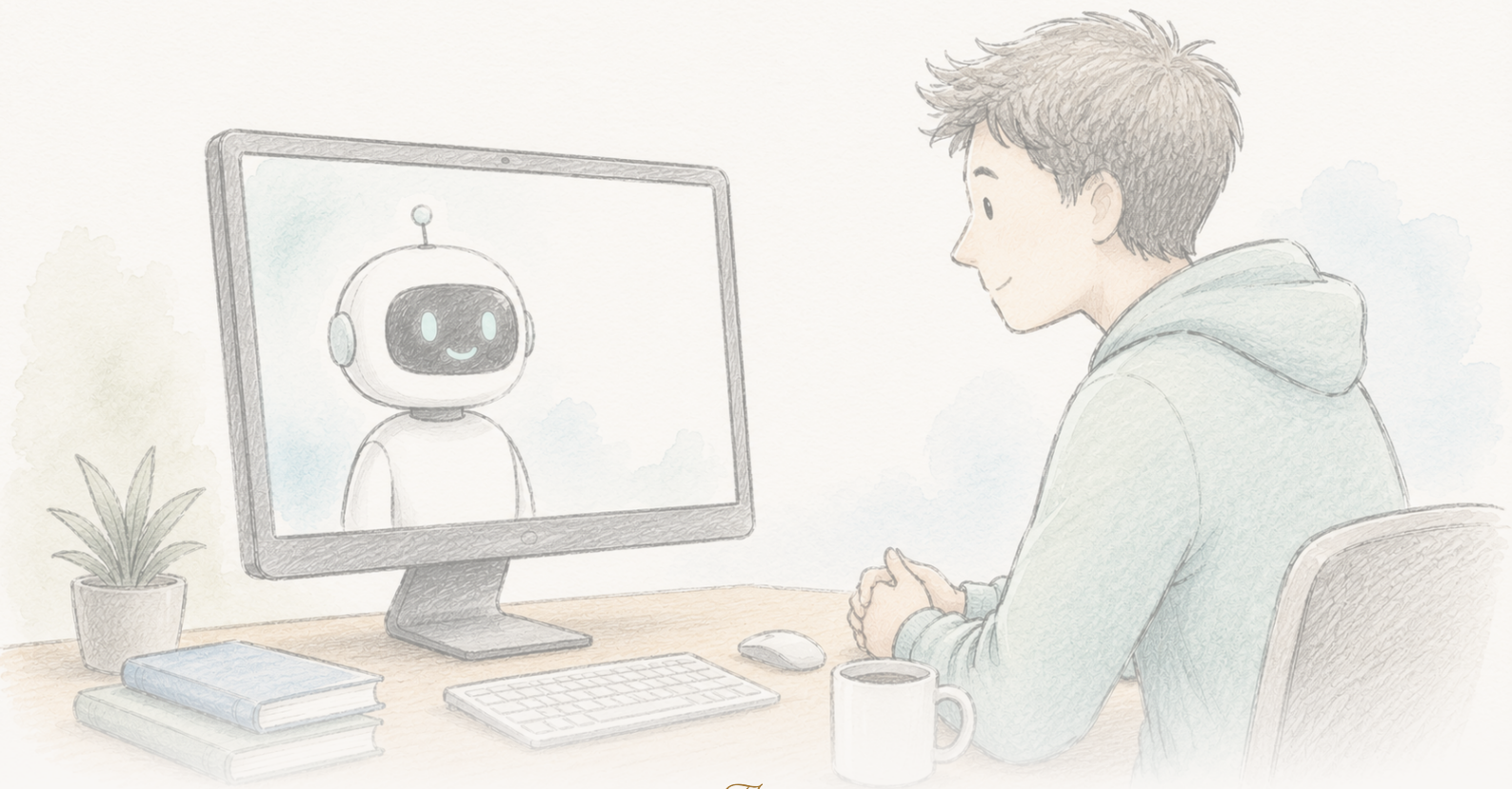




The CLEAN PAPER

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

Chatbot-ը կարծես ամիսներով նվազեցրել է conspiracy belief-ը

2024 թվականի ուսումնասիրությունը գտել էր, որ GPT-4 Turbo-ի հետ երեք փուլով զրույցը մարդկանց մոտ մոտ 20%-ով նվազեցնում է իրենց հավատացած *conspiracy theory*-ի նկատմամբ համոզվածությունը, *effect*-ը պահպանվում է երկու ամիս և տարբեր *conspiracy type*-երի միջև, իսկ AI-ի *factual claim*-ները հիմնականում գնահատվել են ճիշտ: 2026 թվականի հունիսին հոդվածին կցվեց *Editorial Expression of Concern` data-handling ու reproducibility խնդիրների պատճառով*: Հեղինակները ասում են, որ *corrected analysis*-ը պահպանում է արդյունքի ուղղությունը, *significance*-ը և չափը, իսկ *Science*-ը դեռ գնահատում է այն: Հետաքրքիր և լավ կառուցված արդյունք՝ ներկայում *review*-ի տակ, ոչ հաստատված ու ոչ *debunked*:

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-ը հոդվածին կցել է Editorial Expression of Concern, մինչ գնահատում է data-handling և reproducibility հարցերը: Սա retraction չէ և misconduct-ի եզրակացություն չէ. reported result-ը պետք է համարել provisional մինչև review-ի ավարտը:

Editorial Expression of Concern →

Փաստերը, այնուամենայնիվ — և պաշտոնական զգուշացման նշում հոդվածի վրա

