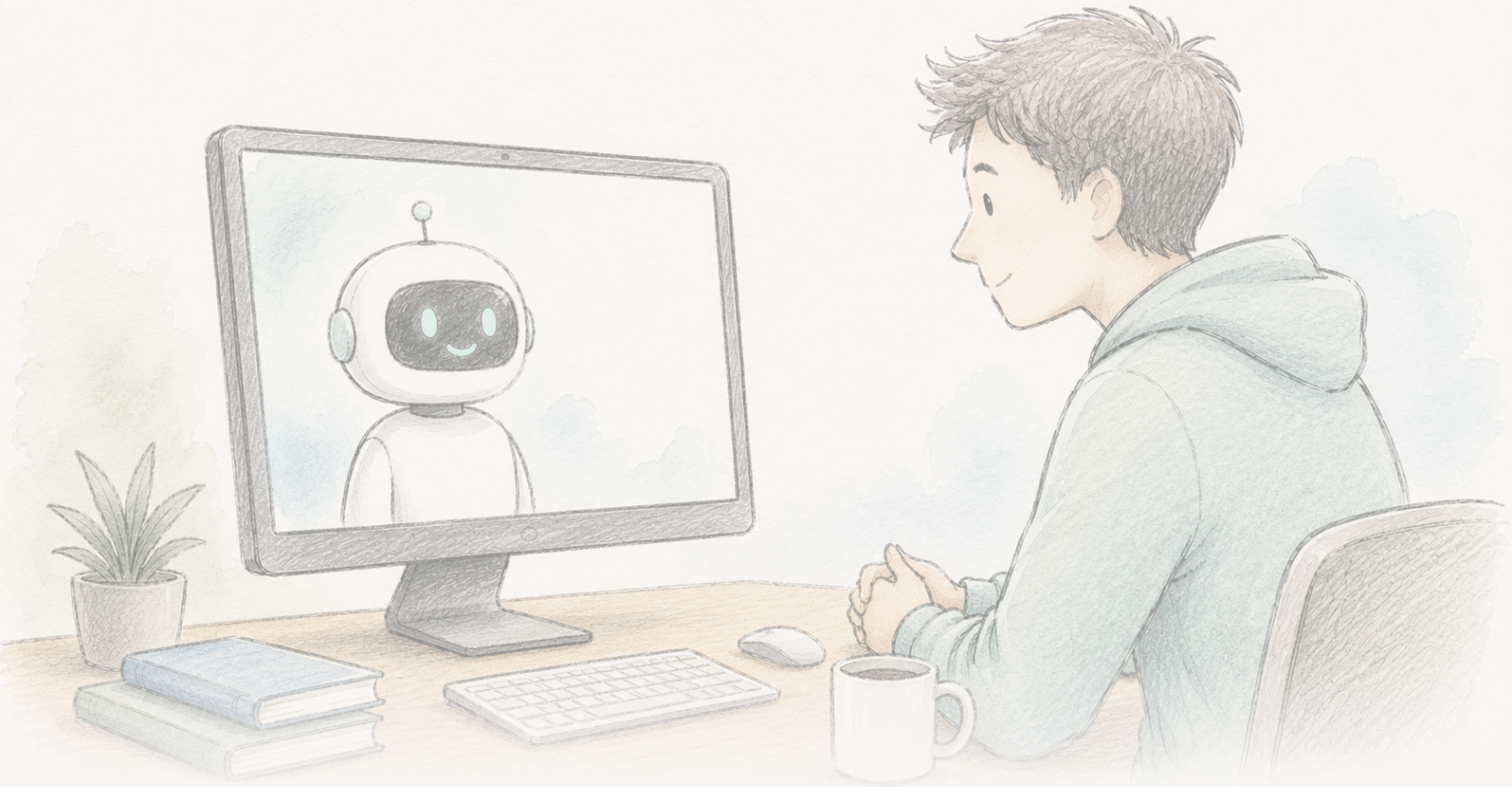


The Clean Paper

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

Chatbot zřejmě na celé měsíce oslaboval víru v konspirace

Studie z roku 2024 zjistila, že tříkolový rozhovor s GPT-4 Turbo snížil víru lidí v jimi zastávanou konspirační teorii asi o 20 %; účinek trval dva měsíce, zobecňoval se napříč typy konspirací a opíral se o tvrzení AI hodnocená jako téměř výhradně přesná. V červnu 2026 byla práce opatřena redakčním vyjádřením obav kvůli problémům se zpracováním dat a reprodukovatelností; autoři tvrdí, že opravená analýza zachovává směr, významnost i velikost výsledku, a Science jej nadále hodnotí. Skutečně zajímavý, pečlivě vystavěný výsledek — nyní prověřovaný, ani ustálený, ani vyvrácený.

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 připojil k práci redakční vyjádření obav, zatímco hodnotí otázky zpracování dat a reprodukovatelnosti. Nejde o stažení článku ani zjištění pochybení; znamená to, že uvedený výsledek je třeba do uzavření hodnocení považovat za předběžný.

Editorial Expression of Concern →

Fakta přece jen — a varovná značka u práce

