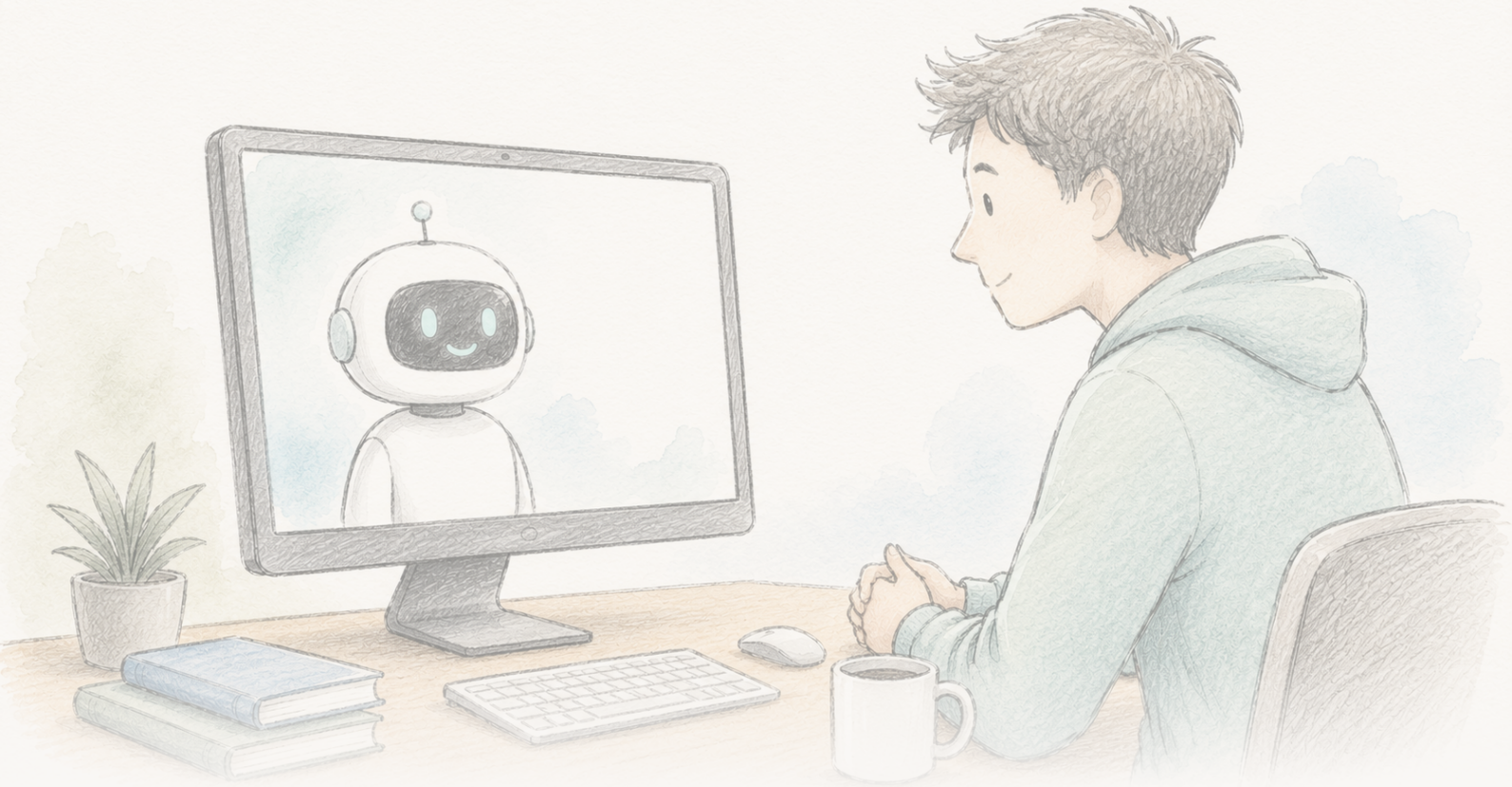


The Clean Paper

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

Keskustelu chatbotin kanssa näytti vähentävän salaliittouskomuksia kuukausiksi

Vuoden 2024 tutkimuksessa kolmen keskustelukierroksen keskustelu GPT-4 Turbon kanssa vähensi ihmisten uskoa heidän omaan salaliittoteoriaansa noin 20 %, vaikutus säilyi kaksi kuukautta, näkyi eri salaliittotyypeissä ja perustui tekoälyn väitteisiin, jotka arvioitiin lähes kauttaaltaan paikkansapitäviksi. Kesäkuussa 2026 Science liitti artikkeliin Editorial Expression of Concern -huomautuksen datankäsittelyyn ja toistettavuuteen liittyvien ongelmien vuoksi. Kirjoittajien mukaan korjattu analyysi säilyttää tuloksen suunnan, tilastollisen merkitsevyyden ja suuruuden, ja Science arvioi asiaa edelleen. Aidosti kiinnostava ja huolellisesti rakennettu tulos — mutta parhaillaan tarkastelussa, ei vahvistettu eikä kumottu.