2024 թվականին հրապարակված ուսումնասիրությունը հակասեց մի հարմարավետ, տարածված համոզմունքի: Եթե մարդուն տալ կարճ՝ երեք փուլից բաղկացած զրույց ԱԲի հետ՝ GPT-4 Turbo, որը հրահանգված է պատասխանելու հենց այն կոնկրետ ապացույցներին, որոնցով մարդը հիմնավորում է իր հավատացած դավադրության տեսությունը, ապա միջինում այդ տեսության նկատմամբ հավատը նվազում է մոտ 20%-ով: Նվազումը պահպանվել էր նաև երկու ամիս անց: Արդյունքը երևացել էր թե՛ դասական դավադրությունների դեպքում՝ John F. Kennedy-ի սպանություն, այլմոլորակայիններ, Illuminati, թե՛ արդիական թեմաների դեպքում՝ COVID-19, 2020 թվականի ԱՄՆ ընտրություններ, և նույնիսկ այն մարդկանց մոտ, որոնց համոզմունքը սկզբում ուժեղ էր ու կապված ինքնության հետ: Երբ պրոֆեսիոնալ fact-checker-ը գնահատել է ԱԲի արած 128 փաստական պնդումները, 127-ը գնահատվել է ճիշտ, մեկը՝ misleading, ոչ մեկը՝ սխալ:

Տարածված վերնագիրը սա էր. մարդկանց հնարավոր է փաստերով դուրս բերել «rabbit խոռոչ»-ից: Այսինքն՝ դավադրություն believer-ները ապացույցից դուրս չեն, պարզապես պետք է նրանց կոնկրետ բերած փաստարկներին պատասխանել բավականաչափ անհատականացված ձևով: Սա իսկապես հետաքրքիր պնդում է, իսկ հետազոտության դիզայնը persuasion-ի մասին բազմաթիվ աշխատանքներից ավելի զգույշ էր:

Հետո՝ 2026 թվականի հունիսին, *Science*-ը հոդվածին կցեց **խմբագրական Expression of Concern**: Վերապատմումների մեծ մասը հենց այս մասն է հեշտությամբ բաց թողնում, մինչդեռ այսօր արդյունքին որքան կշիռ տալու հարցը կախված է հենց դրանից:

