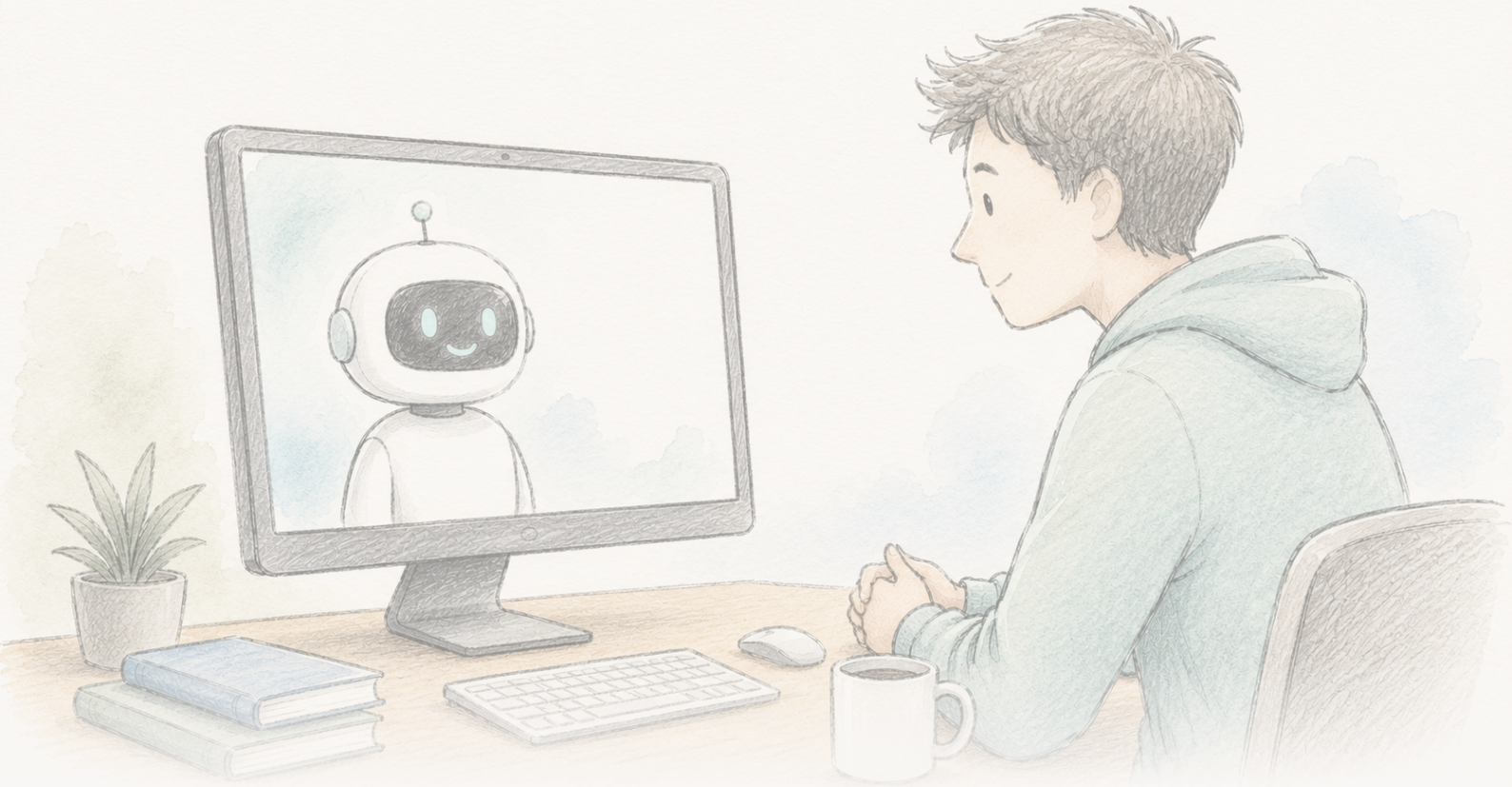


The Clean Paper

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

Bir chatbot komplo inançlarını aylar boyunca azaltmış gibi göründü

2024 tarihli bir çalışma, GPT-4 Turbo ile üç turluk konuşmanın insanların inandığı bir komplo teorisine inancı yaklaşık %20 azalttığını; etkinin iki ay sürdüğünü, farklı komplo türlerinde görüldüğünü ve ezici çoğunlukla doğru puanlanan yapay zekâ iddialarına dayandığını buldu. Haziran 2026'da makale veri işleme ve yeniden üretilebilirlik sorunları nedeniyle Editorial Expression of Concern altına alındı; yazarlar düzeltilmiş analizin sonucu yön, anlamlılık ve büyüklük bakımından koruduğunu söylüyor ve Science hâlâ değerlendiriyor. Gerçekten ilginç, dikkatle kurulmuş bir sonuç — şu anda inceleme altında; ne yerleşmiş ne çürütülmüş.

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, veri işleme ve yeniden üretilebilirlik sorularını değerlendirirken bu makaleye Editorial Expression of Concern ekledi. Bu bir retraction değildir ve suistimal bulgusu değildir; inceleme çözülene kadar bildirilen sonucun geçici okunması gerektiği anlamına gelir.

Editorial Expression of Concern →

Sonuçta gerçekler — ve makalenin üzerinde bir uyarı bayrağı

