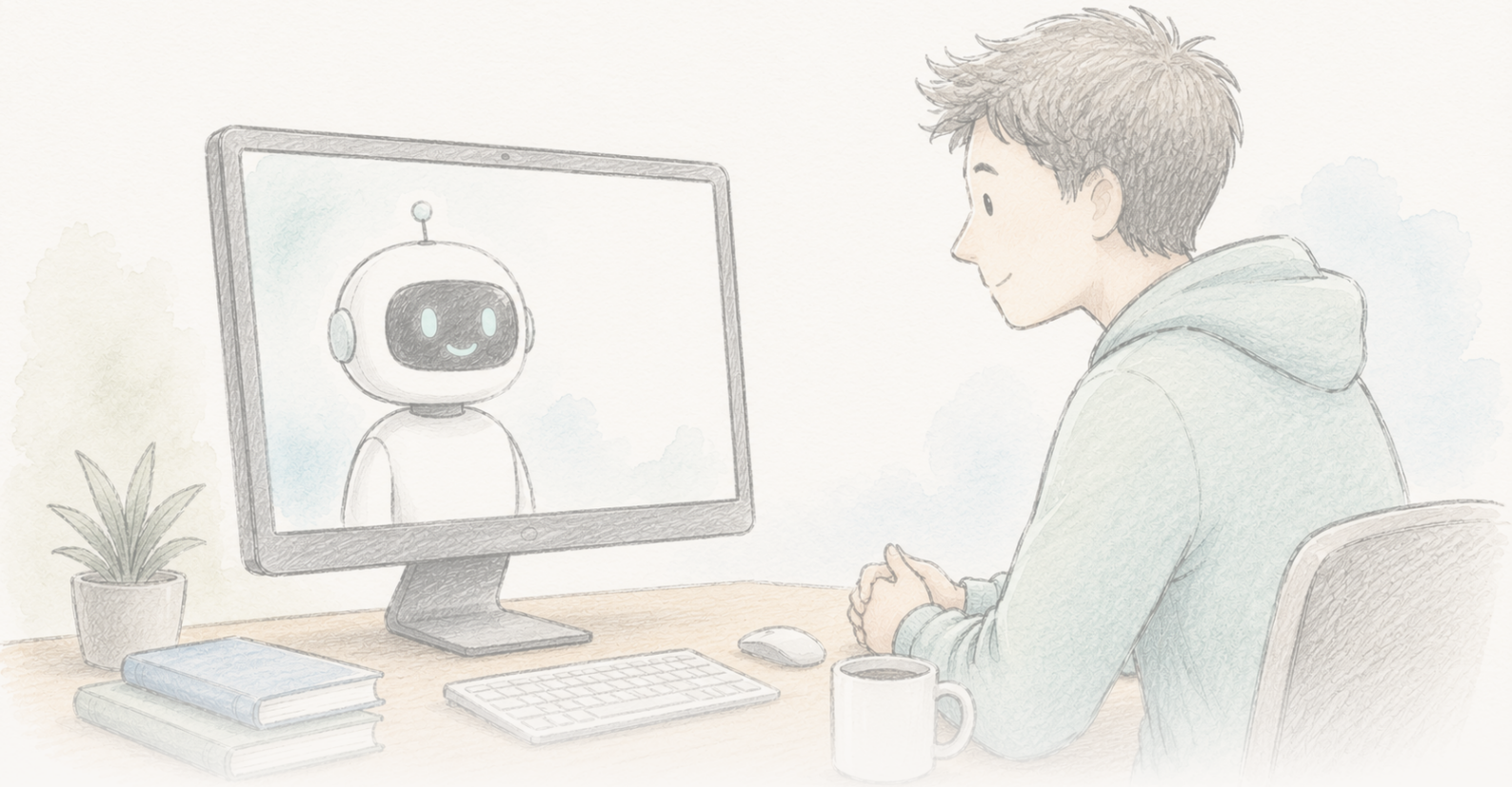


The Clean Paper

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

Egy chatbot hónapokra csökkenteni látszott az összeesküvéses hiedelmeket

Egy 2024-es tanulmány szerint egy háromfordulós GPT-4 Turbo beszélgetés körülbelül 20%-kal csökkentette az emberek által vallott összeesküvés-elméletbe vetett hitet; a hatás két hónapig fennmaradt, különböző elméletekre általánosult, és az MI állításait túlnyomórészt pontosnak értékelték. 2026 júniusában a tanulmány adatkezelési és reprodukálhatósági problémák miatt szerkesztői aggodalomkifejezést kapott; a szerzők szerint a javított elemzés megőrzi az eredmény irányát, szignifikanciáját és nagyságát, a Science pedig még értékeli. Valóban érdekes, gondosan felépített eredmény — jelenleg felülvizsgálat alatt, sem nem lezárva, sem nem megcáfolva.

