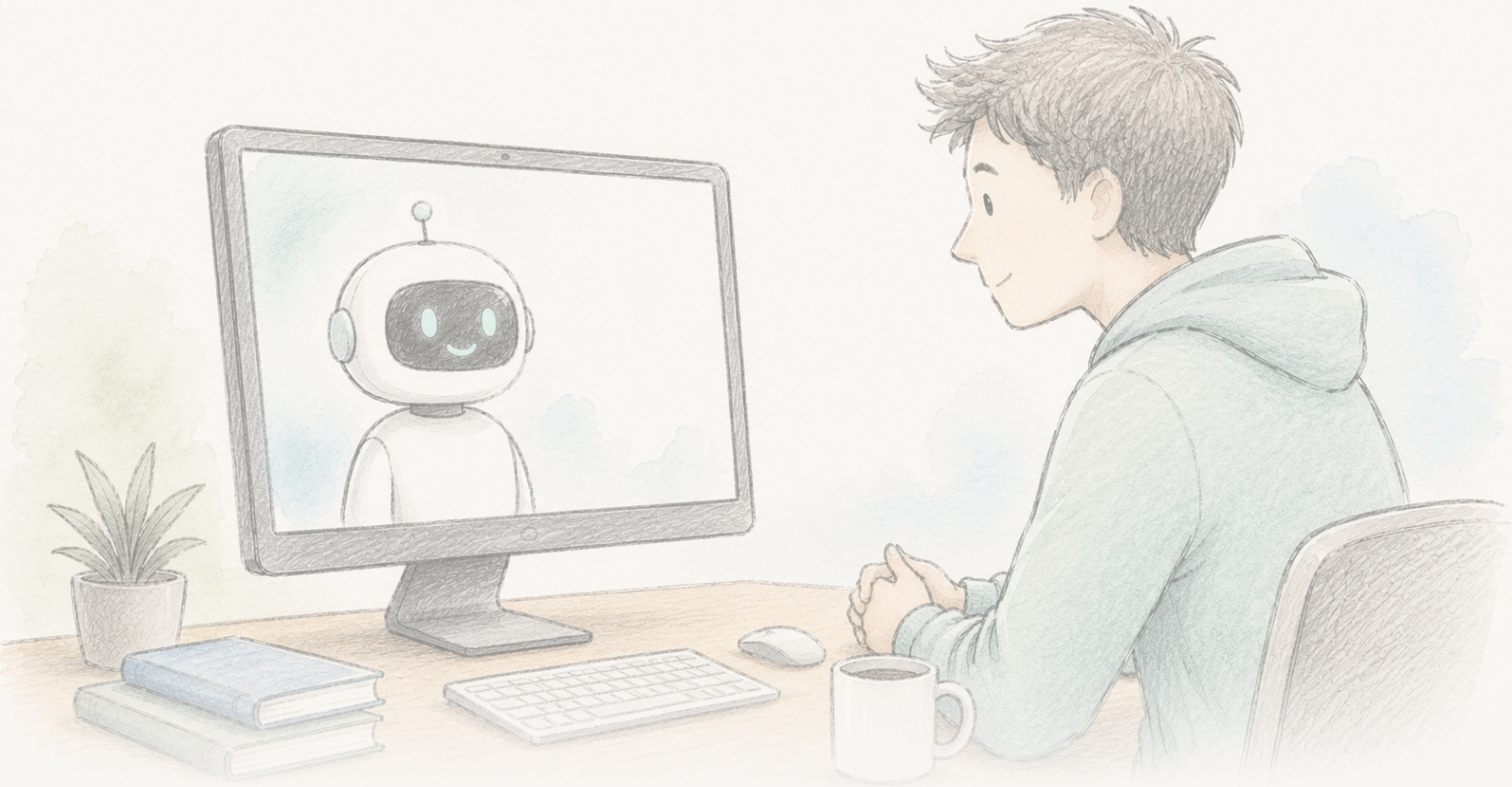


The Clean Paper

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

# Staðreyndir, þegar allt kemur til alls — og formleg viðvörun á greininni

*Rannsókn frá 2024 fann að þrjár umferðir samtals við GPT-4 Turbo minnkuðu trú þátttakenda á eigin samsæriskenningu um um 20%, og áhrifin héldust í tvo mánuði. Í júní 2026 fékk greinin Editorial Expression of Concern vegna spurninga um gagnavinnslu og endurtekningarhæfni; höfundar segja leiðrétta greiningu halda sömu stefnu og stærð, en Science er enn að meta málið. Niðurstaðan er áhugaverð en bráðabirgðakennd, hvorki staðfest til fulls né afsönnuð.*