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 on liittänyt artikkeliin Editorial Expression of Concern -huomautuksen arvioidessaan datankäsittelyyn ja toistettavuuteen liittyviä kysymyksiä. Kyse ei ole artikkelin takaisinvedosta eikä todetusta väärinkäytöksestä; tulos on luettava alustavana, kunnes arviointi ratkeaa.

Editorial Expression of Concern →

Faktoilla sittenkin — ja artikkelissa varoituslippu

Vuonna 2024 julkaistu tutkimus raportoi tuloksen, joka sotii yhtä mukavaa perusolettamusta vastaan. Anna ihmiselle lyhyt, kolmen kierroksen keskustelu tekoälyn kanssa — GPT-4 Turbo, ohjeistettuna vastaamaan juuri niihin todisteisiin, joita henkilö itse esittää uskomansa salaliittoteorian puolesta — ja keskimäärin hänen uskonsa teoriaan vähenee noin 20 %. Lasku näkyi yhä kaksi kuukautta myöhemmin. Vaikutus löytyi sekä klassisista salaliitoista (John F. Kennedyn murha, avaruusolennot, Illuminati) että ajankohtaisista (COVID-19, Yhdysvaltain vuoden 2020 vaalit), myös ihmisillä, joiden usko oli vahvaa ja kietoutunut identiteettiin. Kun ammattimainen faktantarkistaja arvioi 128 tekoälyn esittämää väitettä, 127 oli totta, yksi harhaanjohtava eikä yksikään väärä.

Laajalle levinnyt otsikko oli, että ihmiset *voi* puhua ulos kaninkolosta — salaliittoihin uskovat eivät ole todisteiden tavoittamattomissa, vaan tarvitsevat oikeat todisteet riittävän täsmällisesti kohdennettuina. Se on aidosti kiinnostava väite, ja tutkimus oli rakennettu huolellisemmin kuin suuri osa suostuttelua koskevasta tutkimuksesta.

Sitten kesäkuussa 2026 *Science* lisäsi artikkeliin **Editorial Expression of Concern** -huomautuksen. Tämä on kohta, jonka monet uudelleenkertomukset jättävät pois, vaikka juuri se ratkaisee, kuinka paljon painoa tulos tällä hetkellä kestää.

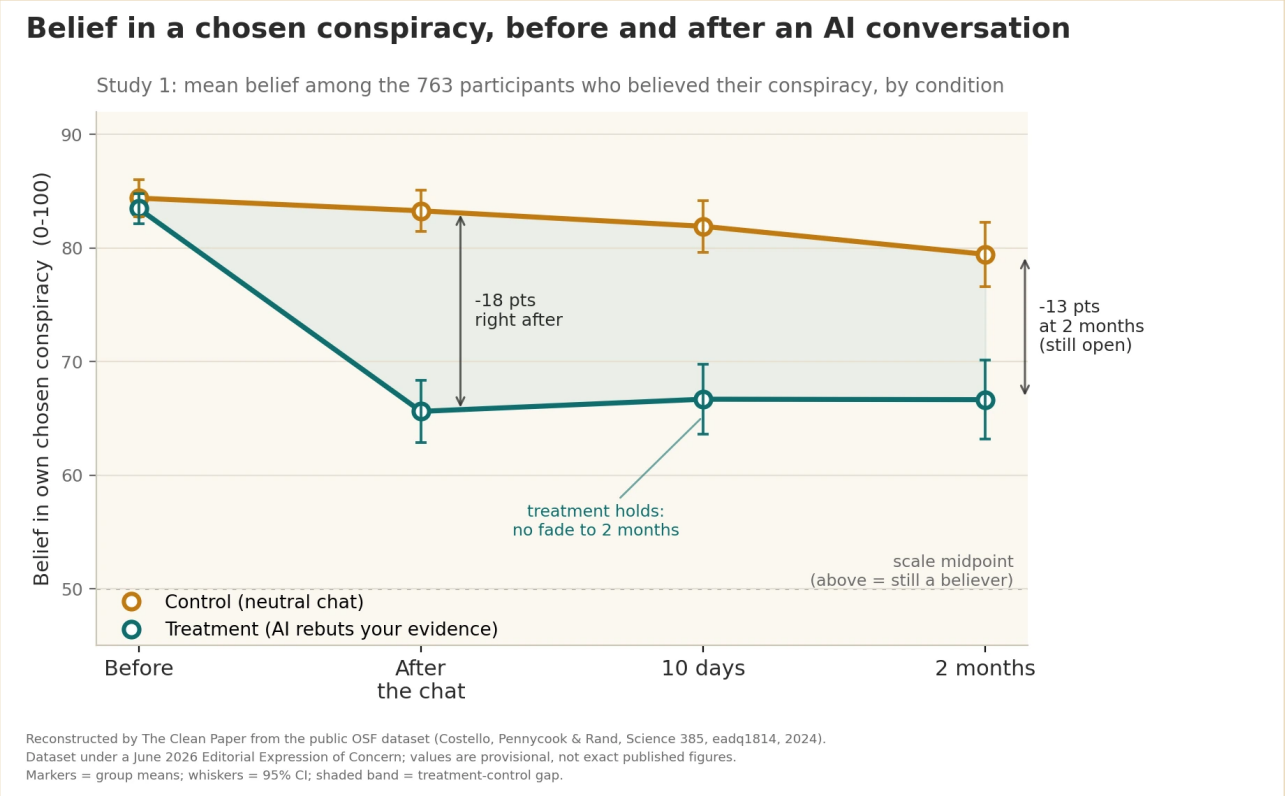
Mitä Expression of Concern tarkoittaa — ja mitä ei

