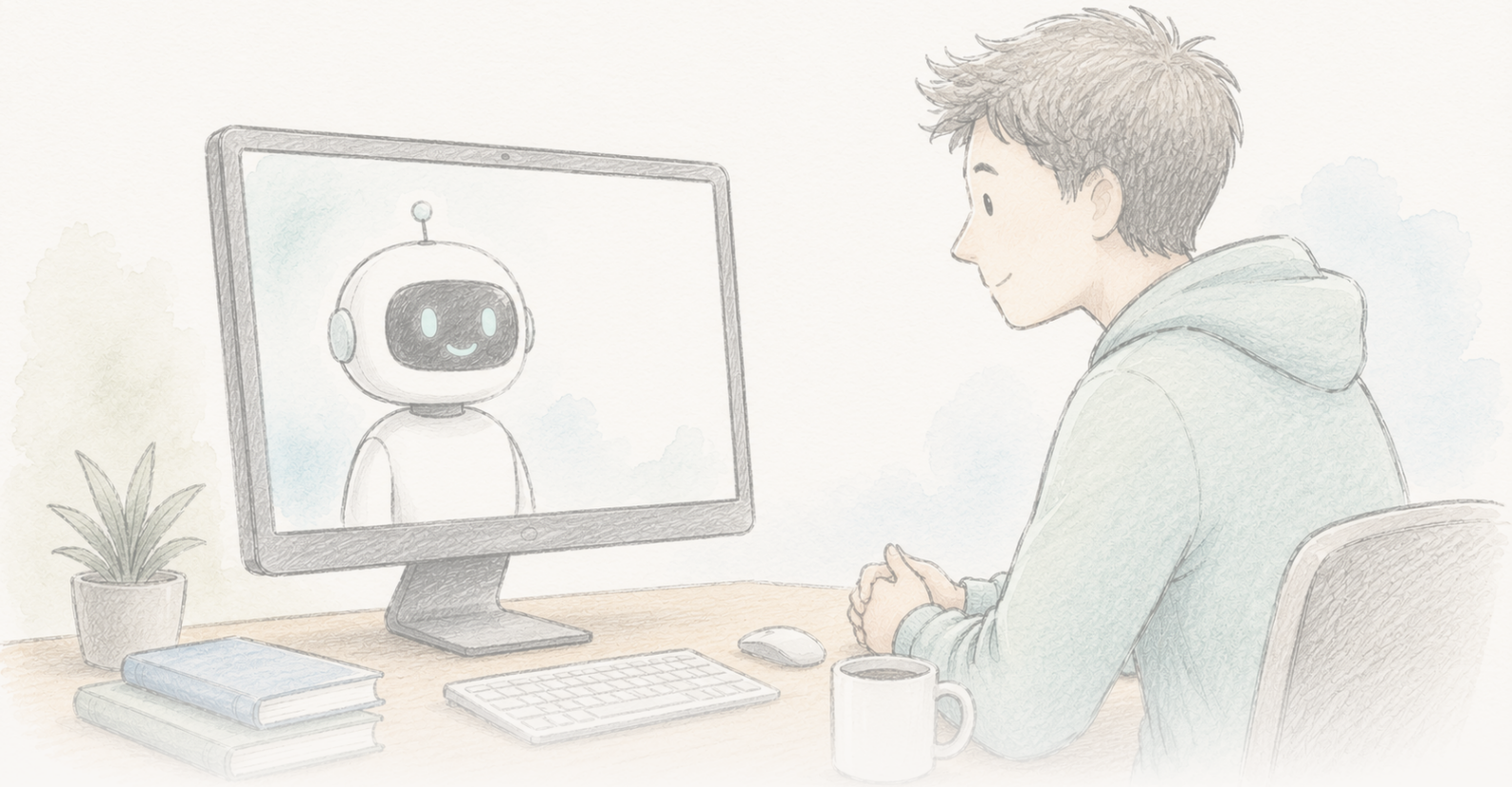


The Clean Paper

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

Sebuah chatbot tampak mengurangi kepercayaan pada teori konspirasi selama berbulan-bulan

