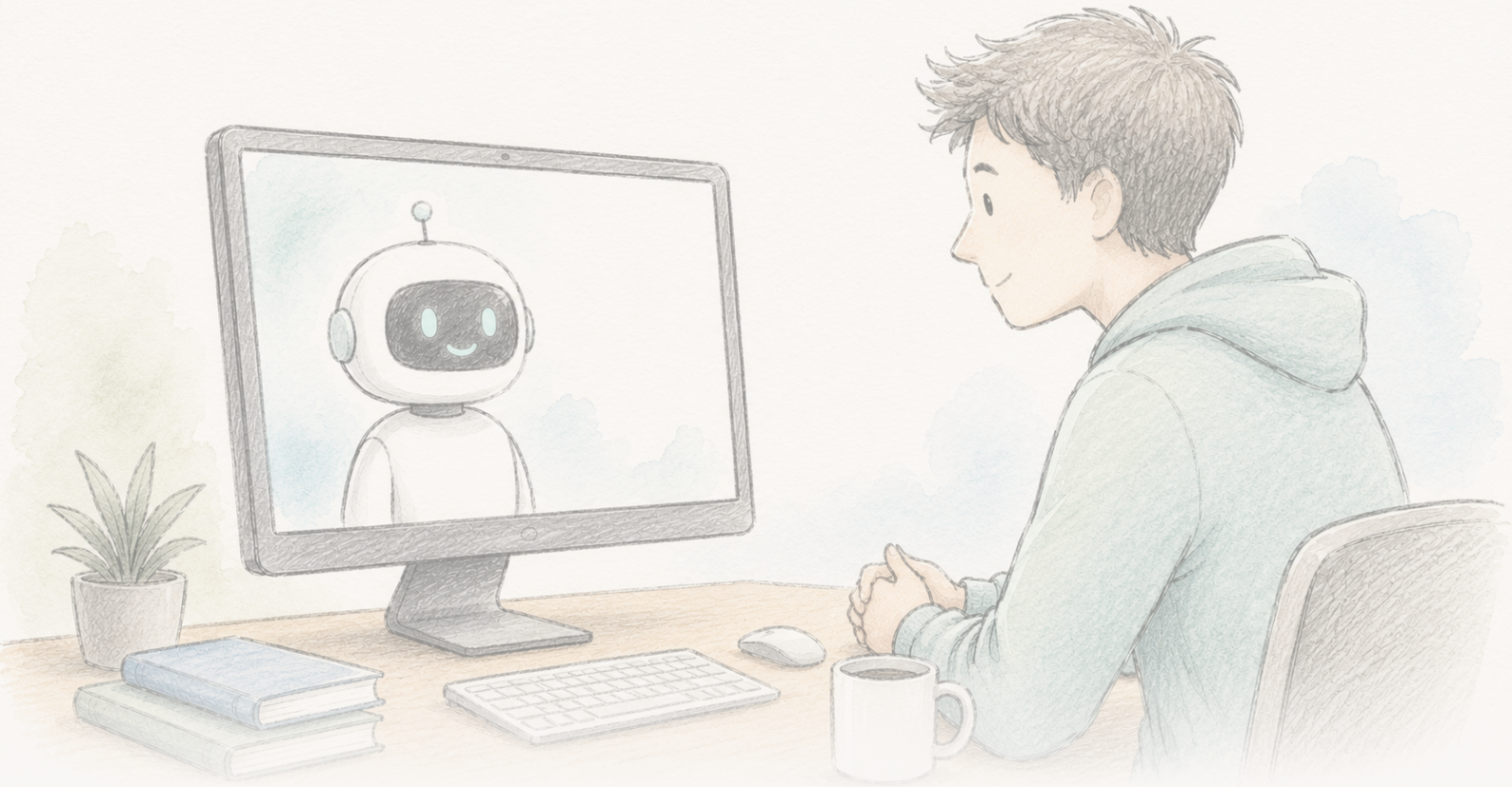


The Clean Paper

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

Bir chatbot konspirasiya inanclarını aylarla azaltmış kimi görünürdü

2024-cü il tədqiqatı GPT-4 Turbo ilə üç raundlu söhbətin insanların inandığı konspirasiya nəzəriyyəsinə inamını təxminən 20% azaltdığını, təsirin iki ay qaldığını, müxtəlif konspirasiya növlərinə yayıldığını və AI-nin iddialarının böyük əksəriyyətlə doğru qiymətləndirildiyini bildirdi. 2026-cı ilin iyununda məqalə məlumatların işlənməsi və reproduktivlik problemlərinə görə Editorial Expression of Concern altına alındı; müəlliflər düzəldilmiş analizin nəticənin istiqamətini, əhəmiyyətini və ölçüsünü qoruduğunu deyirlər, Science isə hələ qiymətləndirir. Həqiqətən maraqlı, diqqətlə qurulmuş nəticə — hazırda yoxlanılır, nə tam təsdiqlənib, nə də təkzib olunub.