Editorial Expression of Concern on lehden artikkeliin liittämä virallinen varoitus, jolla lukijoille kerrotaan esiin nousseista kysymyksistä niiden selvittämisen vielä jatkuessa. Se **ei** ole takaisinvetäminen (artikkelia ei ole poistettu), eikä se itsessään tarkoita todettua väärinkäytöstä. Viesti on: lue tulos tämä varaus mielessä, koska tieteellinen rekisteri voi vielä muuttua. Tässä kirjoittajat tutkivat esiin tulleet ongelmat, raportoivat yksityiskohdat Sciencelle ja toimittivat korjatun analyysin.

Huomautuksen syyt ovat täsmällisiä. Kirjoittajille ilmoitettiin ristiriidoista siinä, miten osallistujien seulontakriteerejä sovellettiin käsikirjoituksen ja julkaistun analyysikoodin välillä, sekä siitä, että julkisessa aineistossa oli ylimääräisiä rivejä, jotka olivat päätyneet mukaan koodin yhdistämisvirheen vuoksi. Tämä teki osasta raportoituja lukuja vaikeita toistaa julkaistusta materiaalista. Kirjoittajat toimittivat *Sciencelle* korjatun analyysiputken ja päivitetyt tulokset, joiden he sanovat vastaavan alkuperäisiä **suunnan, tilastollisen merkitsevyyden ja olennaisen vaikutuskoon osalta**. *Science* arvioi niitä. Kunnes arviointi valmistuu, tarkat luvut ovat kysymysmerkin alla, jonka vain

lehti voi poistaa.

Kyse on siis kahdesta tarinasta yhtä aikaa: kiinnostavasta tuloksesta siitä, voivatko faktat liikuttaa juurtuneita uskomuksia, ja elävästä esimerkistä siitä, miten tieteellinen kirjallisuus korjaa itseään avoimesti. Tämän tekstin tehtävä on pitää tarinat erillään.



Tutkimuksessa kolmen kierroksen tekoälykeskustelun, joka kumosi osallistujan omia salaliittouskomuksensa tueksi esittämiä todisteita, raportoitiin vähentävän valittuun salaliittoon uskomista keskimäärin noin 20 %. Aiheeseen liittymättömästä asiasta keskustelleeseen kontrolliryhmään verrattuna lasku oli yhä mitattavissa 10 päivän ja 2 kuukauden kohdalla. Vaikutus oli pitkäkestoinen mutta osittainen: useimmat osallistujat uskoivat salaliittoon edelleen — vähemmän, eivät nolla. Artikkeli on parhaillaan Editorial Expression of Concern -huomautuksen alainen (*Science*, kesäkuu 2026) raportoinnin ristiriitojen ja julkisen aineiston virheen vuoksi; kirjoittajien mukaan korjattu analyysi säilyttää vaikutuksen suunnan, merkitsevyyden ja koon, mutta *Science* arvioi sitä yhä. The Clean Paper rekonstruoi tämän kaavion julkisesta aineistosta, joten kuvan arvot ovat alustavia, eivät artikkelin lopulliset luvut.