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 hefur sett Editorial Expression of Concern við greinina á meðan tímaritið metur spurningar um gagnavinnslu og endurtekningarhæfni. Þetta er ekki afturköllun og ekki niðurstaða um misferli; það þýðir að lesa ber niðurstöðuna sem bráðabirgðaniðurstöðu þar til yfirferðinni lýkur.

Editorial Expression of Concern →

# Staðreyndir, þegar allt kemur til alls — og formleg viðvörun á greininni

Árið 2024 greindi rannsókn frá niðurstöðu sem gengur gegn þægilegri hefðbundinni hugmynd. Ef einstaklingur á stutt, þriggja umferða samtal við gervigreind — GPT-4 Turbo, leiðbeint um að svara nákvæmlega þeim gögnum sem viðkomandi nefnir fyrir samsæriskenningu sem hann sjálfur trúir — minnkar trúin á kenninguna að meðaltali um um 20%. Lækkunin var enn til staðar tveimur mánuðum síðar. Hún kom fram bæði fyrir klassískar samsæriskenningar (morðið á John F. Kennedy, geimverur, Illuminati) og nýlegar (COVID-19, bandarísku kosningarnar 2020), jafnvel hjá fólki með sterka og sjálfsmyndarlega tengda trú. Þegar faglegur staðreyndaprófari mat 128 staðhæfingar gervigreindarinnar voru 127 sannar, ein villandi og engin röng.

Fyrirsögnin sem fór víða var að *hægt sé* að tala fólk út úr samsærisheimum — að samsæriskenningafólk sé ekki ómóttækilegt fyrir gögnum, það þurfi bara réttu gögnin á nægilega sérsniðinn hátt. Það er í alvöru áhugaverð fullyrðing og rannsóknin var hönnuð af meiri varfærni en margar sannfæringarrannsóknir.

Síðan, í júní 2026, birti *Science* **Editorial Expression of Concern** við greinina. Þetta er sá hluti sem flest endursögn myndi sleppa og einmitt sá hluti sem ræður því hversu mikla þyngd niðurstaðan getur borið núna.