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 məlumatların işlənməsi və reproduktivlik suallarını qiymətləndirərkən bu məqaləyə Editorial Expression of Concern əlavə edib. Bu retraction deyil və pozuntu barədə hökm deyil; nəticənin yoxlama həll olunana qədər provisional oxunmalı olduğunu bildirir.

Editorial Expression of Concern →

Axırda yenə faktlar — və məqalənin üzərində xəbərdarlıq bayrağı

2024-cü ildə bir tədqiqat rahat qəbul edilən fikirlərdən birinə zidd nəticə bildirdi. İnsana qısa, üç raundlu AI söhbəti verin — GPT-4 Turbo, həmin insanın şəxsən inandığı konspirasiya nəzəriyyəsi üçün göstərdiyi konkret sübutlara cavab vermək üçün prompt olunmuş şəkildə — və orta hesabla həmin nəzəriyyəyə inamı təxminən 20% azalır. Azalma iki ay sonra da qalırdı. Bu, klassik konspirasiyalarda (John F. Kennedy-nin qətli, yadplanetlilər, Illuminati) də, aktual mövzularda (COVID-19, ABŞ-ın 2020 seçkisi) da, hətta inamı güclü və kimliyinə bağlı olan insanlarda da işləyirdi. Peşəkar fakt-yoxlayıcı AI-nin etdiyi 128 iddianı qiymətləndirəndə 127-si doğru, biri yanıltıcı, heç biri yalan deyildi.

Yayılan başlıq belə idi: insanı *rabbit hole*-dan danışaraq çıxarmaq mümkündür — konspirasiya inananları sübutun təsir dairəsindən kənarda deyil, sadəcə onlara kifayət qədər konkret, düzgün sübut lazımdır. Bu, həqiqətən maraqlı iddiadır və tədqiqat inandırma işlərinin çoxundan daha diqqətlə qurulmuşdu.

Sonra, 2026-cı ilin iyununda *Science* məqaləyə **Editorial Expression of Concern** əlavə etdi. Əksər təkrar hekayələrin keçəcəyi hissə məhz budur və nəticənin hazırda nə qədər ağırlıq daşıya biləcəyini müəyyən edən də elə budur.

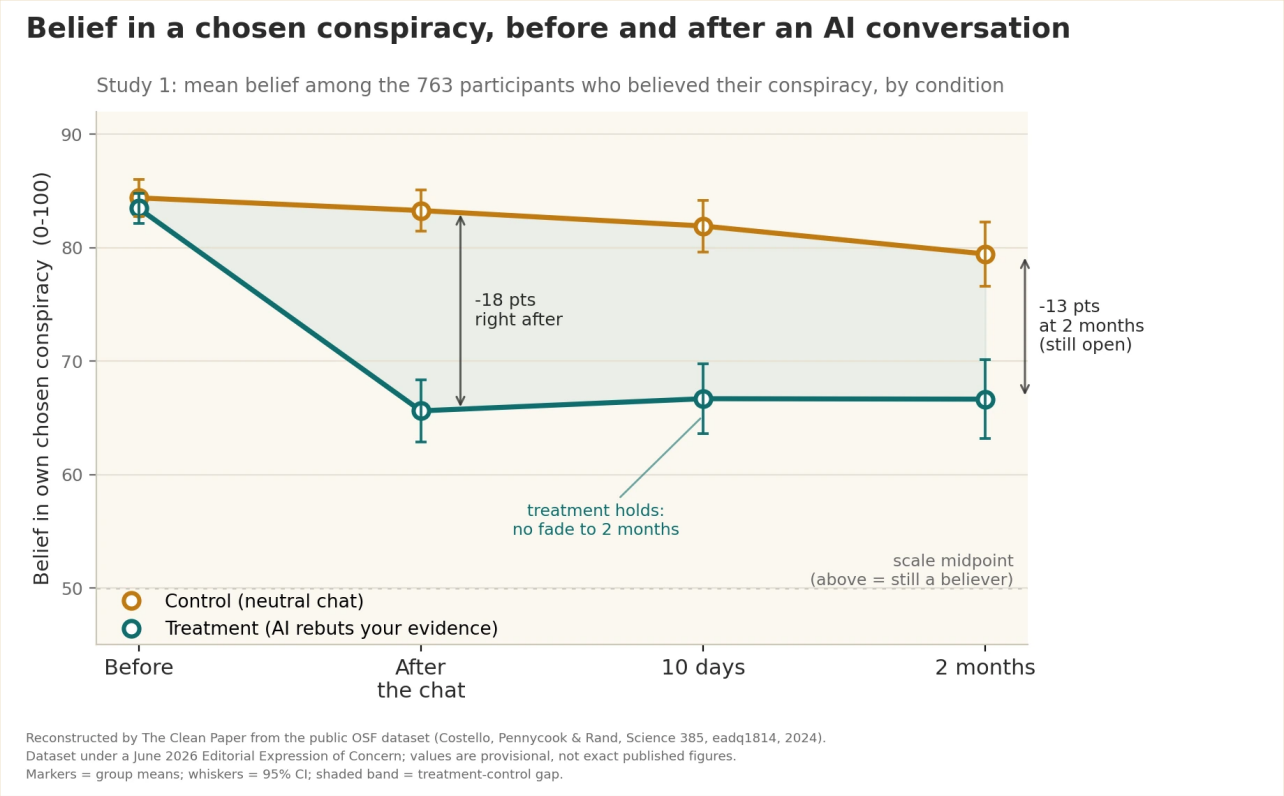
Expression of Concern nədir — və nə deyil