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

A Science szerkesztői aggodalomkifejezést csatolt ehhez a tanulmányhoz, miközben adatkezelési és reprodukálhatósági kérdéseket értékel. Ez nem visszavonás és nem kutatási visszaélés megállapítása; azt jelenti, hogy az eredményt előzetesként kell olvasni, amíg a felülvizsgálat le nem zárul.

Editorial Expression of Concern →

Mégiscsak tények — és egy figyelmeztető jelzés a tanulmányon

Egy 2024-es tanulmány olyasmiről számolt be, ami szembemegy egy kényelmes, elterjedt nézettel. Ha valaki rövid, háromfordulós beszélgetést folytat egy MI-vel — GPT-4 Turbóval, amelyet arra kérnek, hogy az illető által személyesen hitt összeesküvés-elmélet mellett felhozott konkrét bizonyítékokra válaszoljon —, akkor átlagosan körülbelül 20%-kal csökken az elméletbe vetett hite. A csökkenés két hónappal később is megmaradt. Működött klasszikus összeesküvés-elméleteknél (John F. Kennedy meggyilkolása, földönkívüliek, Illuminátusok) és aktuálisaknál (COVID-19, a 2020-as amerikai választás), még olyan embereknél is, akiknek a hite erős volt és identitásukhoz kapcsolódott. Amikor egy hivatásos tényellenőr az MI által tett állítások közül 128-at értékelt, 127 igaz volt, egy félrevezető, és egy sem hamis.

Az a cím terjedt el, hogy *ki lehet beszélni* az embereket a nyúlüregből — hogy az összeesküvés-hívők nincsenek a bizonyítékok hatókörén kívül, csak megfelelő, kellően személyre szabott bizonyítékot kell kapniuk. Ez valóban érdekes állítás, és a tanulmány gondosabban épült fel, mint a meggyőzéskutatások nagy része.

Aztán 2026 júniusában a *Science* **szerkesztői aggodalomkifejezést (Editorial Expression of Concern)** csatolt a tanulmányhoz. A legtöbb újramesélés ezt a részt kihagyja majd, pedig éppen ez dönti el, mekkora súlyt bír jelenleg az eredmény.