Sebuah studi 2024 menemukan bahwa percakapan tiga ronde dengan GPT-4 Turbo menurunkan keyakinan seseorang terhadap teori konspirasi yang mereka percayai sekitar 20%, efek yang bertahan dua bulan, muncul pada berbagai jenis konspirasi, dan didukung klaim AI yang dinilai hampir seluruhnya akurat. Pada Juni 2026 paper ditempatkan di bawah Editorial Expression of Concern karena masalah penanganan data dan reproduksibilitas; para penulis mengatakan analisis yang dikoreksi mempertahankan arah, signifikansi, dan ukuran temuan, sementara Science masih mengevaluasi. Hasil yang sungguh menarik dan dibangun hati-hati — saat ini masih ditinjau, belum selesai dan belum dibantah.

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 memasang Editorial Expression of Concern pada paper ini sementara mengevaluasi pertanyaan tentang penanganan data dan reproduksibilitas. Ini bukan retraction dan bukan temuan misconduct; artinya hasil yang dilaporkan harus dibaca sebagai provisional sampai tinjauan selesai.

Editorial Expression of Concern →

Fakta ternyata bisa bekerja — dan makalah ini sedang diberi tanda peringatan

Pada 2024, sebuah studi melaporkan sesuatu yang berlawanan dengan satu anggapan nyaman yang sudah umum. Beri seseorang percakapan singkat tiga ronde dengan AI — GPT-4 Turbo, diarahkan untuk menanggapi bukti spesifik yang mereka sendiri sebutkan untuk teori konspirasi yang mereka percayai — dan, secara rata-rata, keyakinan terhadap teori itu turun sekitar 20%. Penurunan masih terlihat dua bulan kemudian. Efeknya muncul pada konspirasi klasik (pembunuhan John F. Kennedy, alien, Illuminati) maupun topik aktual (COVID-19, pemilu AS 2020), bahkan pada orang dengan keyakinan kuat dan terkait identitas. Ketika seorang pemeriksa fakta profesional menilai 128 klaim yang dibuat AI, 127 dinilai benar, satu menyesatkan, dan tidak ada yang salah.

Judul yang kemudian beredar adalah bahwa orang *dapat* diajak keluar dari rabbit hole — bahwa penganut teori konspirasi bukan kebal terhadap bukti, melainkan mungkin membutuhkan bukti yang tepat dan cukup spesifik. Itu klaim yang benar-benar menarik, dan studi ini dirancang lebih hati-hati daripada kebanyakan riset persuasi.

Lalu, pada Juni 2026, *Science* memasang **Editorial Expression of Concern** pada makalah tersebut. Bagian inilah yang kemungkinan besar akan dilewatkan banyak penceritaan ulang, padahal justru bagian ini yang menentukan seberapa berat hasil tersebut boleh dibawa sekarang.