2024'te yayımlanan bir çalışma, rahatlatıcı bir genel kabule ters düşen bir şey bildirdi. Bir kişiye, gerçekten inandığı bir komplo teorisi için kendi sunduğu kanıtlara özel yanıt vermesi istenmiş GPT-4 Turbo ile kısa, üç turluk bir konuşma yaptırın; ortalama olarak o teoriye inancı yaklaşık %20 azalıyor. Düşüş iki ay sonra da duruyordu. Etki klasik komplolarda (John F. Kennedy suikastı, uzaylılar, Illuminati) ve güncel olanlarda (COVID-19, 2020 ABD seçimi) görüldü; başlangıçta inancı güçlü ve kimliğiyle bağlantılı kişilerde bile. Profesyonel bir doğruluk denetçisi yapay zekânın ileri sürdüğü 128 iddiayı değerlendirdiğinde 127'sini doğru, birini yanıltıcı buldu; hiçbiri yanlış değildi.

Dolaşıma giren başlık, insanları tavşan deliğinden *kanıtla konuşarak çıkarabileceğiniz* oldu — komplo inananlarının kanıta bütünüyle kapalı olmadığı, yalnızca doğru ve yeterince kişiselleştirilmiş kanıta ihtiyaç duyduğu. Bu gerçekten ilginç bir iddia ve çalışma ikna araştırmalarının çoğundan daha dikkatli kurulmuştu.

Sonra Haziran 2026'da *Science* makaleye bir **Editorial Expression of Concern** ekledi. Çoğu yeniden anlatımın atlayacağı kısım bu ve sonucun şu anda ne kadar ağırlık taşıyabileceğine karar veren de tam olarak bu.