Original figure — The Clean Paper CC BY 4.0

Mitä tutkijat tekivät

- Toteuttivat kaksi koetta yhteensä 2190 osallistujalla. Jokainen nimesi omin sanoin salaliittoteorian, johon uskoi, sekä todisteet, joiden hän katsoi tukevan sitä.
- Jokainen osallistuja kävi kolmen kierroksen keskustelun GPT-4 Turbon kanssa. **Hoitoryhmässä** tekoälyä ohjeistettiin kumoamaan juuri osallistujan esittämät todisteet; **kontrolliryhmässä**

keskusteltiin aiheeseen liittymättömästä asiasta.
- Mittasivat uskoa valittuun salaliittoon ennen keskustelua ja heti sen jälkeen sekä tavoittivat osallistujat uudelleen **10 päivän ja 2 kuukauden** kohdalla.
- Ammattimainen faktantarkistaja arvioi otoksessa kaikki tekoälyn esittämät 128 faktaväitettä.
- Tutkivat, siirtyikö vaikutus muihin, asiaan liittymättömiin salaliittoihin — ja spesifisyyden testinä myös sitä, vähensikö interventio uskoa salaliittoihin, jotka sattuvat olemaan **tosia**.

Mitä he löysivät

- **Noin 20 prosentin keskimääräinen lasku** valittuun salaliittoon uskomisessa hoitoryhmässä verrattuna kontrolliin (tutkimus 1: 95 %:n luottamusväli [13.8, 19.7] pistettä 100 pisteen asteikolla, $P < 0.001$).
- **Vaikutus säilyi.** Hoitoryhmässä lasku ei hävinnyt: uskomus pysyi 10 päivän ja kahden kuukauden kohdalla lähellä heti keskustelun jälkeistä tasoa eikä noussut takaisin, samalla kun kontrolliryhmä pysyi korkeammalla. Artikkeli kuvaa vaikutusta valittuun salaliittoon vähintään kaksi kuukautta heikkenemättä säilyneeksi, ei hetkelliseksi heilahdukseksi.
- **Vaikutus yleistyi** monenlaisiin osallistujien nimeämiin salaliittoihin ja näkyi myös niillä, joiden alkuperäinen usko oli vahva.
- **Tekoälyn väitteet olivat tässä asetelmassa paikkansapitäviä:** 127/128 (99.2 %) arvioitiin tosiksi, 1 (0.8 %) harhaanjohtavaksi, eikä yhtäkään vääräksi.
- **Vaikutus oli spesifinen**, ei yleistä epäluuloa lisäävä: interventio **ei** vähentänyt uskoa tosiksi osoittautuneisiin salaliittoihin, mikä viittaa siihen, että se liikutti perusteettomia uskomuksia eikä tehnyt ihmisistä kyynisiä kaikkea kohtaan.
- **Vaikutus levisi maltillisesti** myös muihin salaliittoihin — usko asiaan liittymättömiin salaliittoihin laski noin 12 %, ja osallistujat ilmoittivat suuremmasta aikomuksesta vastustaa salaliittoväitteitä. Päätulos toistui tutkimuksessa 2 ja osoitti samaan suuntaan pienemmässä aineistossa, jossa ei ollut kontrolliryhmää (heikompi näyttö, mutta linjassa).

Mitä tämä ei todista

- Tulos **ei** tällä hetkellä seiso ilman kysymysmerkkiä. Artikkelissa on Editorial Expression of Concern datankäsittely- ja toistettavuusongelmien vuoksi, ja *Science* arvioi yhä korjattuja lukuja. Täsmällisiä arvoja kannattaa pitää alustavina, kunnes arviointi valmistuu.
- Se **ei** osoita tekoälyn ”parantavan” salaliittouskomuksia. Noin 20 prosentin keskimääräinen lasku on merkittävä siirtymä, ei uskomuksen poistaminen — useimmat uskoivat teoriaan keskustelun jälkeenkin, vain vähemmän.
- Kyse **ei** ole tosielämän käyttöönotosta. Osallistujat ilmoittautuivat tutkimukseen ja keskustelivat tekoälyn kanssa hyvässä uskossa; todellisessa maailmassa vahvimmin salaliittoon sitoutuneet ihmiset ovat juuri niitä, jotka epätodennäköisimmin hakeutuvat chatbotin luo väittelemään siitä.
- 99.2 prosentin tarkkuus on **tämän tehtävän, mallin ja huolellisen promptauksen tulos** — ei yleinen takuu siitä, että kielimallit puhuvat totta. Kirjoittajat korostavat, että sama yksilöity

suostutteluvoima voisi vahvistaa *vääriä* uskomuksia, jos sitä käytettäisiin siihen.
- ”Pitkäkestoinen” tarkoittaa tässä **kahta kuukautta**, mikä on yhdelle keskustelulle huomionarvoista mutta ei sama asia kuin pysyvä.