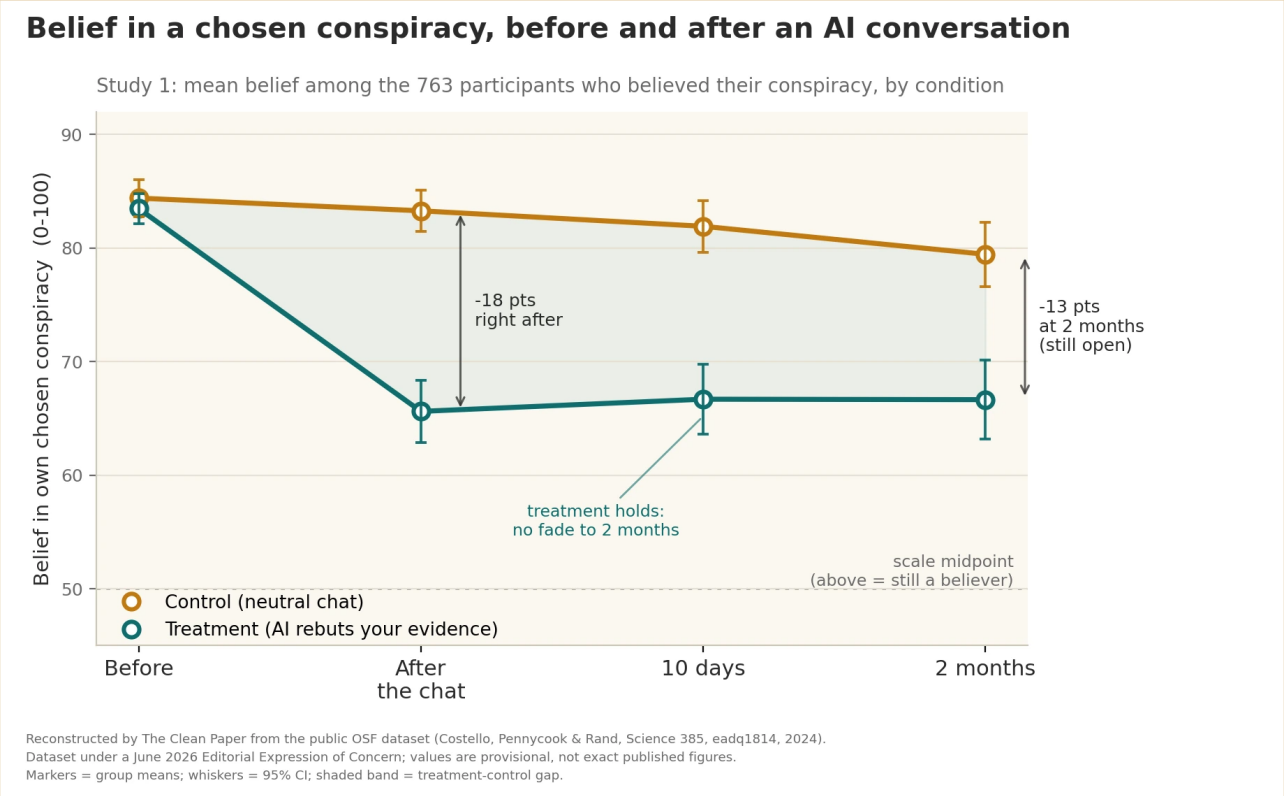
Apa itu Expression of Concern — dan apa yang bukan

Editorial Expression of Concern adalah tanda formal yang ditempelkan jurnal pada sebuah makalah untuk memperingatkan pembaca bahwa ada pertanyaan yang sedang diselidiki. Ini **bukan** retraction — makalah belum ditarik — dan bukan dengan sendirinya temuan misconduct. Artinya: baca dengan caveat ini karena catatan ilmiahnya masih dapat berubah. Dalam kasus ini, setelah diberi tahu mengenai masalah, para penulis menyelidikinya, melaporkan detail kepada Science, dan memberikan analisis yang dikoreksi.

Masalah yang ditandai cukup spesifik. Para penulis diberi tahu mengenai ketidakkonsistenan penerapan kriteria screening peserta antara manuskrip dan kode analisis yang dipublikasikan, serta baris data tambahan yang terselip dalam kumpulan data publik akibat kesalahan penggabungan kode — masalah yang membuat sebagian angka sulit direproduksi dari materi yang dirilis. Para

penulis memberikan pipeline analisis yang diperbaiki dan serangkaian hasil baru kepada *Science*, yang menurut mereka tetap sama **dalam arah, signifikansi statistik, dan besaran substantif**. *Science* masih mengevaluasinya. Sampai evaluasi itu selesai, angka persisnya tetap memiliki tanda tanya yang hanya dapat diangkat oleh jurnal.

Jadi ada dua cerita sekaligus: hasil menarik tentang apakah fakta dapat menggeser keyakinan yang mengakar, dan contoh langsung bagaimana catatan ilmiah mengoreksi dirinya secara terbuka. Tujuan artikel ini adalah menjaga keduanya tetap terpisah.



Dalam studi, percakapan AI tiga ronde yang membantah bukti yang disebut peserta sendiri dilaporkan menurunkan keyakinan terhadap konspirasi pilihannya sekitar 20% secara rata-rata; tidak seperti percakapan kontrol tentang topik tak terkait, penurunan masih terukur pada 10 hari dan 2 bulan. Efek yang dilaporkan bertahan tetapi parsial: kebanyakan peserta tetap mempercayai konspirasinya setelah itu — sebuah pengurangan, bukan penyembuhan. Makalah saat ini berada di bawah Editorial Expression of Concern (*Science*, Juni 2026) karena ketidakkonsistenan pelaporan dan kesalahan pada kumpulan data publik; para penulis mengatakan analisis yang dikoreksi mempertahankan arah, signifikansi, dan ukuran efek sementara *Science* terus mengevaluasi. Grafik ini direkonstruksi The Clean Paper dari kumpulan data publik tersebut, jadi nilai yang ditampilkan masih provisional, bukan angka final makalah.