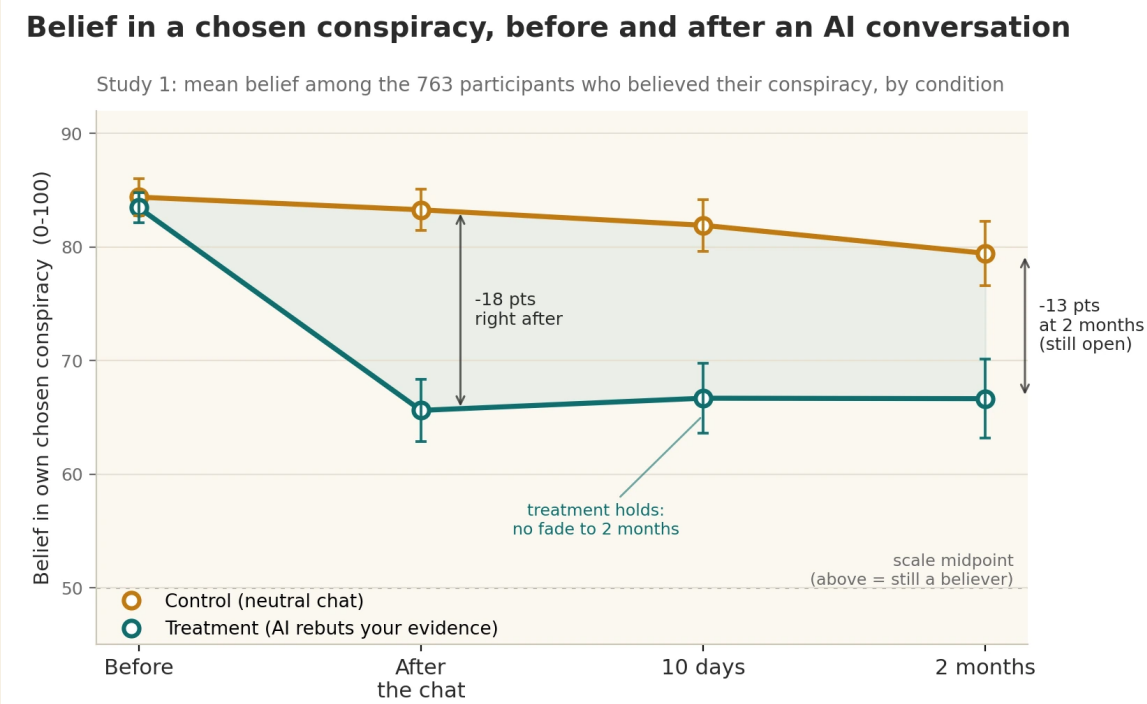
Expression of Concern nedir — ve ne değildir?

Editorial Expression of Concern, bir derginin bir makaleye, sorular ortaya çıktığını fakat bunların hâlâ incelendiğini okuyucuya bildirmek için eklediği resmi uyarıdır. **Retraction değildir** (makale geri çekilmiş değildir) ve tek başına suistimal bulgusu değildir. Anlamı şudur: kayıt değişebileceği için bunu uyarıyı akılda tutarak okuyun. Burada yazarlar sorunlardan haberdar edildikten sonra konuyu araştırdı, ayrıntıları *Science*'a bildirdi ve düzeltilmiş bir analiz sundu.

İşaretlenen sorunlar spesifik. Yazarlar, katılımcı eleme ölçütlerinin makale ile yayımlanan analiz kodu arasında tutarsız uygulanmış olduğundan ve kamuya açık veri setinde kod birleştirme hatası nedeniyle araya yanlışlıkla eklenmiş fazladan satırlar bulunduğundan haberdar edildi — bu sorunlar yayımlanan bazı sayıların açık materyallerden yeniden üretilmesini zorlaştırdı. Yazarlar *Science*'a düzeltilmiş analiz hattını ve güncellenmiş sonuçları verdi; bunların orijinal sonuçlarla **yön, istatistiksel anlamlılık ve esas büyüklük bakımından uyuştuğunu** söylüyorlar. *Science* bunları

değerlendiriyor. Bu değerlendirme bitene kadar kesin rakamların üzerinde yalnızca derginin kaldırabileceği bir soru işareti var.

Dolayısıyla burada aynı anda iki hikâye var: gerçeklerin kökleşmiş inançları hareket ettirip ettiremeyeceğine dair ilginç bir sonuç ve bilimsel kaydın kendini açıkça düzeltmesinin canlı bir örneği. Bu yazının amacı ikisini birbirine karıştırmamak.



Çalışmada, katılımcının kendi sunduğu kanıtı çürüten üç turluk yapay zekâ konuşmasının o kişinin seçtiği komploya inancı ortalama yaklaşık %20 azalttığı bildirildi; alakasız bir konudaki kontrol sohbetinden farklı olarak düşüş 10 gün ve 2 ay sonra da ölçülebiliyordu. Bildirilen etki kalıcı ama kısmiydi: katılımcıların çoğu daha sonra da komploya inanıyordu — tedavi değil, azalma. Makale şu anda raporlama tutarsızlıkları ve kamuya açık veri setindeki bir hata nedeniyle Editorial Expression of Concern altındadır (*Science*, Haziran 2026); yazarlar düzeltilmiş analizin etkinin yönünü, anlamlılığını ve büyüklüğünü koruduğunu söylüyor, *Science* ise değerlendirmeyi sürdürüyor. Bu grafik The Clean Paper tarafından kamuya açık veri setinden yeniden oluşturuldu; dolayısıyla gösterilen değerler geçicidir, makalenin nihai rakamları değildir.

