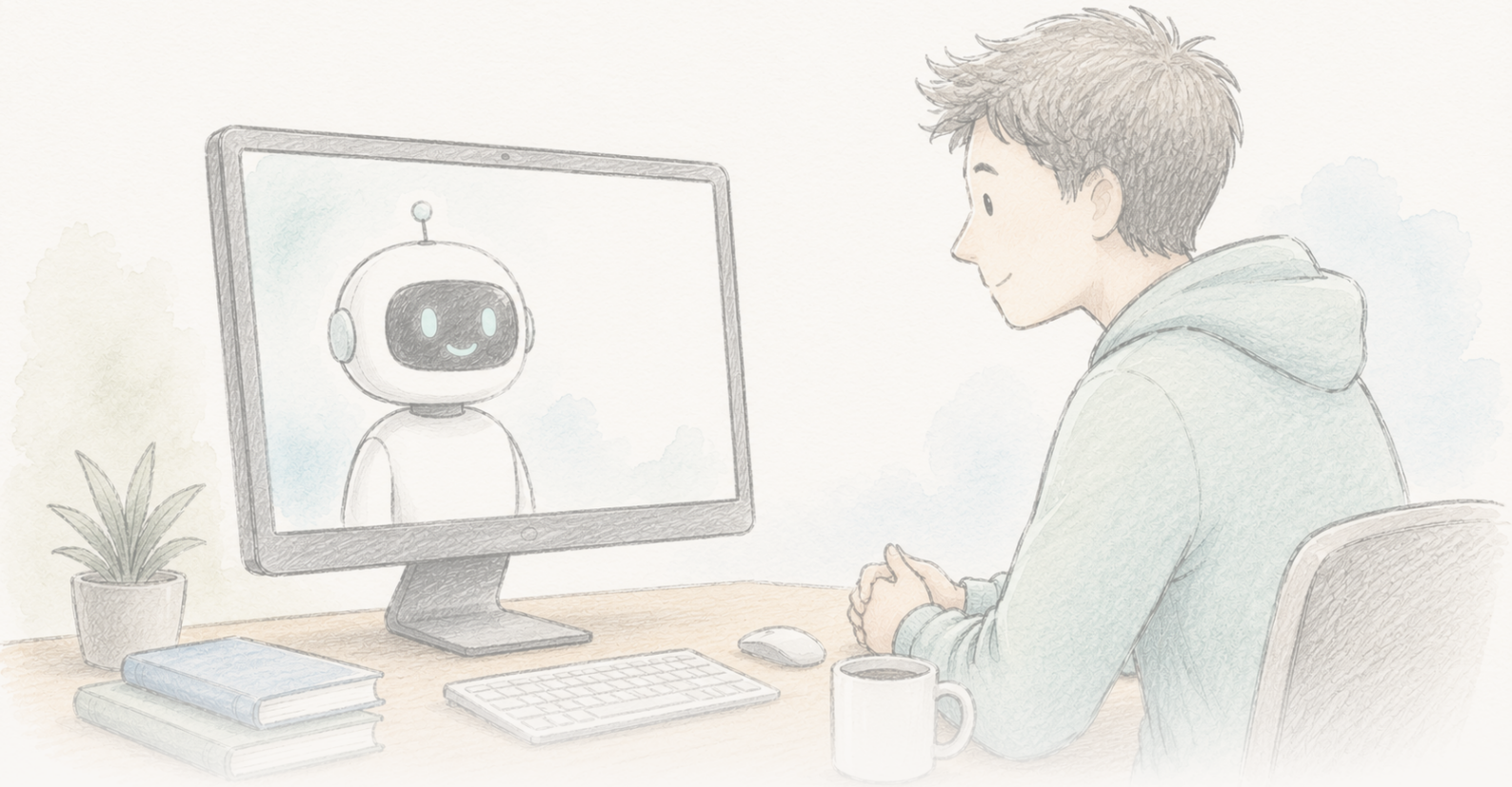


The Clean Paper

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

Sebuah chatbot kelihatan mengurangkan kepercayaan konspirasi selama berbulan-bulan

Satu kajian 2024 mendapati bahawa perbualan tiga pusingan dengan GPT-4 Turbo mengurangkan kepercayaan seseorang terhadap teori konspirasi yang mereka pegang kira-kira 20%, kesan yang bertahan dua bulan, berlaku merentas jenis konspirasi, dan bergantung pada dakwaan AI yang dinilai hampir semuanya tepat. Pada Jun 2026, kertas itu diberi Editorial Expression of Concern kerana masalah pengendalian data dan kebolehulangan; para penulis mengatakan analisis yang dibetulkan mengekalkan arah, keertian statistik dan saiz penemuan, sementara Science masih menilai. Satu hasil yang benar-benar menarik dan dibina dengan teliti — kini masih dalam semakan, belum disahkan sepenuhnya dan belum disangkal.