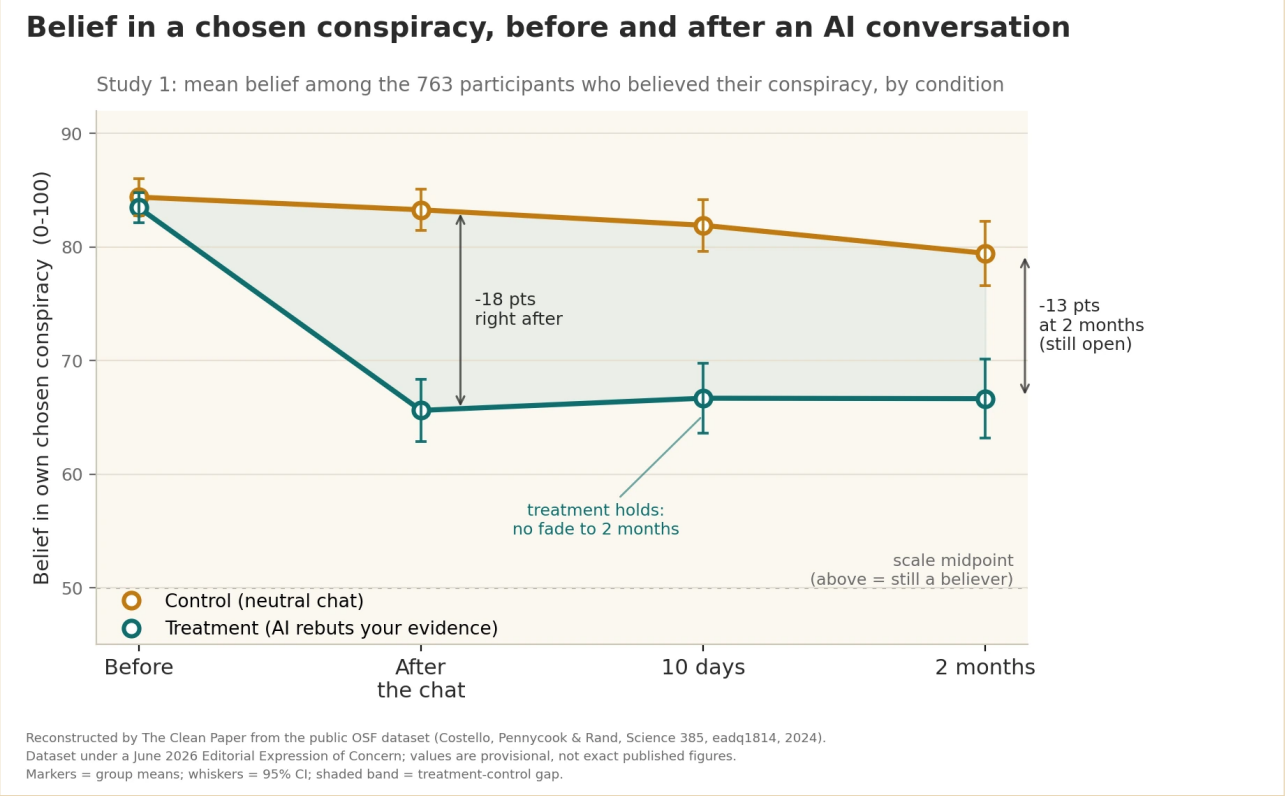
Mi az Expression of Concern — és mi nem az?

A szerkesztői aggodalomkifejezés egy hivatalos jelzés, amelyet a folyóirat akkor csatol egy tanulmányhoz, ha kérdések merültek fel vele kapcsolatban, de ezek vizsgálata még folyamatban van. **Nem** visszavonás — a cikket nem vonták vissza —, és önmagában nem jelent kutatási visszaélést. Azt üzeni: olvasd ezt a fenntartást szem előtt tartva, mert a tudományos nyilvántartás még változhat. Ebben az esetben a problémákról értesülő szerzők vizsgálatot végeztek, jelentették a részleteket a *Science*-nek, és javított elemzést adtak át.

A megjelölt problémák konkrétak. A szerzők tudomást szereztek arról, hogy a résztvevők szűrési kritériumait nem teljesen ugyanúgy alkalmazták a kéziratban leírtak és a publikált elemzőkód között, a nyilvános adathalmazba pedig egy kódösszefésülési hiba miatt oda nem tartozó, duplikált sorok

kerültek. Emiatt a közzétett anyagokból egyes közölt számokat nehéz volt reprodukálni. A szerzők javított elemzési folyamatot és frissített eredményeket adtak a *Science*-nek, amelyek állításuk szerint **irányukban, statisztikai szignifikanciájukban és érdemi nagyságukban** megegyeznek az eredetivel. A *Science* jelenleg értékeli ezeket. Amíg ez nem zárul le, a pontos számok fölött kérdőjel marad, amelyet csak a folyóirat tud levenni.

Ez tehát egyszerre két történet: érdekes eredmény arról, hogy a tények képesek-e elmozdítani mélyen rögzült meggyőződéseket, és élő példája annak, ahogy a tudományos nyilvántartás nyilvánosan korrigálja önmagát. Ennek a cikknek az a dolga, hogy a kettőt ne keverje össze.



A vizsgálatban az a háromfordulós MI-beszélgetés, amely a résztvevő saját maga által megadott bizonyítékait cáfolta, a beszámoló szerint átlagosan körülbelül 20%-kal csökkentette a választott összeesküvés-elméletbe vetett hitet; egy más témáról szóló kontrollbeszélgetéssel szemben a csökkenés 10 nap és 2 hónap múlva is mérhető maradt. A közölt hatás tartós, de részleges volt: a legtöbb résztvevő később is hitt az elméletben — csökkent a hit, nem tűnt el. A tanulmány jelenleg szerkesztői aggodalomkifejezés alatt áll (*Science*, 2026. június) jelentési következetlenségek és a nyilvános adathalmaz hibája miatt; a szerzők szerint a javított elemzés megőrzi a hatás irányát, szignifikanciáját és nagyságát, miközben a *Science* tovább értékeli az anyagot. Ezt a grafikont a The Clean Paper rekonstruálta a nyilvános adathalmazból, ezért az értékek előzetesek, nem a cikk végleges számai.