V roce 2024 vyšla studie, která jde proti pohodlné součásti běžného mínění. Nechte člověka absolvovat krátký rozhovor o třech kolech s umělou inteligencí — GPT-4 Turbo, které dostalo pokyn odpovědět na konkrétní důkazy, jimiž dotyčný podporuje konspirační teorii, které sám věří — a jeho víra v tuto teorii klesne v průměru asi o 20 %. Pokles byl patrný i o dva měsíce později. Fungovalo to u klasických konspirací (atentát na Johna F. Kennedyho, mimozemšťané, iluminláti) i současných (COVID-19, americké volby v roce 2020), dokonce u lidí se silnou vírou spojenou s vlastní identitou. Když profesionální ověřovatel faktů ohodnotil 128 tvrzení AI, 127 bylo pravdivých, jedno zavádějící a žádné nepravdivé.

Rozšířil se titulek, že lidi *lze* z králičí nory vyvést rozhovorem — že zastánci konspirací nejsou mimo dosah důkazů, jen potřebují správné důkazy podané dostatečně konkrétně. Je to skutečně zajímavé tvrzení a studie byla navržena pečlivěji než většina výzkumu přesvědčování.

V červnu 2026 však časopis *Science* k práci připojil **redakční vyjádření obav** (Editorial Expression of Concern). Právě tuto část většina převyprávění vynechá, a přitom rozhoduje o tom, jakou váhu dnes výsledek unese.