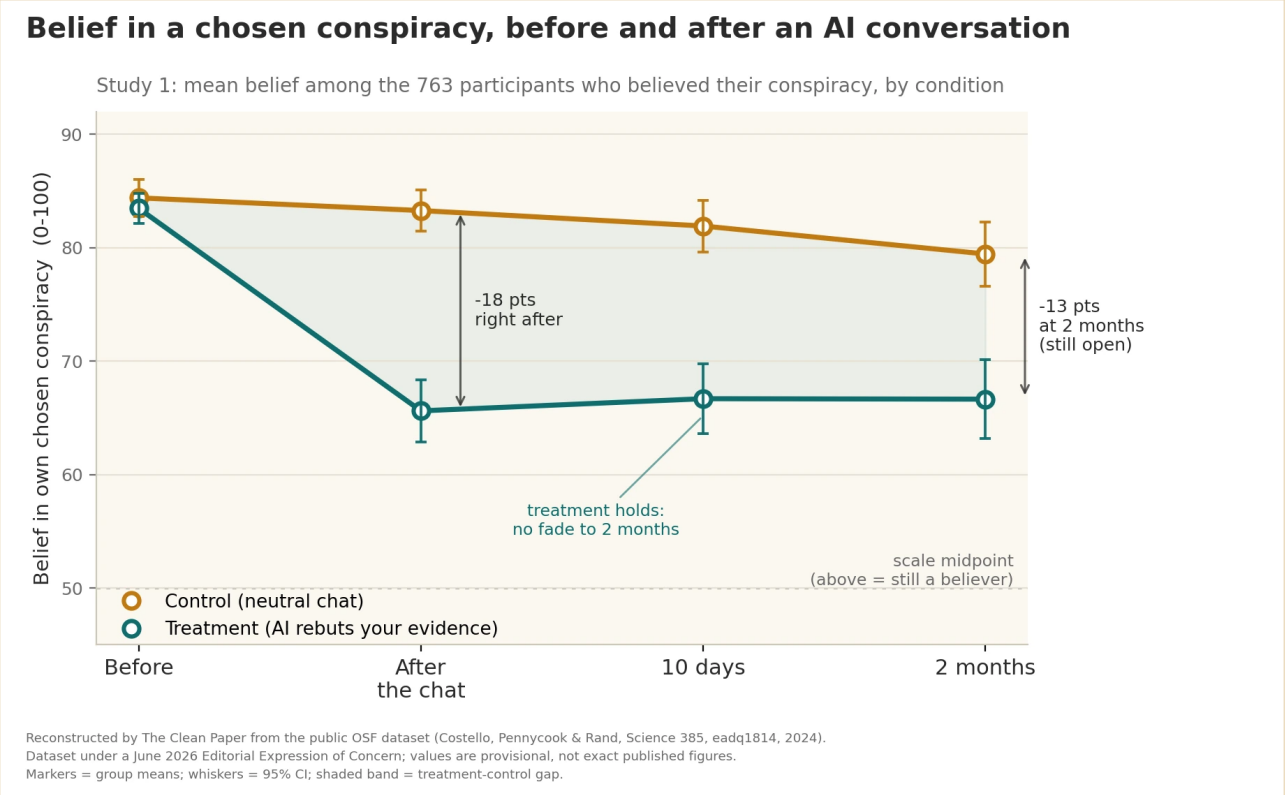
**Hvað Expression of Concern er — og er ekki**

Editorial Expression of Concern er formleg viðvörunarmerking sem tímarit setur við grein til að upplýsa lesendur um að spurningar hafi komið upp meðan þær eru enn í rannsókn. Hún er **ekki** afturköllun (greinin hefur ekki verið dregin til baka) og er ekki sjálfkrafa niðurstaða um misferli. Hún þýðir: lestu greinina með þessum fyrirvara í huga, því vísindaskráin getur breyst. Hér rannsökuðu höfundarnir sjálfir atriðin eftir að þeim var bent á þau, tilkynntu *Science* um þau og lögðu fram leiðrétta greiningu.

Það sem var merkt er afmarkað. Höfundunum var bent á ósamræmi í því hvernig síunarskilyrðum þátttakenda var beitt milli handritsins og birtu greiningarkóðans, og að opinbera gagnasafnið innihélt auka raðir sem höfðu límst inn vegna villu við samruna kóða. Þetta gerði erfitt að endurskapa sumar birtar tölur úr opinberu gögnunum. Höfundarnir sendu *Science* leiðrétta greiningarleið og

uppfærðar niðurstöður sem þeir segja passa við upprunalegu niðurstöðuna **í stefnu, tölfræðilegri marktækni og efnislegri stærð**. *Science* metur þær enn. Þar til því mati lýkur eru nákvæmu tölurnar undir spurningarmerki sem aðeins tímaritið getur fjarlægt.

Þetta eru því tvær sögur í einu: áhugaverð niðurstaða um hvort staðreyndir geti hreyft við rótgróinni trú og lifandi dæmi um hvernig vísindaskrá leiðréttir sjálfa sig opinberlega. Tilgangurinn hér er að halda þessu tvennu aðskildu.




Í rannsókninni var greint frá því að þriggja umferða gervigreindarsamtal sem svaraði eigin rökum þátttakanda fyrir samsæriskenningu drægi trú hans á valda kenningu niður um um 20% að meðaltali. Ólíkt viðmiðunarsamtali um ótengt efni var lækkunin enn mælanleg eftir 10 daga og 2 mánuði. Áhrifin voru varanleg að þessu marki en ekki algjör: flestir trúðu enn kenningunni eftir á — minni trú, ekki „lækning“. Greinin er nú undir Editorial Expression of Concern (*Science*, júní 2026) vegna ósamræmis í skýrslugjöf og villu í opinbera gagnasafninu; höfundarnir segja leiðrétta greiningu varðveita stefnu, marktækni og stærð áhrifanna meðan *Science* metur málið. Þetta línurit er endurgert af The Clean Paper úr sama opinbera gagnasafni og gildin eru því bráðabirgðatölur, ekki endanlegar tölur greinarinnar.