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 telah melampirkan Editorial Expression of Concern pada kertas ini sementara ia menilai persoalan tentang pengendalian data dan kebolehulangan. Ini bukan penarikan balik dan bukan dapatan salah laku; ia bermaksud hasil yang dilaporkan perlu dibaca sebagai sementara sehingga semakan itu selesai.

Editorial Expression of Concern →

Fakta, rupanya — dan satu tanda amaran pada kertas itu

Pada 2024, satu kajian melaporkan sesuatu yang bertentangan dengan sebahagian kebijaksanaan umum yang selesa. Berikan seseorang perbualan singkat tiga pusingan dengan AI — GPT-4 Turbo, diarahkan untuk menjawab bukti khusus yang mereka kemukakan bagi teori konspirasi yang mereka sendiri percayai — dan secara purata kepercayaan mereka terhadap teori itu turun kira-kira 20%. Penurunan itu masih ada dua bulan kemudian. Ia berlaku pada konspirasi klasik (pembunuhan John F. Kennedy, makhluk asing, Illuminati) dan isu semasa (COVID-19, pilihan raya AS 2020), malah pada orang yang kepercayaannya kuat dan terikat pada identiti mereka. Apabila pemeriksa fakta profesional menilai 128 dakwaan yang dibuat AI, 127 adalah benar, satu mengelirukan, dan tiada yang palsu.

Tajuk utama yang tersebar ialah bahawa anda *boleh* bercakap dengan seseorang sehingga mereka keluar daripada lubang arnab konspirasi — bahawa penganut konspirasi bukan di luar jangkauan bukti, mereka cuma memerlukan bukti yang tepat, disampaikan dengan cukup khusus. Itu dakwaan yang benar-benar menarik, dan kajian ini dibina dengan lebih teliti daripada kebanyakan penyelidikan tentang pemujukan.

Kemudian, pada Jun 2026, *Science* meletakkan **Editorial Expression of Concern** pada kertas itu. Inilah bahagian yang akan ditinggalkan oleh banyak penceritaan semula, dan tepat bahagian inilah yang menentukan berapa banyak berat yang boleh dipikul oleh hasil tersebut buat masa ini.

