodnoty sú predbežné, nie konečné čísla z práce.

Čo autori urobili

- Vykonali dva experimenty s 2190 účastníkmi, z ktorých každý vlastnými slovami pomenoval konšpiračnú teóriu, ktorej verí, a dôkazy, ktoré ju podľa neho podporujú.
- Každý účastník viedol trojkolový rozhovor s GPT-4 Turbo. V **experimentálnej** podmienke mala AI vyvracať konkrétne dôkazy účastníka; v **kontrolnej** podmienke sa rozprávala o nesúvisiacej téme.
- Zmerali vieru vo zvolenú konšpiráciu pred rozhovorom a bezprostredne po ňom a potom účastníkov znovu kontaktovali po **10 dňoch a 2 mesiacoch**.
- Profesionálny overovateľ faktov posúdil presnosť všetkých 128 faktických tvrdení, ktoré AI vo vzorke uviedla.

- Overili, či sa účinok prenáša na iné, nesúvisiace konšpirácie — a ako test špecifickosti aj to, či znižuje vieru v konšpirácie, ktoré sú skutočne **pravdivé**.

Čo zistili

- **Priemerný pokles približne o 20 %** vo viere vo zvolenú konšpiráciu v experimentálnej skupine oproti kontrolnej (štúdia 1: 95 % interval spoľahlivosti [13,8; 19,7] bodu na 100-bodovej škále, $P < 0,001$).
- **Účinok pretrval.** V experimentálnej skupine pokles nevyprchal: po 10 dňoch a dvoch mesiacoch zostala viera blízko úrovne bezprostredne po rozhovore namiesto návratu, zatiaľ čo kontrolná skupina zostala celý čas vyššie. Práca opisuje účinok na zvolenú konšpiráciu ako neoslabený najmenej dva mesiace, nie ako chvíľkové zakolísanie.
- **Účinok sa zovšeobecňoval** naprieč konšpiráciami, ktoré ľudia uvádzali, a platil aj pre účastníkov s pôvodne silnou vierou.
- **Tvrdenia AI boli v tomto prostredí presné:** 127 zo 128 (99,2 %) bolo hodnotených ako pravdivých, 1 (0,8 %) ako zavádzajúce a žiadne ako nepravdivé.
- **Účinok bol špecifický**, nie plošným pochybovaním: intervencia **neznížila** vieru v pravdivé konšpirácie, čo naznačuje, že posúvala nepodložené presvedčenia namiesto vytvárania všeobecného cynizmu.
- **Mierne sa preniesol aj inde** — viera v ďalšie nesúvisiace konšpirácie klesla asi o 12 % a účastníci uvádzali väčší úmysel odporovať konšpiračným tvrdeniam. Hlavný účinok sa potvrdil v štúdii 2 a ukazoval rovnakým smerom aj v menšej vzorke bez kontrolnej skupiny (slabší, ale súhlasný dôkaz).

Čo to nedokazuje

- Práca teraz **nie je bez otázok**. Je označená redakčným vyjadrením obáv pre problémy so spracovaním údajov a reprodukovateľnosťou a opravené čísla stále vyhodnocuje *Science*. Konkrétne hodnoty treba do uzavretia hodnotenia považovať za predbežné.
- Výsledok **neukazuje**, že AI „lieči“ vieru v konšpirácie. Priemerný pokles približne o 20 % je významný posun, nie vymazanie — väčšina účastníkov teórii naďalej verila, len menej.
- **Nejde** o výsledok nasadenia v reálnom svete. Účastníci sa dobrovoľne zapojili do štúdie a konali s AI v dobrej viere; v bežnom prostredí majú ľudia najpevnejšie oddaní konšpirácii najmenšiu motiváciu vyhľadať chatbota, ktorý sa s nimi bude sporiť.
- Presnosť 99,2 % je **špecifická pre túto úlohu, model a starostlivé zadanie** — nie je všeobecnou zárukou, že jazykové modely uvádzajú pravdu. Autori výslovne hovoria, že rovnaká personalizovaná presviedčacia sila by mohla posilniť *nepravdivé* presvedčenie, keby na to bola zameraná.
- „Trvalý“ tu znamená **dva mesiace**, čo je po jedinom rozhovore pôsobivé, ale nie je to to isté ako navždy.