Editorial Expression of Concern jurnalın suallar ortaya çıxarkən və hələ araşdırılarkən oxucuları xəbərdar etmək üçün məqaləyə əlavə etdiyi rəsmi işarədir. Bu, **retraction deyil** (məqalə geri çəkilməyib) və öz-özlüyündə elmi pozuntu barədə hökm deyil. Mənası budur: qeydi ehtiyatla oxuyun, çünki elmi rekord dəyişə bilər. Burada müəlliflər problemlər barədə məlumat aldıqdan sonra araşdırma apardılar, detalları *Science*-a bildirdilər və düzəldilmiş analiz təqdim etdilər.

Bayraqlanan problem konkret idi. Müəlliflər iştirakçı skrininq meyarlarının əlyazma ilə yayımlanmış analiz kodunda fərqli tətbiq edildiyindən və açıq dataset-də kodların birləşdirilməsi xətası nəticəsində əlavə sətirlərin araya qarışdığından xəbərdar oldular — bu problemlər yayımlanmış materiallardan bəzi nəticələri təkrar çıxarmağı çətinləşdirirdi.

Müəlliflər *Science*-a düzəldilmiş analiz pipeline-ı və yenilənmiş nəticələr verdilər; onların sözlərinə görə yeni nəticələr orijinalla **istiqamət, statistik əhəmiyyət və praktik ölçü baxımından** üst-üstə düşür. *Science* bunları qiymətləndirir. Qiymətləndirmə bitənə qədər dəqiq rəqəmlər yalnız jurnalın götürə biləcəyi sual işarəsinin altındadır.

Beləliklə, eyni vaxtda iki hekayə var: faktların kök salmış inancları dəyişdirib-dəyişdirə bilməsi barədə maraqlı nəticə və elmi qeydin açıq şəkildə özünü düzəltməsinin canlı nümunəsi. Bu məqalənin məqsədi ikisini bir-birindən ayırmaqdır.



Tədqiqatda iştirakçının öz göstərdiyi sübutu təkzib edən üç raundlu AI söhbətinin seçdiyi konspirasiya nəzəriyyəsinə inamı orta hesabla təxminən 20% azaltdığı bildirilmişdi; əlaqəsiz mövzudakı nəzarət söhbətindən fərqli olaraq azalma 10 gün və 2 ay sonra da ölçülürdü. Bildirilən təsir davamlı, amma qismən idi: iştirakçıların çoxu sonradan da konspirasiyaya inanırdı — müalicə yox, azalma. Məqalə hazırda hesabat uyğunsuzluqları və açıq dataset xətası səbəbindən Editorial Expression of Concern altındadır (*Science*, iyun 2026); müəlliflər düzəldilmiş analizin təsirin istiqamətini, statistik əhəmiyyətini və ölçüsünü qoruduğunu deyirlər, *Science* isə qiymətləndirməni davam etdirir. Bu diaqram The Clean Paper tərəfindən həmin açıq dataset-dən yenidən qurulub, ona görə göstərilən qiymətlər provisionaldır, məqalənin yekun rəqəmləri deyil.