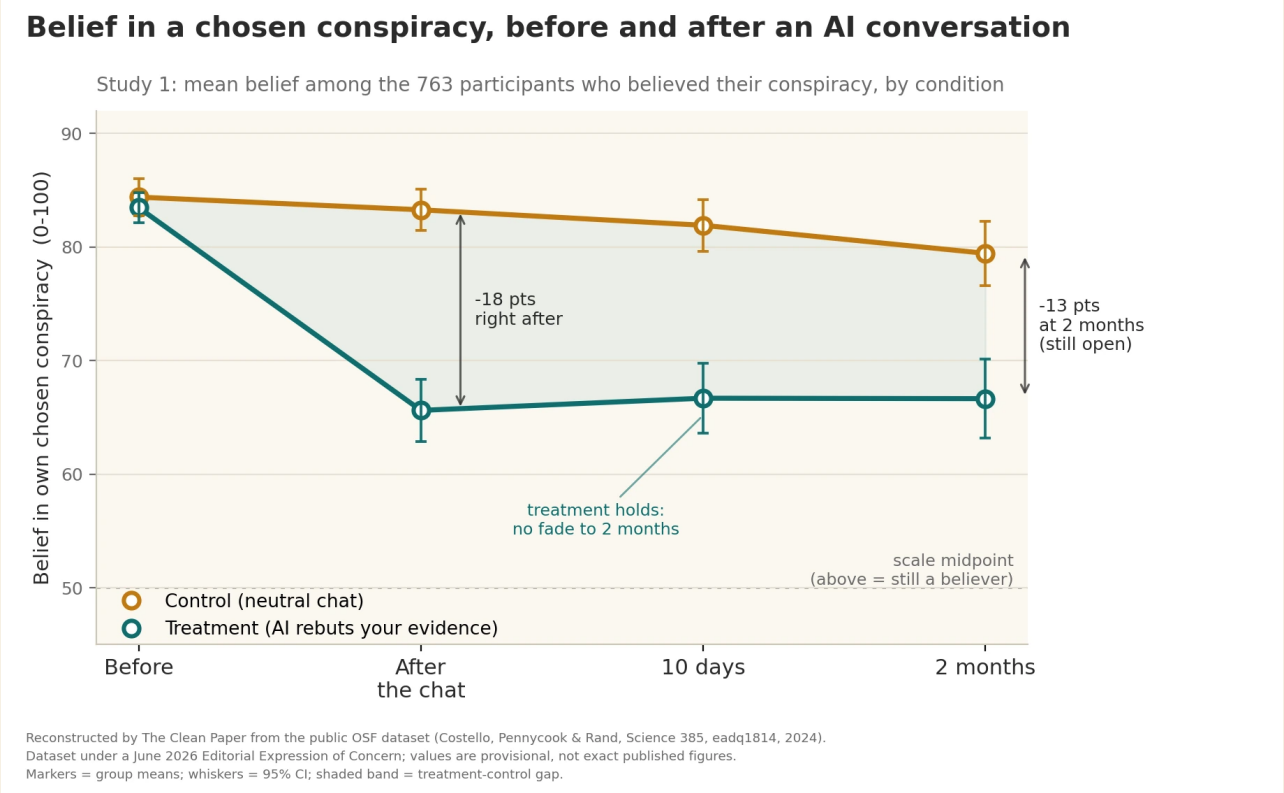
Apakah Expression of Concern — dan apakah yang bukan

Editorial Expression of Concern ialah tanda rasmi yang dilampirkan jurnal pada sesuatu kertas untuk memaklumkan pembaca bahawa persoalan telah dibangkitkan, sementara persoalan itu masih sedang diselesaikan. Ia **bukan** penarikan balik (kertas itu tidak ditarik), dan dengan sendirinya bukan dapatan salah laku. Maksudnya: baca dengan syarat ini dalam fikiran, kerana rekod mungkin berubah. Di sini, selepas isu tersebut dimaklumkan kepada mereka, para penulis menyiasat, melaporkan butirannya kepada Science, dan menyerahkan analisis yang dibetulkan.

Perkara yang ditandai itu khusus. Para penulis dimaklumkan tentang ketidakselarasan dalam cara kriteria saringan peserta digunakan antara manuskrip dan kod analisis yang diterbitkan, serta bahawa set data awam mengandungi baris tambahan yang tidak sepatutnya kerana ralat penggabungan

kod — masalah yang menjadikan sebahagian nombor yang dilaporkan sukar dihasilkan semula daripada bahan yang dikeluarkan. Para penulis memberikan *Science* saluran analisis yang dibetulkan dan satu set hasil dikemas kini, yang menurut mereka sepadan dengan hasil asal **dari segi arah, keertian statistik dan saiz substantif**. *Science* sedang menilainya. Sehingga penilaian itu selesai, angka tepat berada di bawah tanda soal yang hanya jurnal itu boleh angkat.

Jadi ini sebenarnya dua cerita serentak: satu hasil menarik tentang sama ada fakta boleh menggerakkan kepercayaan yang telah berakar, dan satu contoh langsung bagaimana rekod saintifik membetulkan dirinya secara terbuka. Tujuan tulisan ini ialah memastikan kedua-duanya tidak bercampur.



Dalam kajian itu, perbualan AI tiga pusingan yang menyangkal bukti yang dinyatakan sendiri oleh peserta dilaporkan mengurangkan kepercayaan terhadap konspirasi pilihan mereka kira-kira 20% secara purata; berbeza daripada sembang kawalan tentang topik tidak berkaitan, penurunan itu masih boleh diukur selepas 10 hari dan 2 bulan. Kesan yang dilaporkan tahan lama tetapi hanya separa: kebanyakan peserta masih mempercayai konspirasi itu selepasnya — satu pengurangan, bukan penawar. Kertas ini kini berada di bawah Editorial Expression of Concern (*Science*, Jun 2026) kerana ketidakselarasan pelaporan dan satu ralat dalam set data awam; para penulis mengatakan analisis yang dibetulkan mengekalkan arah, keertian dan saiz kesan, sementara *Science* terus menilai. Carta ini dibina semula oleh The Clean Paper daripada set data awam tersebut, jadi nilai yang ditunjukkan adalah sementara, bukan angka akhir kertas itu.