Kuinka vahvaa näyttö on

- **Tutkimusasetelma on todellinen vahvuus.** Kontrolloidut hoito-vastaan-kontrolli-kokeet, puhdas plasebon kaltainen kontrolli, toisto kahdessa tutkimuksessa ja erillisessä aineistossa, pitkäkestoisuuden seuranta sekä tekoälyn omien väitteiden eksplisiittinen faktantarkistus tekevät tästä huolellisempaa kuin suuren osan suostuttelututkimuksesta. Vaikutuksen suunta on ollut johdonmukainen kaikkialla, missä sitä testattiin.
- **Mutta Expression of Concern ei ole alaviite.** Luotettavuuteen kuuluu, että julkaistusta datasta ja koodista pystytään tuottamaan raportoidut luvut uudelleen, ja juuri tässä tapahtui katkos — seulontakriteerien ristiriidat sekä koodin yhdistämisvirheestä syntyneet ylimääräiset rivit. Kirjoittajien väite siitä, että korjattu putki säilyttää tuloksen, on rohkaiseva ja uskottava, mutta se on heidän oma raporttinsa, ei vielä riippumaton ratkaisu. *Sciencen* arviointi ratkaisee asian.
- **Rehellinen tila on ”merkitty tarkastettavaksi”, ei ”kumottu” tai ”vahvistettu”.** Hyvin rakennettu ja vaikuttava tutkimus, jonka tarkat luvut ovat virallisessa arvioinnissa. Hyvän tarinan jättäminen tähän epämukavaan välitilaan on tarkin tapa kertoa siitä.

Miksi tällä on merkitystä

Vuosien ajan vaikutusvaltainen näkemys on ollut, että salaliittouskomuksia ajavat psykologiset tarpeet ja identiteetti, joten ihmisiä ei voi puhua niistä ulos faktoilla. Jos pitkäkestoinen, näyttöön perustuva lasku kestää tarkastelun, se on merkittävä vastapaino tälle pessimismille: ehkä ongelma oli osittain siinä, ettei yleinen faktantarkistus koskaan tarttunut niihin *täsmällisiin* todisteisiin, joihin uskova itse vetoaa — asiaan, jonka LLM pystyy tekemään suuressa mittakaavassa.

Tulos on tärkeä myös esimerkkinä siitä, miten tieteen pitäisi toimia. Expression of Concern ei opeta, että huomioidut tulokset ovat arvottomia; se näyttää, että virheitä tuodaan esiin, data tutkitaan uudelleen ja väitteet tarkistetaan julkisesti. Tämän koneiston toiminta on ominaisuus, ei skandaali — ja siksi oikea suhtautuminen tulokseen on kärsivällisyys, ei aplodit eikä hylkäys.

Ja tekoälyn kannalta vaikutus toimii molempiin suuntiin. Sama kyky heikentää väärää uskomusta räätälöidyllä argumentilla voisi yhtä hyvin vahvistaa väärää uskomusta. Tekniikka on neutraali; tavoite ei ole.

Tiivistelmä

Vuoden 2024 tutkimuksessa kolmen kierroksen keskustelu GPT-4 Turbon kanssa vähensi ihmisten uskoa heidän omaan salaliittoteoriaansa noin 20 %, vaikutus säilyi kaksi kuukautta ja näkyi eri salaliittotyypeissä, ja tekoälyn faktaväitteet arvioitiin lähes täysin oikeiksi. Kesäkuussa 2026 artikkeli sai Editorial Expression of Concern -huomautuksen datankäsittelyyn ja toistettavuuteen liittyvien ongelmien vuoksi. Kirjoittajien mukaan korjattu analyysi säilyttää tuloksen suunnan, merkitsevyyden ja suuruuden, ja *Science* arvioi asiaa edelleen. Se kannattaa lukea aidosti kiinnostavana ja huolellisesti rakennettuna tuloksena, joka on juuri nyt tarkastelussa — ei ratkaistuna eikä kumottuna. Rehellisin lause siitä on sama, jonka itse prosessi kirjoittaa: tarkista uudelleen.

Lähteet

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