Original figure — The Clean Paper CC BY 4.0

Apa yang dilakukan para penulis

- Menjalankan dua eksperimen dengan 2.190 peserta yang masing-masing menyebutkan — dengan kata-kata sendiri — teori konspirasi yang mereka percayai dan bukti yang menurut mereka mendukungnya.
- Setiap peserta melakukan percakapan tiga ronde dengan GPT-4 Turbo. Dalam kondisi **treatment**, AI diarahkan untuk membantah bukti spesifik peserta; pada kondisi **control**, AI membahas topik yang tidak terkait.
- Mengukur keyakinan terhadap konspirasi pilihan sebelum dan segera setelah percakapan, lalu menghubungi peserta kembali pada **10 hari dan 2 bulan**.
- Meminta pemeriksa fakta profesional menilai keakuratan seluruh 128 klaim faktual AI dalam satu sampel.
- Memeriksa apakah efek meluas ke konspirasi lain yang tidak terkait — dan sebagai tes spesifisitas, apakah ia juga menurunkan keyakinan terhadap konspirasi yang kebetulan **benar**.

Apa yang mereka temukan

- **Penurunan rata-rata sekitar 20%** pada keyakinan terhadap konspirasi pilihan dalam kelompok treatment relatif terhadap kontrol (studi 1: interval kepercayaan 95% [13,8, 19,7] poin pada skala 100 poin, $P < 0{,}001$).
- **Efek bertahan.** Dalam kelompok treatment, penurunan tidak memudar: keyakinan tetap dekat tingkat segera setelah percakapan pada 10 hari dan dua bulan, sedangkan kontrol tetap lebih tinggi. Makalah menggambarkan efek pada konspirasi pilihan bertahan tanpa berkurang selama setidaknya dua bulan, bukan goyangan sesaat.
- **Efek berlaku luas** pada berbagai jenis konspirasi yang disebut peserta dan tetap terlihat bahkan pada mereka yang keyakinan awalnya kuat.
- **Klaim AI akurat dalam setting ini:** 127 dari 128 (99,2%) dinilai benar, 1 (0,8%) menyesatkan, dan tidak ada yang salah.
- **Efek spesifik**, bukan sekadar membuat orang meragukan semuanya: intervensi **tidak** menurunkan keyakinan pada konspirasi yang benar, mengisyaratkan bahwa yang bergeser adalah keyakinan tak berdasar, bukan skeptisisme umum.
- **Ada spillover kecil** — keyakinan pada konspirasi lain yang tidak terkait turun sekitar 12%, dan peserta melaporkan niat lebih besar untuk menentang klaim konspirasi. Efek utama bertahan pada studi 2, dan bergerak ke arah yang sama pada sampel lebih kecil tanpa kelompok kontrol — bukti lebih lemah, tetapi konsisten.

Apa yang *tidak* dibuktikan

- Hasil ini **tidak** berdiri tanpa pertanyaan saat ini. Makalah berada di bawah Editorial Expression of Concern karena masalah penanganan data dan reproduksibilitas, sementara angka yang diko-

reksi masih dievaluasi *Science*. Perlakukan nilai spesifik sebagai provisional sampai tinjauan selesai.
- Studi ini **tidak** menunjukkan AI "menyembuhkan" kepercayaan konspirasi. Penurunan rata-rata ~20% adalah perubahan bermakna, bukan penghapusan — kebanyakan peserta masih percaya setelahnya, hanya lebih sedikit.
- Ini **bukan** hasil penerapan dunia nyata. Peserta memilih ikut studi dan berinteraksi dengan AI dengan itikad baik; di dunia nyata, orang yang paling terikat pada teori konspirasi justru paling kecil kemungkinannya mencari chatbot yang akan membantahnya.
- Akurasi 99,2% **khusus untuk tugas, model, dan prompting yang hati-hati ini** — bukan jaminan umum bahwa language model mengatakan hal benar. Para penulis tegas bahwa kekuatan persuasi personal yang sama dapat mendorong keyakinan *salah* jika diarahkan ke sana.
- "Bertahan" di sini berarti **dua bulan**, mengesankan untuk satu percakapan tetapi tidak sama dengan permanen.