Aké silné sú dôkazy

- **Návrh je skutočnou prednosťou.** Kontrolované experimenty s experimentálnou a kontrolnou skupinou, čistá kontrola podobná placebu, zopakovanie v dvoch štúdiách aj nezávislej vzorke, následné merania trvácnosti a výslovná kontrola presnosti vlastných tvrdení AI — to je starostlivejšie než väčšina výskumu presviedčania a smer účinku je konzistentný všade, kde sa testoval.
- **Redakčné vyjadrenie obáv však nie je poznámka pod čiarou.** Schopnosť znovu vytvoriť uvedené čísla zo zverejnených údajov a kódu patrí k dôveryhodnosti výsledku a práve tá zlyhala — pre nezrovnalosti vo výberových kritériách a duplicitné riadky z chyby pri spájaní kódu. Tvrdenie autorov, že opravený postup výsledok zachováva, je povzbudivé a vierohodné, no ide o ich vlastný opis, nie nezávislý verdikt. Treba počkať na hodnotenie *Science*.
- **Poctivý stav je „označené výhradou“, nie „vyvrátené“ ani „potvrdené“.** Dobre vystavaná, nápadná štúdia, ktorej presné čísla sa formálne preverujú. Je nepohodlné nechať dobrý príbeh práve tu, ale je to presné.

Prečo na tom záleží

Roky prevládal názor, že zástancov konšpirácií poháňajú psychologické potreby a identita, a preto ich z presvedčenia nemožno vyviesť faktami. Trvalé oslabenie založené na dôkazoch — ak obstojí — je významnou protiváhou tohto pesimizmu: naznačuje, že problém sčasti spočíval v tom, že všeobecné vyvracanie nikdy nereagovalo na *konkrétne* dôkazy, ktoré veriaci uvádza, zatiaľ čo jazykový model to dokáže vo veľkom rozsahu.

Výsledok je dôležitý aj ako prípadová štúdia správneho fungovania vedy. Poučenie z redakčného vyjadrenia obáv nie je, že známe výsledky sú bezcenné; chyby vychádzajú na povrch, údaje sa znovu skúmajú a tvrdenia sa verejne overujú. To, že tento mechanizmus funguje, je prednosť, nie škandál — a preto je správnym postojom k výsledku trpezlivosť namiesto okamžitého potlesku alebo odmietnutia.

Pri AI je navyše ostrie obojstranné. Rovnaká schopnosť oslabiť nepravdivé presvedčenie argumentom na mieru by iné mohla rovnako účinne posilniť. Technika je neutrálna; cieľ nie.

Čisté zhrnutie

Štúdia z roku 2024 zistila, že trojkolový rozhovor s GPT-4 Turbo znížil vieru ľudí v nimi zastávanú konšpiračnú teóriu približne o 20 %, pričom účinok vydržal dva mesiace a zovšeobecňoval sa naprieč typmi konšpirácií a faktické tvrdenia AI boli hodnotené ako takmer výhradne presné. V júni 2026 bola práca označená redakčným vyjadrením obáv pre problémy so spracovaním údajov a reprodukovateľnosťou; autori uvádzajú, že opravená analýza zachováva smer, významnosť aj veľkosť výsledku, a *Science* ho naďalej hodnotí. Čítajte ju ako skutočne zaujímavý, starostlivo vystavaný výsledok, ktorý sa práve preveruje — ani ustálený, ani vyvrátený. Najpoctivejšiu vetu píše samotný proces: skontrolujme

to znova.

Zdroje

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