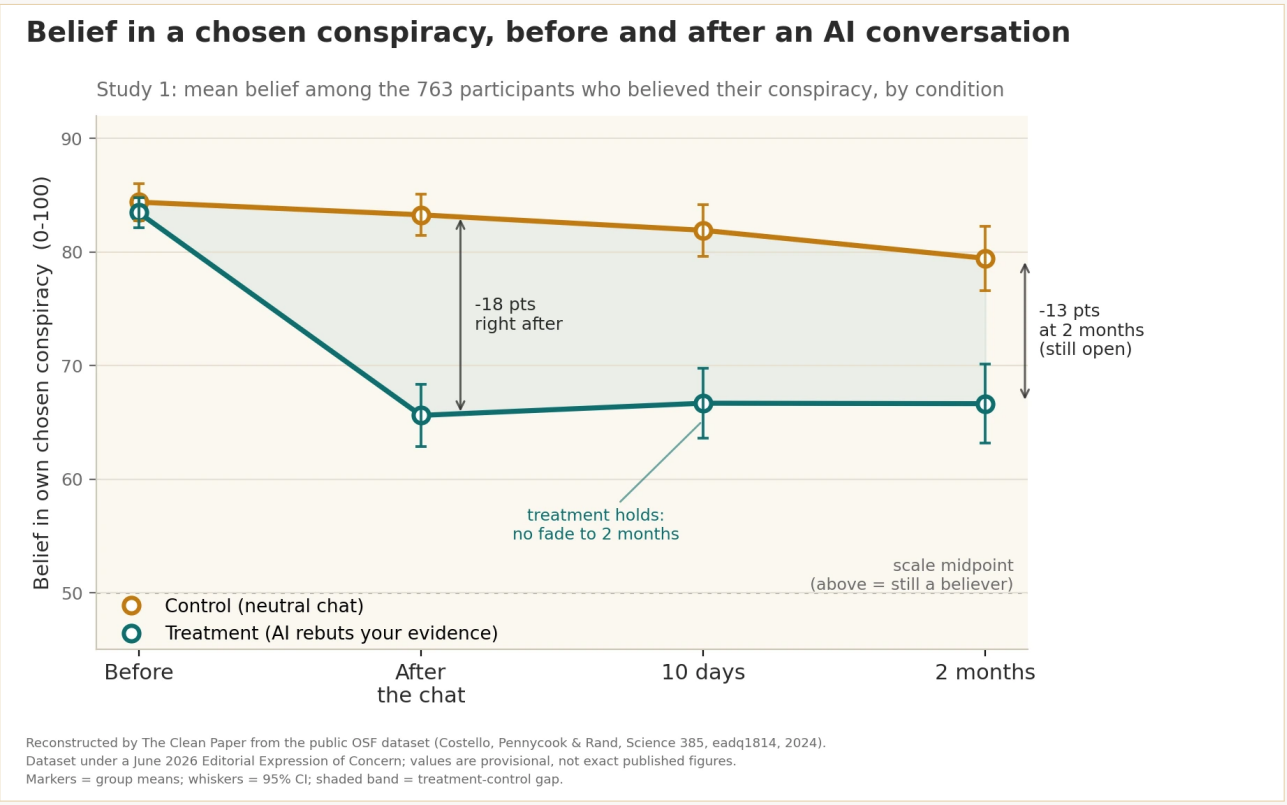
Co je — a není — redakční vyjádření obav

Redakční vyjádření obav je formální upozornění, které časopis připojí k práci, aby čtenáře informoval o vznesených otázkách, dokud se tyto otázky řeší. **Není** to stažení článku a samo o sobě ani zjištění pochybení. Znamená: čtěte s touto výhradou, protože se vědecký záznam může změnit. V tomto případě autoři po upozornění na problémy provedli šetření, sdělili podrobnosti časopisu *Science* a dodali opravenou analýzu.

Vznesené výhrady jsou konkrétní. Autoři byli upozorněni na nesrovnalosti mezi rukopisem a zveřejněným analytickým kódem ve způsobu použití kritérií výběru účastníků a na to, že veřejný soubor dat obsahoval nadbytečné řádky vložené chybou při spojování kódu — problémy, kvůli nimž bylo obtížné reprodukovat některá uvedená čísla ze zveřejněných materiálů. Autoři poskytli *Science* opravený analytický postup a soubor aktualizovaných výsledků, které podle nich odpovídají původním **směrem, statistickou významností i věcnou velikostí**. *Science* je vyhodnocuje. Dokud hodnocení neskončí, visí nad přesnými hodnotami otazník, který může odstranit pouze časopis.

Jsou to tedy dva příběhy současně: zajímavý výsledek o tom, zda fakta dokážou pohnout zakořeněným

přesvědčením, a živý příklad toho, jak se vědecký záznam veřejně opravuje. Smyslem tohoto článku je oba příběhy nezaměnit.