Apa yang dilakukan oleh para penulis

- Menjalankan dua eksperimen dengan 2190 peserta yang masing-masing menamakan — dengan kata-kata mereka sendiri — satu teori konspirasi yang mereka percayai dan bukti yang mereka fikir menyokongnya.
- Meminta setiap peserta menjalankan perbualan tiga pusingan dengan GPT-4 Turbo. Dalam keadaan **rawatan**, AI diarahkan untuk menyangkal bukti khusus peserta; dalam keadaan **kawalan**, ia membincangkan topik yang tidak berkaitan.
- Mengukur kepercayaan terhadap konspirasi pilihan sebelum dan sejurus selepas perbualan, kemudian menghubungi peserta semula selepas **10 hari dan 2 bulan**.
- Meminta pemeriksa fakta profesional menilai ketepatan semua 128 dakwaan fakta yang dibuat AI dalam satu sampel.
- Memeriksa sama ada kesan itu merebak kepada konspirasi lain yang tidak berkaitan — dan, sebagai ujian kekhususan, sama ada ia turut menurunkan kepercayaan terhadap konspirasi yang kebetulan **benar**.

Apa yang mereka temui

- **Pengurangan purata kira-kira 20%** dalam kepercayaan terhadap konspirasi pilihan dalam kumpulan rawatan berbanding kawalan (kajian 1: selang keyakinan 95% [13.8, 19.7] mata pada skala 100 mata, $P < 0.001$).
- **Ia bertahan.** Dalam kumpulan rawatan, penurunan itu tidak pudar: kepercayaan kekal hampir pada tahap selepas sembang pada 10 hari dan dua bulan, bukannya naik semula, manakala kumpulan kawalan kekal lebih tinggi sepanjang masa — kertas itu menggambarkan kesan terhadap konspirasi pilihan sebagai bertahan tanpa berkurang sekurang-kurangnya dua bulan, bukan goyangan sesaat.
- **Ia berlaku secara meluas** merentas pelbagai konspirasi yang dinamakan orang, dan kekal walaupun bagi peserta yang pada awalnya sangat yakin.
- **Dakwaan AI tepat** dalam keadaan ini: 127 daripada 128 (99.2%) dinilai benar, 1 (0.8%) mengelirukan, tiada yang palsu.
- **Ia khusus**, bukan keraguan menyeluruh: intervensi **tidak** mengurangkan kepercayaan terhadap konspirasi yang benar, menunjukkan ia menggerakkan kepercayaan yang tidak disokong dan bukannya menjadikan orang sinis terhadap segala-galanya.
- **Terdapat sedikit kesan limpahan** — kepercayaan terhadap konspirasi lain yang tidak berkaitan turun sekitar 12%, dan peserta melaporkan niat lebih besar untuk mencabar dakwaan konspirasi. Kesan utama kekal dalam kajian 2, dan menunjuk ke arah sama dalam sampel lebih kecil tanpa kumpulan kawalan (bukti lebih lemah, tetapi konsisten).

Apa yang tidak dibuktikan oleh kajian ini

- Ia **tidak** boleh dianggap tanpa persoalan buat masa ini. Kertas tersebut berada di bawah Editorial Expression of Concern kerana masalah pengendalian data dan kebolehulangan, dan nombor yang dibetulkan masih dinilai oleh *Science*. Anggap nilai khusus sebagai sementara sehingga semakan itu selesai.
- Ia **tidak** menunjukkan AI "menyembuhkan" kepercayaan konspirasi. Pengurangan purata ~20% ialah perubahan bermakna, bukan penghapusan — kebanyakan peserta masih mempercayai teori itu selepasnya, cuma kurang kuat.
- Ia **bukan** hasil penggunaan dunia sebenar. Peserta memilih untuk masuk kajian dan berinteraksi dengan AI dengan niat baik; di dunia sebenar, orang yang paling kuat terikat kepada sesuatu konspirasi mungkin paling tidak cenderung mencari chatbot yang akan berhujah dengan mereka.
- Ketepatan 99.2% itu **khusus kepada tugasan, model dan arahan berhati-hati ini** — bukan jaminan umum bahawa model bahasa menyatakan perkara yang benar. Para penulis secara jelas mengatakan kuasa pemujukan diperibadikan yang sama juga boleh mendorong kepercayaan *palsu* jika diarahkan ke situ.
- "Tahan lama" di sini bermaksud **dua bulan**, sesuatu yang mengagumkan untuk satu perbualan, tetapi tidak sama dengan kekal selama-lamanya.