Original figure — The Clean Paper CC BY 4.0

Müəlliflər nə etdilər

- Hər biri öz sözləri ilə inandığı konspirasiya nəzəriyyəsini və onu dəstəklədiyini düşündüyü sübutu yazan 2190 iştirakçı ilə iki eksperiment apardılar.

- Hər iştirakçıya GPT-4 Turbo ilə üç raundlu söhbət etdirdilər. **Müdaxilə** şəraitində AI iştirakçının konkret sübutunu təkzib etmək üçün prompt olunurdu; **nəzarət** şəraitində isə əlaqəsiz mövzu müzakirə edilirdi.
- Seçilmiş konspirasiyaya inamı söhbətdən əvvəl və dərhal sonra ölçdülər, sonra iştirakçılarla **10 gün və 2 ay** sonra yenidən əlaqə saxladılar.
- Peşəkar fakt-yoxlayıcıya nümunədə AI-nin etdiyi bütün 128 fakt iddiasının düzgünlüyünü qiymətləndirdilər.
- Təsirin başqa, əlaqəsiz konspirasiyalara yayılıb-yayılmadığını və spesifiklik testi kimi, təsadüfən **doğru olan** konspirasiyalara inamı da azaldıb-azaltmadığını yoxladılar.

Nə tapdılar

- Nəzarətlə müqayisədə müdaxilə qrupunda seçilmiş konspirasiyaya inamda **orta hesabla təxminən 20% azalma** (tədqiqat 1: 100 ballıq şkalada 95% etibar intervalı [13.8, 19.7] bal, $P < 0.001$).
- **Təsir qaldı.** Müdaxilə qrupunda azalma itmirdi: 10 gün və iki ay sonra inam söhbətdən dərhal sonrakı səviyyəyə yaxın qalırdı, nəzarət qrupu isə bütün müddət daha yüksək idi — məqalə seçilmiş konspirasiyadakı təsiri ani tərəddüd yox, ən azı iki ay azalmadan davam edən effekt kimi təsvir edir.
- **Ümumiləşdi:** insanların adlandırdığı müxtəlif konspirasiyalarda və başlanğıc inamı güclü olan iştirakçılarda da nəticə saxlanıldı.
- **AI-nin iddiaları bu şəraitdə düzgün idi:** 128-dən 127-si (99.2%) doğru, 1-i (0.8%) yanıltıcı, heç biri yalan qiymətləndirildi.
- **Spesifik idi**, ümumi şübhəçilik deyildi: müdaxilə doğru konspirasiyalara inamı **azaltmadı**; bu, insanları hər şeyə qarşı sinik etmək əvəzinə əsassız inancları dəyişdirdiyini göstərir.
- **Müəyyən qədər yayıldı** — başqa, əlaqəsiz konspirasiyalara inam təxminən 12% azaldı və iştirakçılar konspirasiya iddialarına etiraz etmək niyyətinin artdığını bildirdilər. Əsas effekt tədqiqat 2-də də qaldı və nəzarət qrupu olmayan daha kiçik nümunədə də eyni istiqamətə işarə etdi (daha zəif sübut, amma uyğundur).

Bu nəyi sübut etmir

- Hazırda nəticə **sualsız qəbul edilmir**. Məqalə məlumatların işlənməsi və reproduktivlik problemlərinə görə Editorial Expression of Concern altındadır və düzəldilmiş rəqəmlər hələ *Science* tərəfindən qiymətləndirilir. Həmin yoxlama bitənə qədər konkret rəqəmləri provisional sayın.
- AI-nin konspirasiya inamını "**müalicə etdiyini**" göstərmir. ~20% orta azalma mənalı dəyişiklikdir, silinmə deyil — iştirakçıların çoxu sonradan nəzəriyyəyə yenə inanırdı, sadəcə daha az.
- Bu, **real dünyada tətbiq nəticəsi deyil**. İştirakçılar tədqiqata könüllü qoşulub AI ilə

yaxşı niyyətlə söhbət etdilər; real həyatda konspirasiyaya ən çox bağlı olanların onlarla mübahisə edəcək chatbot axtarma ehtimalı ən aşağı ola bilər.
- 99.2% düzgünlük **bu tapşırığa, modelə və diqqətli prompting-ə xasdır** — dil modellərinin ümumən doğru danışdığına zəmanət deyil. Müəlliflər açıq deyirlər ki, eyni fərdiləşdirilmiş inandırma gücü yanlış inancları da, məqsəd o olsa, gücləndirə bilər.
- "Davamlı" burada **iki ay** deməkdir; bir söhbət üçün təsirlidir, amma daimi demək deyil.