Original figure — The Clean Paper CC BY 4.0

Mit csináltak a szerzők?

- Két kísérletet futtattak **2190 résztvevővel**, akik saját szavaikkal megneveztek egy összeesküvés-elméletet, amelyben hittek, és azokat a bizonyítékokat, amelyek szerintük alátámasztották.
- Minden résztvevő háromfordulós beszélgetést folytatott GPT-4 Turbóval. A **kezelési** feltételben az MI-t arra utasították, hogy a résztvevő konkrét bizonyítékait cáfolja; a **kontroll** feltételben egy nem kapcsolódó témáról beszélt.
- A választott összeesküvés-elméletbe vetett hitet a beszélgetés előtt és közvetlenül utána mérték, majd **10 nap és 2 hónap** múlva újra megkeresték a résztvevőket.
- Hivatásos tényellenőrrel értékeltették a mintában az MI által tett mind a 128 tényszerű állítás pontosságát.
- Megvizsgálták, átterjed-e a hatás más, nem kapcsolódó összeesküvés-elméletekre, és specifikussági tesztként azt is, csökkenti-e a véletlenül **igaznak bizonyuló** összeesküvésekbe vetett hitet.

Mit találtak?

- **Körülbelül 20%-os átlagos csökkenést** a választott összeesküvés-elméletbe vetett hitben a kezelési csoportban a kontrollhoz képest (1. vizsgálat: 95%-os konfidenciaintervallum [13.8, 19.7] pont egy 100 pontos skálán, $P < 0.001$).
- **A hatás megmaradt.** A kezelési csoportban a csökkenés nem tűnt el: 10 nap és két hónap múlva a hit közel maradt a közvetlenül beszélgetés utáni szinthez, miközben a kontrollcsoport végig magasabb maradt. A tanulmány úgy írja le, hogy a kiválasztott elméletre gyakorolt hatás legalább két hónapig lényegében változatlanul fennmaradt, nem csak pillanatnyi megingás volt.
- **Többféle összeesküvés-elméletre általánosult**, és azoknál a résztvevőknél is fennállt, akik kezdetben erősen hittek az adott elméletben.
- **Ebben a feladatban az MI állításai pontosak voltak:** 128-ból 127-et (99.2%) igaznak, 1-et (0.8%) félrevezetőnek értékeltek, hamisat egyet sem.
- **Specifikus volt**, nem általános kételkedés: a beavatkozás **nem** csökkentette az igaz összeesküvésekbe vetett hitet, ami arra utal, hogy a nem alátámasztott hiedelmeket mozdította el, nem egyszerűen minden iránt cinikussá tette az embereket.
- **Mérsékelten átterjedt más hiedelmekre is** — más, nem kapcsolódó összeesküvés-elméletekbe vetett hit körülbelül 12%-kal csökkent, és a résztvevők nagyobb hajlandóságot jeleztek arra, hogy ellenvéleményt mondjanak összeesküvés-állításokra. A fő hatás a második vizsgálatban is fennmaradt, és ugyanabba az irányba mutatott egy kisebb, kontrollcsoport nélküli mintában is — ez gyengébb, de konzisztens bizonyíték.

Mit nem bizonyít ez?

- Jelenleg **nem áll meg kérdések nélkül**. A cikk adatkezelési és reprodukálhatósági problémák miatt szerkesztői aggodalomkifejezés alatt áll, a javított számokat pedig a *Science* még értékeli. A konkrét értékeket addig előzetesnek kell tekinteni.
- **Nem** mutatja, hogy az MI „meggyógyítja" az összeesküvéses hitet. Körülbelül 20%-os átlagos csökkenés jelentős elmozdulás, nem eltörlés — a legtöbb résztvevő utána is hitt az elméletben, csak kevésbé.
- **Nem** valós világban bevezetett eredmény. A résztvevők önként jelentkeztek vizsgálatra, és jóhiszeműen foglalkoztak az MI-vel; a való életben éppen a legelkötelezettebb összeesküvés-hívők a legkevésbé valószínű, hogy olyan chatbotot keresnek, amely vitatkozni fog velük.
- A 99.2%-os pontosság **ehhez a feladathoz, modellhez és gondos promptoláshoz kötődik** — nem általános garancia arra, hogy a nyelvi modellek igaz állításokat tesznek. A szerzők kifejezetten megjegyzik, hogy ugyanaz a személyre szabott meggyőzőerő *hamis* hiteket is erősíthetne, ha arra irányítanák.
- A „tartós" itt **két hónapot** jelent, ami egyetlen beszélgetéshez képest figyelemre méltó, de nem azonos az állandóval.