Original figure — The Clean Paper CC BY 4.0

# Hvað höfundarnir gerðu

- Framkvæmdu tvær tilraunir með 2.190 þátttakendum sem hver nefndi, með eigin orðum, samsæriskenningu sem hann trúði og þau gögn sem hann taldi styðja hana.
- Létu hvern þátttakanda eiga þrjár umferðir af samtali við GPT-4 Turbo. Í **meðferðarhópnum** var

gervigreindin beðin að svara sértækum gögnum þátttakandans; í **viðmiðunarhópnum** ræddi hún óskylt efni.
- Mældu trú á valda samsæriskenningu fyrir og strax eftir samtalið og höfðu aftur samband eftir **10 daga og 2 mánuði**.
- Létu faglegan staðreyndaprófara meta nákvæmni allra 128 staðreyndalegra fullyrðinga gervigreindarinnar í úrtaki.
- Prófuðu hvort áhrifin færðust yfir á aðrar, ótengdar samsæriskenningar — og sem sértækniþróf hvort þau lækkuðu einnig trú á samsæri sem eru í raun **sönn**.

# Hvað þeir fundu

- **Um 20% meðalminnkun** á trú á valda samsæriskenningu í meðferðarhópi miðað við viðmið (rannsókn 1: 95% öryggisbil [13,8, 19,7] stig á 100 stiga kvarða, P < 0,001).
- **Áhrifin héldust.** Í meðferðarhópnum hvarf lækkunin ekki fljótt; trú var enn nálægt gildinu strax eftir samtal bæði eftir 10 daga og tvo mánuði, meðan viðmiðunarhópurinn var hærri. Greinin lýsir áhrifunum á valda kenningu sem óskertum í að minnsta kosti tvo mánuði, ekki tímabundinni sveiflu.
- **Áhrifin náðu yfir margs konar kenningar** og komu fram jafnvel hjá þátttakendum sem höfðu mjög sterka trú í upphafi.
- **Fullyrðingar gervigreindarinnar voru nákvæmar** í þessu verkefni: 127 af 128 (99,2%) voru metnar sannar, 1 (0,8%) villandi og engin röng.
- **Áhrifin voru sértæk**, ekki almenn efahyggja: inngripið lækkaði **ekki** trú á sannar samsærisfullyrðingar, sem bendir til að það hafi hreyft við órökstuddri trú fremur en að gera fólk tortryggið gagnvart öllu.
- **Smávægileg yfirfærsluáhrif** komu fram — trú á aðrar, ótengdar samsæriskenningar lækkaði um um 12%, og þátttakendur sögðust líklegri til að mótmæla samsærisfullyrðingum. Megináhrifin héldust í rannsókn 2 og bentu í sömu átt í minna úrtaki án viðmiðunarhóps (veikari gögn en samkvæm).

# Hvað þetta sannar ekki

- Niðurstaðan stendur **ekki óumdeild núna**. Greinin er undir Editorial Expression of Concern vegna gagnameðferðar- og endurtekningarvandamála, og leiðréttu tölurnar eru enn til mats hjá *Science*. Meðhöndlaðu nákvæmu gildin sem bráðabirgða þar til því mati lýkur.
- Hún sýnir **ekki** að gervigreind „lækni“ samsæriskenningatrú. Um 20% meðalminnkun er merkingarbær breyting, ekki eyðing — flestir trúðu kenningunni enn eftir á, bara minna.
- Þetta er **ekki** niðurstaða úr raunverulegri útbreiddri notkun. Þátttakendur skráðu sig í rannsókn og tóku þátt í samtalinu af góðum vilja; í raunheimi er líklegt að harðasti kjarni samsæriskenningafólks sé síst líklegur til að leita til spjallmennis sem muni rökræða við það.
- 99,2% nákvæmnin er **bundin þessu verkefni, þessu líkani og vönduðum fyrirmælum** — ekki almenn trygging um að mállíkön segi sannleikann. Höfundarnir benda sjálfir á að sama sérsniðna

sannfæringargeta gæti styrkt *rangar* hugmyndir ef hún væri notuð í þá átt.
- „Varanlegt“ þýðir hér **tveir mánuðir**, sem er athyglisvert eftir eitt samtal en ekki það sama og varanlegt til æviloka.