Sübut nə qədər güclüdür

- **Dizayn həqiqi güclü tərəfdir.** Nəzarətli müdaxilə-nəzarət eksperimentləri, təmiz placebo-tipli nəzarət, iki tədqiqat və müstəqil nümunədə replikasiya, davamlılıq follow-up-ları və AI-nin öz iddialarının düzgünlüyünün açıq yoxlanması — bunlar inandırma tədqiqatlarının çoxundan daha diqqətlidir və təsirin istiqaməti sınandığı hər yerdə eynidir.
- **Amma Expression of Concern dipnot deyil.** Yayımlanmış data və koddan bildirildiyi rəqəmləri yenidən əldə edə bilmək nəticəni etibarlı edən şeylərin bir hissəsidir və məhz bu pozuldu — skrininq meyarlarının uyğunsuzluğu və kod birləşməsi səhvindən yaranmış təkrarlanan/əlavə sətirlər. Müəlliflərin düzəldilmiş pipeline nəticəni qoruyur deməsi ümidverici və mümkündür, amma hələ müstəqil hökm deyil, müəlliflərin öz hesabatıdır. Gözlənməli şey *Science*-ın qiymətləndirməsidir.
- **Dürüst status "bayraqlanıb"dır, "təkzib olunub" və ya "təsdiqlənib" yox.** Yaxşı qurulmuş, diqqətçəkən tədqiqatın dəqiq rəqəmləri rəsmi yoxlamadadır. Yaxşı hekayəni belə yarımçıq saxlamaq rahat deyil, amma doğru vəziyyət budur.

Niyə vacibdir

İllərlə təsirli bir baxış konspirasiya inananlarının psixoloji ehtiyaclar və kimlik tərəfindən idarə olunduğunu, buna görə də faktlarla inanclarından daşındırıla bilməyəcəyini deyirdi. Davamlı, sübuta əsaslanan azalma — nəticə yoxlamadan keçsə — bu pessimizmə mənalı əks çəki olar: problem qismən generic debunking-in insanın həqiqətən göstərdiyi *konkret* sübutla heç vaxt məşğul olmaması ola bilər; LLM isə bunu miqyasda edə bilir.

Bu, elmin necə işləməli olduğuna dair nümunə kimi də vacibdir. Expression of Concern-un dərsi məşhur nəticələrin dəyərsiz olması deyil; səhvlərin üzə çıxması, məlumatların yenidən nəzərdən keçirilməsi və iddiaların açıq şəkildə təkrar yoxlanmasıdır. Bu mexanizmin işləməsi qalmaqal yox, sistemin xüsusiyyətidir — və nəticəyə doğru münasibətin alqış və ya rədd yox, səbir olmasının səbəbi budur.

AI baxımından da iki tərəfli qılıncdır. Fərdiləşdirilmiş arqumentlə yanlış inamı söndürmək qabiliyyəti eyni dərəcədə onu şişirdə bilər. Texnika neytraldır; məqsəd deyil.

Təmiz xülasə

2024-cü il tədqiqatı GPT-4 Turbo ilə üç raundlu söhbətin insanların inandığı konspirasiya nəzəriyyəsinə inamını təxminən 20% azaltdığını, təsirin iki ay qaldığını və müxtəlif konspirasiya növlərinə yayıldığını bildirdi; AI-nin fakt iddiaları böyük əksəriyyətlə doğru qiymətləndirilmişdi. 2026-cı ilin iyununda məqalə məlumatların işlənməsi və reproduktivlik problemlərinə görə Editorial Expression of Concern altına alındı; müəlliflər düzəldilmiş analizin nəticənin istiqamətini, statistik əhəmiyyətini və ölçüsünü qoruduğunu bildirirlər, *Science* isə hələ qiymətləndirir. Bunu hazırda yoxlanılan, həqiqətən maraqlı və diqqətlə qurulmuş nəticə kimi oxuyun — nə bağlanmış məsələ, nə də təkzib. Bu barədə ən dürüst cümləni elmi prosesin özü yazır: yenidən yoxlayın.

Mənbələr

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