Mennyire erős a bizonyíték?

- **A kísérleti terv valódi erősség.** Kontrollált kezelés-vs-kontroll kísérletek, tiszta placebojellegű kontroll, ismétlés két vizsgálatban és egy független mintában, tartóssági utánkövetések, valamint az MI állításainak explicit pontosságellenőrzése — ez gondosabb a meggyőzéskutatások nagy részénél, és a hatás iránya minden tesztben konzisztens volt.
- **De az Expression of Concern nem lábjegyzet.** A közölt számok reprodukálhatósága a kiadott adatokból és kódból része annak, amitől egy eredmény megbízhatóvá válik, és éppen ez sérült: szűrési kritériumok következetlenségei és kódösszefésülési hiba miatt bekerült duplikált sorok. Biztató és hihető, hogy a szerzők szerint a javított folyamat megőrzi az eredményt, de ez egyelőre a szerzők saját beszámolója, nem független ítélet. A *Science* értékelését kell megvárni.
- **Az őszinte státusz: „megjelölve", nem „megcáfolva" vagy „megerősítve".** Jól felépített, feltűnő tanulmány, amelynek pontos számai hivatalos felülvizsgálat alatt állnak. Kényelmetlen hely egy jó történet számára — és jelenleg ez a pontos hely.

Miért fontos?

Évekig befolyásos nézet volt, hogy az összeesküvés-hívőket pszichológiai szükségletek és identitás vezérli, ezért tényekkel nem lehet őket „kiérvelni" a hiedelmeikből. Ha tartós, bizonyítékalapú csökkenés valóban fennmarad az ellenőrzés után is, az fontos ellensúlya ennek a pesszimizmusnak: arra utal, hogy részben talán az volt a probléma, hogy az általános cáfolatok soha nem foglalkoztak azzal a *konkrét* bizonyítékkal, amelyet az adott ember valóban idéz — amit egy LLM nagy léptékben

képes megtenni.

A tudomány működésének esettanulmányaként is számít. A szerkesztői aggodalomkifejezés tanulsága nem az, hogy a sokat ünnepelt eredmények értéktelenek; hanem az, hogy a hibák felszínre kerülnek, az adatokat újravizsgálják, az állításokat nyilvánosan újraellenőrzik. Az, hogy ez a gépezet működik, tulajdonság, nem botrány — és ezért a helyes hozzáállás ehhez az eredményhez türelem, nem taps vagy elutasítás.

Az MI szempontjából pedig kétélű. Ugyanaz a képesség, amely személyre szabott érveléssel le tud ereszteni egy hamis hiedelmet, ugyanolyan hatékonyan fel is fújhat egyet. A technika semleges; a célja nem az.

Röviden

Egy 2024-es tanulmány szerint egy háromfordulós GPT-4 Turbo beszélgetés átlagosan körülbelül 20%-kal csökkentette az emberek által vallott összeesküvés-elméletbe vetett hitet; a hatás két hónapig fennmaradt, különböző összeesküvés-típusokra általánosult, és az MI tényszerű állításait túlnyomórészt pontosnak értékelték. 2026 júniusában a cikk adatkezelési és reprodukálhatósági problémák miatt szerkesztői aggodalomkifejezést kapott; a szerzők szerint a javított elemzés irányában, szignifikanciájában és nagyságában megőrzi az eredményt, a *Science* pedig még értékeli az anyagot. Olyan valóban érdekes, gondosan felépített eredményként érdemes olvasni, amely jelenleg felülvizsgálat alatt áll — nem lezárt és nem megcáfolt eredményként. A legőszintébb mondat róla az, amelyet maga a folyamat ír: ellenőrizzük újra.

Források

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