Sejauh mana kukuh buktinya

- **Reka bentuk ialah kekuatan sebenar.** Eksperimen rawatan-lawan-kawalan yang terkawal, kawalan gaya plasebo yang bersih, replikasi merentas dua kajian dan satu sampel bebas, susulan ketahanan kesan, serta pemeriksaan ketepatan eksplisit terhadap dakwaan AI sendiri — ini lebih teliti daripada kebanyakan penyelidikan pemujukan, dan arah kesan konsisten di mana sahaja ia diuji.
- **Tetapi Expression of Concern bukan nota kaki.** Keupayaan menghasilkan semula nombor yang dilaporkan daripada data dan kod yang dikeluarkan ialah sebahagian daripada apa yang menjadikan sesuatu hasil boleh dipercayai, dan tepat bahagian itulah yang rosak — ketidakselarasan kriteria saringan dan baris pendua akibat ralat penggabungan kod. Kenyataan para penulis bahawa saluran analisis dibetulkan mengekalkan hasil adalah menggalakkan dan munasabah, tetapi itu masih laporan mereka sendiri, belum keputusan bebas. Penilaian *Science* ialah perkara yang perlu ditunggu.
- **Status jujurnya ialah "ditandai", bukan "disangkal" atau "disahkan".** Satu kajian yang dibina dengan baik dan menarik, tetapi nombor tepatnya sedang berada di bawah semakan rasmi. Itu tempat yang tidak selesa untuk meninggalkan cerita yang bagus, dan itulah tempat yang tepat.

Mengapa ia penting

Selama bertahun-tahun, satu pandangan berpengaruh berpendapat bahawa penganut konspirasi didorong oleh keperluan psikologi dan identiti, maka mereka tidak boleh dihujahkan keluar daripada kepercayaan dengan fakta. Pengurangan yang tahan lama dan berasaskan bukti — jika ia bertahan selepas semakan — menjadi imbangan penting kepada pesimisme itu: ia mencadangkan masalahnya sebahagiannya ialah penyangkalan umum tidak pernah berinteraksi dengan *bukti khusus* yang benar-benar dikemukakan oleh seseorang penganut, sesuatu yang boleh dilakukan LLM pada skala besar.

Ia juga penting sebagai kajian kes tentang bagaimana sains sepatutnya berfungsi. Pelajaran daripada Expression of Concern bukan bahawa hasil terkenal tidak bernilai; pelajarannya ialah ralat dibawa ke permukaan, data diperiksa semula, dan dakwaan diuji semula secara terbuka. Mesin pembetulan itu sedang berfungsi ialah satu ciri, bukan skandal — dan itulah sebab sikap yang betul terhadap hasil ini ialah sabar, bukan tepukan atau penolakan segera.

Dan perkara ini memotong dua arah bagi AI. Keupayaan yang sama untuk melemahkan kepercayaan palsu dengan hujah yang disesuaikan juga boleh menguatkan kepercayaan palsu dengan sama berkesan. Teknik itu neutral; tujuannya tidak.

Ringkasan utama

Satu kajian 2024 mendapati bahawa perbualan tiga pusingan dengan GPT-4 Turbo mengurangkan kepercayaan seseorang terhadap teori konspirasi yang mereka pegang kira-kira 20%, kesan yang bertahan dua bulan dan berlaku merentas jenis konspirasi, dengan dakwaan fakta AI dinilai hampir semuanya tepat. Pada Jun 2026 kertas itu diberi Editorial Expression of Concern kerana masalah pengendalian data dan kebolehulangan; para penulis melaporkan bahawa analisis yang dibetulkan mengekalkan arah, keertian dan saiz penemuan, dan *Science* masih menilai. Bacalah sebagai satu hasil yang benar-benar menarik dan dibina dengan teliti tetapi kini sedang disemak — belum diputuskan, belum disangkal. Ayat paling jujur tentangnya ialah ayat yang sedang ditulis oleh proses itu sendiri: periksa sekali lagi.

Sumber

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`