Original figure — The Clean Paper CC BY 4.0

Yazarlar ne yaptı?

- Her biri kendi sözleriyle inandığı bir komplo teorisini ve onu desteklediğini düşündüğü kanıtı belirten 2.190 katılımcıyla iki deney yürüttüler.
- Her katılımcıya GPT-4 Turbo ile üç turluk konuşma yaptırdılar. **Müdahale** koşulunda yapay zekâ katılımcının belirli kanıtlarını çürütmek üzere yönlendirildi; **kontrol** koşulunda alakasız bir konu

konuşuldu.
- Seçilen komploya inancı konuşmadan önce ve hemen sonra ölçtüler, ardından katılımcılarla **10 gün ve 2 ay** sonra yeniden iletişim kurdular.
- Profesyonel bir doğruluk denetçisine, örneklemde yapay zekânın ileri sürdüğü 128 olgusal iddianın tamamının doğruluğunu değerlendirdiler.
- Etkinin başka, ilgisiz komplolara taşıp taşmadığını — ve özgüllük testi olarak tesadüfen **doğru olan** komplolara inancı da azaltıp azaltmadığını — kontrol ettiler.

Ne buldular?

- Kontrole göre müdahale grubunda seçilen komploya inançta **ortalama yaklaşık %20 azalma** (çalışma 1: 100 puanlık ölçekte %95 güven aralığı [13,8, 19,7] puan, $P < 0,001$).
- **Sürdü.** Müdahale grubunda düşüş geri dönmedi: inanç 10 gün ve iki ayda konuşma-sonrası düzeyine yakın kaldı; kontrol boyunca daha yüksekti — makale seçilen komplo üzerindeki etkinin en az iki ay boyunca azalmadan sürdüğünü, anlık bir sallantı olmadığını söylüyor.
- İnsanların adlandırdığı komplo türlerinin genelinde **genellendi** ve başlangıçtaki inancı güçlü katılımcılarda bile görüldü.
- **Yapay zekânın iddiaları bu ortamda doğruydu:** 128'in 127'si (%99,2) doğru, 1'i (%0,8) yanıltıcı, hiçbiri yanlış olarak puanlandı.
- **Özgüldü**, genel şüphecilik değildi: müdahale doğru komplolara inancı **azaltmadı**; bu, insanları her şey hakkında alaycı yapmak yerine desteklenmeyen inançları hareket ettirdiğini düşündürüyor.
- **Mütevazı biçimde başka alanlara da taştı** — ilgisiz başka komplolara inanç yaklaşık %12 düştü ve katılımcılar komplo iddialarına karşı çıkma niyetlerinin arttığını bildirdi. Ana etki çalışma 2'de de kaldı ve kontrol grubu olmayan daha küçük bir örneklemde aynı yöne işaret etti (daha zayıf kanıt ama tutarlı).

Bunun kanıtlamadığı şeyler

- Şu anda **sorgusuz bir sonuç değildir**. Makale veri işleme ve yeniden üretilebilirlik sorunları nedeniyle Editorial Expression of Concern altında; düzeltilmiş sayılar hâlâ *Science* tarafından değerlendiriliyor. İnceleme bitene kadar spesifik değerleri geçici kabul edin.
- Yapay zekânın komplo inancını "**tedavi ettiğini**" göstermez. Ortalama ~%20 azalma anlamlı bir değişimdir, silme değil — çoğu katılımcı daha sonra da teoriye inanıyordu, yalnızca daha az.
- Bu **gerçek dünya kullanım sonucu değildir**. Katılımcılar çalışmaya gönüllü oldu ve yapay zekâ ile iyi niyetle etkileşti; sahada bir komploya en bağlı kişilerin kendileriyle tartışacak bir chatbota gönüllü gitme olasılığı en düşüktür.
- %99,2 doğruluk **bu görev, model ve dikkatli isteme özgüdür** — dil modellerinin genelde doğru konuştuğuna dair garanti değil. Yazarlar aynı kişiselleştirilmiş ikna gücünün yanlış inançları da hedeflendiğinde büyütebileceğini açıkça söylüyor.
- "Kalıcı" burada **iki ay** demek; tek konuşma için etkileyici ama kalıcı/sonsuz demek değil.