# Hversu sterk eru gögnin?

- **Rannsóknarsniðið er raunverulegur styrkleiki.** Meðferðar- og viðmiðunartilraunir, hreinn lyfleysulíkur samanburður, endurtekning í tveimur rannsóknum og óháðu úrtaki, eftirfylgni yfir tíma og sérstakt sannleikapróf á fullyrðingum gervigreindarinnar — þetta er vandaðri hönnun en algengt er í sannfæringarrannsóknum og stefna áhrifanna er samkvæm þar sem hún var prófuð.
- **En Expression of Concern er ekki neðanmálsgrein.** Að geta endurskapað birtar tölur úr opinberum gögnum og kóða er hluti af trausti á niðurstöðu, og það er einmitt þar sem vandinn kom fram — ósamræmi í skimunarskilyrðum og tvíteknar/aukalegar raðir vegna samrunavillu í kóða. Fullyrðing höfundanna um að leiðrétt greining varðveiti niðurstöðuna er hvetjandi og möguleg, en hún er þeirra eigin frásögn, ekki enn óháður úrskurður. Mat *Science* er það sem ræður.
- **Heiðarleg staða er „merkt“, ekki „hrakin“ eða „staðfest“.** Vönduð og sláandi rannsókn þar sem nákvæmu tölurnar eru nú undir formlegri endurskoðun. Það er óþægileg staða fyrir góða sögu en rétta staðan.

# Af hverju þetta skiptir máli

Í mörg ár hefur áhrifamikil skoðun verið sú að samsæriskenningafólk sé knúið áfram af sálfræðilegum þörfum og sjálfsmynd og því sé ekki hægt að rökræða það úr trú með staðreyndum. Varanleg, gagnadrifin lækkun — ef hún stenst endurskoðun — væri mikilvæg mótvægi við þá svartsýni: hún bendir til að vandinn hafi að hluta verið sá að almenn staðreyndaleiðrétting svaraði aldrei *sértæku* „sönnunargögnunum“ sem einstaklingurinn vitnar í, eitthvað sem LLM getur gert í stórum skala.

Rannsóknin skiptir einnig máli sem dæmi um hvernig vísindi eiga að virka. Lærdómur Expression of Concern er ekki að vel þekktar niðurstöður séu einskis virði; hann er að villur komi upp á yfirborðið, gögn séu endurskoðuð og fullyrðingar prófaðar aftur fyrir opnum tjöldum. Að þetta kerfi virki er styrkleiki, ekki hneyksli — og ástæðan fyrir því að rétta afstaðan til niðurstöðunnar er þolinmæði fremur en annaðhvort fagnaðarlæti eða afskrifun.

Og þetta sker í báðar áttir varðandi gervigreind. Sama geta til að veikja ranga trú með sérsniðnum rökum gæti styrkt ranga trú alveg jafn vel. Aðferðin sjálf er hlutlaus; markmiðið er það ekki.

# Í stuttu máli

Rannsókn frá 2024 greindi frá því að þriggja umferða samtal við GPT-4 Turbo hefði minnkað trú fólks á samsæriskenningu sem það sjálft trúði um um 20%, að áhrifin héldust í tvo mánuði og næðu yfir margs konar kenningar, en fullyrðingar gervigreindarinnar voru metnar nær alltaf réttar. Í júní 2026 fékk greinin Editorial Expression of Concern vegna vandamála í gagnameðferð og endurtekningarhæfni; höfundarnir segja leiðrétta greiningu varðveita stefnu, marktækni og stærð niðurstöðunnar og *Science* er enn að meta málið. Lestu þetta sem raunverulega áhugaverða og vel hannaða niðurstöðu sem er nú undir endurskoðun — ekki útkljáða og ekki hrakta. Heiðarlegasta setningin er sú sem vísindaferlið sjálft er að skrifa: athugaðu hana aftur.

## Heimildir

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