Ve studii měl rozhovor se třemi koly, v němž AI vyvracela důkazy uvedené samotným účastníkem, snižovat víru v jím zvolenou konspiraci v průměru asi o 20 %; na rozdíl od kontrolního rozhovoru o nesouvisejícím tématu byl pokles měřitelný i po 10 dnech a 2 měsících. Uváděný účinek byl trvalý, ale částečný: většina účastníků konspiraci věřila i poté — šlo o oslabení, ne léčbu. Práce je nyní opatřena redakčním vyjádřením obav (*Science*, červen 2026) kvůli nesrovnalostem ve vykazování a chybě ve veřejném souboru dat; autoři uvádějí, že opravená analýza zachovává směr, významnost i velikost účinku, zatímco *Science* pokračuje v hodnocení. Graf vytvořil The Clean Paper z veřejného souboru dat, takže zobrazené hodnoty jsou předběžné, nikoli konečná čísla z práce.

Co autoři udělali

- Provedli dva experimenty s 2190 účastníky, z nichž každý vlastními slovy pojmenoval konspirační teorii, které věří, a důkazy, jež ji podle něj podporují.
- Každý účastník vedl tříkolový rozhovor s GPT-4 Turbo. V **experimentální** podmínce měla AI vyvracet konkrétní důkazy účastníka; v **kontrolní** podmínce se bavila o nesouvisejícím tématu.
- Změřili víru ve zvolenou konspiraci před rozhovorem a bezprostředně po něm a poté účastníky znovu kontaktovali po **10 dnech a 2 měsících**.
- Profesionální ověřovatel faktů posoudil přesnost všech 128 faktických tvrzení, která AI ve vzorku uvedla.

- Ověřili, zda se účinek přenáší na jiné, nesouvisející konspirace — a jako test specifičnosti také zda snižuje víru v konspirace, které jsou skutečně **pravdivé**.

Co zjistili

- **Průměrný pokles přibližně o 20 %** ve víře ve zvolenou konspiraci v experimentální skupině oproti kontrolní (studie 1: 95% interval spolehlivosti [13,8; 19,7] bodu na 100bodové škále, P < 0,001).
- **Účinek přetrval.** V experimentální skupině pokles nevyprchal: po 10 dnech a dvou měsících zůstala víra blízko úrovně bezprostředně po rozhovoru, místo aby se vracela, zatímco kontrolní skupina zůstala po celou dobu výše. Práce popisuje účinek na zvolenou konspiraci jako neoslabený nejméně po dva měsíce, nikoli jako chvilkové zakolísání.
- **Účinek se zobecňoval** napříč konspiracemi, které lidé uváděli, a platil i pro účastníky s původně silnou vírou.
- **Tvrzení AI byla v tomto prostředí přesná:** 127 ze 128 (99,2 %) bylo hodnoceno jako pravdivých, 1 (0,8 %) jako zavádějící a žádné jako nepravdivé.
- **Účinek byl specifický**, nikoli plošným pochybováním: intervence **nesnížila** víru v pravdivé konspirace, což naznačuje, že posouvala nepodložená přesvědčení, místo aby z lidí činila cyniky vůči všemu.
- **Mírně se přenesl jinam** — víra v další nesouvisející konspirace klesla asi o 12 % a účastníci uváděli větší úmysl odporovat konspiračním tvrzením. Hlavní účinek se potvrdil ve studii 2 a ukazoval stejným směrem i v menším vzorku bez kontrolní skupiny (slabší, ale souhlasný důkaz).