Seberapa kuat buktinya?

- **Desainnya merupakan kekuatan nyata.** Eksperimen treatment-versus-control, kontrol bergaya plasebo yang bersih, replikasi di dua studi dan satu sampel independen, follow-up ketahanan efek, serta pemeriksaan eksplisit atas akurasi klaim AI — lebih hati-hati daripada kebanyakan riset persuasi, dengan arah efek yang konsisten di seluruh pengujian.
- **Namun Expression of Concern bukan catatan kaki.** Kemampuan menghasilkan kembali angka dari data dan kode yang dirilis merupakan bagian dari kepercayaan pada hasil, dan tepat di situlah masalah terjadi — inkonsistensi kriteria screening dan baris duplikat akibat kesalahan penggabungan kode. Pernyataan penulis bahwa pipeline yang diperbaiki mempertahankan hasil memang menenangkan dan masuk akal, tetapi masih merupakan laporan mereka sendiri, belum putusan independen. Evaluasi *Science* adalah hal yang perlu ditunggu.
- **Status jujurnya "flagged", bukan "dibantah" atau "dikonfirmasi".** Studi yang dibangun baik dan mencolok, tetapi angka persisnya sedang berada dalam tinjauan formal. Itu tempat yang tidak nyaman untuk meninggalkan cerita bagus, dan justru itulah tempat yang akurat.

Mengapa ini penting

Selama bertahun-tahun, pandangan berpengaruh mengatakan penganut teori konspirasi digerakkan kebutuhan psikologis dan identitas, sehingga tidak dapat diajak keluar dari keyakinan dengan fakta. Jika bertahan setelah tinjauan, penurunan yang tahan lama dan berbasis bukti ini menjadi penyeimbang bermakna terhadap pesimisme tersebut: mungkin masalahnya sebagian karena debunking generik tidak pernah menyentuh *bukti spesifik* yang benar-benar dikutip seseorang, sesuatu yang dapat dilakukan LLM dalam skala besar.

Hasil ini juga penting sebagai studi kasus tentang bagaimana sains seharusnya bekerja. Pelajaran dari Expression of Concern bukan bahwa hasil terkenal tidak berguna; melainkan bahwa kesalahan

diangkat, data diperiksa kembali, dan klaim diuji ulang secara terbuka. Mesin koreksi ini adalah fitur, bukan skandal — dan alasan sikap yang tepat terhadap hasil ini adalah sabar, bukan tepuk tangan atau penolakan prematur.

Dan AI-nya bermata dua. Kemampuan yang sama untuk meruntuhkan keyakinan salah dengan argumen yang disesuaikan dapat memperkuat keyakinan salah dengan sama efektifnya. Tekniknya netral; tujuannya tidak.

Singkatnya

Sebuah studi 2024 menemukan bahwa percakapan tiga ronde dengan GPT-4 Turbo menurunkan keyakinan seseorang terhadap teori konspirasi yang mereka percayai sekitar 20%, efek yang bertahan dua bulan dan muncul pada berbagai jenis konspirasi, dengan klaim faktual AI dinilai hampir semuanya akurat. Pada Juni 2026 makalah ditempatkan di bawah Editorial Expression of Concern karena masalah penanganan data dan reproduksibilitas; para penulis melaporkan bahwa analisis yang dikoreksi mempertahankan arah, signifikansi, dan ukuran temuan, sementara *Science* masih mengevaluasinya. Baca sebagai hasil yang benar-benar menarik dan dirancang hati-hati tetapi saat ini masih ditinjau — belum selesai, juga belum dibantah. Kalimat paling jujur adalah kalimat yang sedang ditulis proses ilmiah sendiri: periksa 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