Kanıt ne kadar güçlü?

- **Tasarım gerçek bir güçlü yan.** Müdahale-kontrol deneyleri, temiz placebo benzeri kontrol, iki çalışma ve bağımsız örneklemde tekrar, kalıcılık takipleri ve yapay zekânın kendi iddialarına açık doğruluk kontrolü — bu, ikna araştırmalarının çoğundan daha dikkatli ve etkinin yönü test edildiği her yerde tutarlı.
- **Ama Expression of Concern dipnot değildir.** Yayımlanan sayıları açık veri ve koddan yeniden üretebilmek sonucun güvenilirliğinin bir parçasıdır ve bozulan şey tam olarak budur — eleme ölçütü tutarsızlıkları ve kod-birleştirme hatasından çoğaltılmış satırlar. Yazarların düzeltilmiş hattın sonucu koruduğu açıklaması cesaret verici ve makul, ama şu anda yazarların kendi anlatımı; bağımsız hüküm değil. Beklenmesi gereken *Science*'ın değerlendirmesi.
- **Dürüst durum "işaretli"; ne "çürütüldü" ne "onaylandı".** İyi kurulmuş, çarpıcı bir çalışma; kesin rakamları resmi incelemede. İyi bir hikâyeyi bu rahatsız edici yerde bırakmak zor, ama doğru yer burası.

Neden önemli?

Yıllardır etkili bir görüş, komplo inananlarının psikolojik ihtiyaçlar ve kimlik tarafından yönlendirildiğini ve bu yüzden gerçeklerle ikna edilemeyeceklerini savundu. Dayanıklı, kanıta dayalı bir azalma — eğer doğrulanırsa — bu kötümserliğe anlamlı karşı ağırlık olur: sorunun kısmen genel çürütmelerin kişinin gerçekten dayandığı *özgül* kanıtla hiç uğraşmaması olduğunu, LLM'nin bunu ölçekli biçimde yapabildiğini düşündürür.

Bilimin nasıl işlemesi gerektiğine dair vaka olarak da önemli. Expression of Concern'ın dersi ünlü sonuçların değersiz olduğu değil; hataların yüzeye çıkması, verinin yeniden incelenmesi ve iddiaların kamu önünde yeniden kontrol edilmesi. Bu mekanizmanın çalışması skandal değil, özelliktir — ve sonuç karşısındaki doğru tutumun alkış ya da reddetme değil sabır olmasının nedeni.

Yapay zekâ açısından da iki yönlü. Özelleştirilmiş bir argümanla yanlış inancı söndürebilen aynı kapasite, yanlış bir inancı aynı etkililikle şişirebilir. Teknik nötr; amaç değil.

Kısa özet

2024 tarihli bir çalışma, GPT-4 Turbo ile üç turluk konuşmanın insanların inandığı bir komplo teorisine inancı yaklaşık %20 azalttığını; etkinin iki ay sürdüğünü ve komplo türleri arasında genellendiğini, yapay zekânın olgusal iddialarının da ezici çoğunlukla doğru puanlandığını buldu. Haziran 2026'da makale veri işleme ve yeniden üretilebilirlik sorunları nedeniyle Editorial Expression of Concern altına alındı; yazarlar düzeltilmiş analizin sonucu yön, anlamlılık ve büyüklük bakımından koruduğunu bildiriyor ve *Science* hâlâ değerlendiriyor. Bunu şu anda inceleme altında olan, gerçekten ilginç ve dikkatle kurulmuş bir sonuç olarak okuyun — ne yerleşmiş ne çürütülmüş. Hakkındaki en dürüst cümle bilimsel sürecin kendisinin yazdığı: bir daha kontrol et.

Kaynaklar

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