Co to nedokazuje

- Práce nyní **není bez otázek**. Je opatřena redakčním vyjádřením obav kvůli problémům se zpracováním dat a reprodukovatelností a opravená čísla stále vyhodnocuje *Science*. Konkrétní hodnoty je třeba do uzavření hodnocení považovat za předběžné.
- Výsledek **neukazuje**, že AI „léčí“ víru v konspirace. Průměrný pokles přibližně o 20 % je významný posun, nikoli vymazání — většina účastníků teorii nadále věřila, jen méně.
- **Nejde** o výsledek nasazení v reálném světě. Účastníci se dobrovolně zapojili do studie a jednali s AI v dobré víře; v běžném prostředí mají lidé nejpevněji oddaní konspiraci nejmenší motivaci vyhledat chatbota, který se s nimi bude přít.
- Přesnost 99,2 % je **specifická pro tento úkol, model a pečlivé zadání** — není obecnou zárukou, že jazykové modely uvádějí pravdu. Autoři výslovně říkají, že stejná personalizovaná přesvědčovací síla by mohla posílit *nepravdivé* přesvědčení, kdyby k tomu byla zaměřena.
- „Trvalý“ zde znamená **dva měsíce**, což je po jediném rozhovoru působivé, ale není to totéž co navždy.

Jak silné jsou důkazy

- **Návrh je skutečnou předností.** Kontrolované experimenty s experimentální a kontrolní skupinou, dobře navržená kontrolní podmínka podobná placebu, opakování ve dvou studiích i nezávislém vzorku, následná měření trvanlivosti a výslovná kontrola přesnosti vlastních tvrzení AI — to je pečlivější než většina výzkumu přesvědčování a směr účinku je konzistentní všude, kde se testoval.
- **Redakční vyjádření obav však není poznámka pod čarou.** Schopnost znovu vytvořit uvedená čísla ze zveřejněných dat a kódu patří k důvěryhodnosti výsledku, a právě ta selhala — kvůli nesrovnalostem ve výběrových kritériích a duplikovaným řádkům z chyby při spojování kódu. Tvrzení autorů, že opravený postup výsledek zachovává, je povzbudivé a věrohodné, jde však o jejich vlastní popis, nikoli nezávislý verdikt. Je třeba počkat na hodnocení *Science*.
- **Poctivý stav je „označeno výhradou“, nikoli „vyvráceno“ nebo „potvrzeno“.** Dobře vystavěná, nápadná studie, jejíž přesná čísla se formálně prověřují. Je nepohodlné nechat dobrý příběh právě zde, ale je to přesné.

Proč na tom záleží

Po léta převládal názor, že zastánce konspirací pohánějí psychologické potřeby a identita, a proto je z jejich přesvědčení nelze vyvést fakty. Trvalé oslabení založené na důkazech — pokud obstojí — je významnou protiváhou tohoto pesimismu: naznačuje, že problém částečně spočíval v tom, že obecná vyvrácení nikdy nereagovala na *konkrétní* důkazy, které věřící uvádí, zatímco jazykový model to dokáže ve velkém měřítku.

Výsledek je důležitý také jako případová studie správného fungování vědy. Poučení z redakčního vyjádření obav není, že známé výsledky jsou bezcenné; chyby vycházejí na povrch, data se znovu zkoumají a tvrzení se veřejně ověřují. To, že tento mechanismus běží, je přednost, ne skandál — a proto je správným postojem k výsledku trpělivost místo okamžitého potlesku či odmítnutí.

U AI je navíc ostří oboustranné. Stejná schopnost oslabit nepravdivé přesvědčení argumentem na míru by jiné mohla stejně účinně posílit. Technika je neutrální; cíl nikoli.

Stručné shrnutí

Studie z roku 2024 zjistila, že tříkolový rozhovor s GPT-4 Turbo snížil víru lidí v jimi zastávanou konspirační teorii asi o 20 %, přičemž účinek vydržel dva měsíce a zobecňoval se napříč typy konspirací a faktická tvrzení AI byla hodnocena jako téměř výhradně přesná. V červnu 2026 byla práce opatřena redakčním vyjádřením obav kvůli problémům se zpracováním dat a reprodukovatelností; autoři uvádějí, že opravená analýza zachovává směr, významnost i velikost výsledku, a *Science* jej nadále hodnotí. Čtěte ji jako skutečně zajímavý, pečlivě vystavěný výsledek, který se právě prověřuje — ani ustálený, ani vyvrácený. Nejpoctivější větu píše sám proces: zkontrolujme to znovu.

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