Ի՞նչ է Expression of Concern-ը — և ինչ չէ

Խմբագրական Expression of Concern-ը ամսագրի պաշտոնական զգուշացումն է, որը ընթերցողին տեղեկացնում է, որ հոդվածի վերաբերյալ հարցեր են առաջացել, մինչ այդ հարցերը դեռ քննվում են: Այն **հետևանք չէ՝** հոդվածը հետ չի կանչվել, և ինքնին գիտական խախտման եզրակացություն

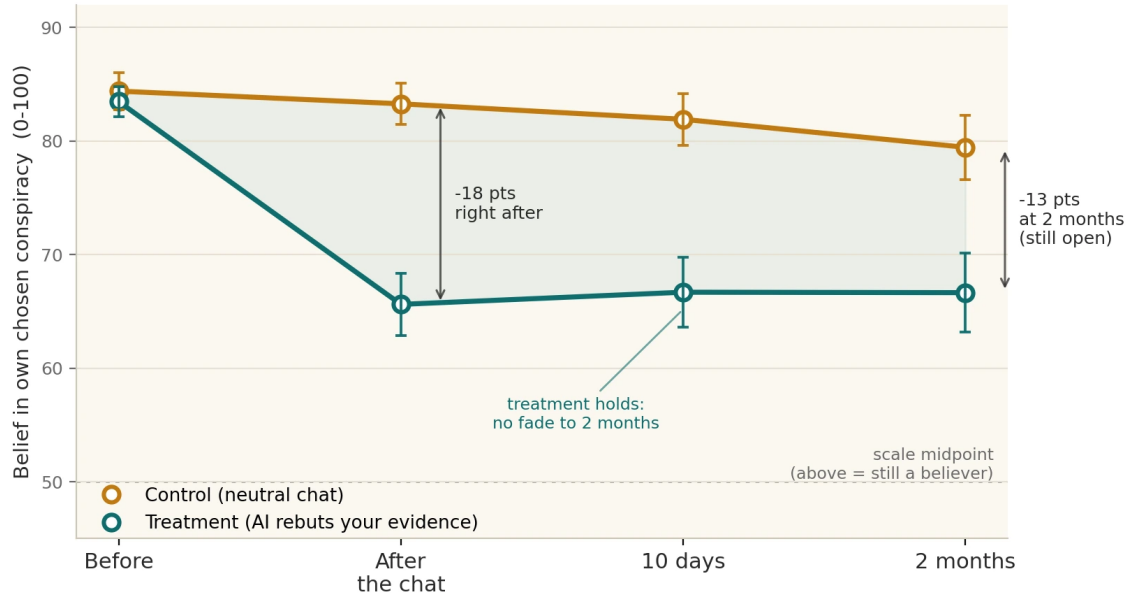
չէ: Այն նշանակում է. կարդացեք այս վերապահումը հաշվի առնելով, որովհետև գիտական գրառումը դեռ կարող է փոխվել: Այստեղ խնդիրների մասին տեղեկանալուց հետո հեղինակները հետաքննել են, մանրամասները փոխանցել *Science*-ին և ներկայացրել ուղղված վերլուծություն:

Flag արված խնդիրները կոնկրետ են: Հեղինակներին տեղեկացրել են, որ մասնակից-screening criteria-ն ձեռագիրում և հրապարակված վերլուծություն կողում նույն կերպ չի կիրառվել, և որ հանրային տվյալների հավաքածուում կող-merging սխալի պատճառով հայտնվել են ավելորդ շարքեր: Դրա պատճառով հրապարակված նյութերից որոշ թվեր վերաբրտադրելը դժվար էր: Հեղինակները *Science*-ին ներկայացրել են ուղղված վերլուծություն pipeline և updated արդյունքներ, որոնք, իրենց խոսքով, բնօրինակին համապատասխան են **ուղղությամբ, վիճակագրական նշանակալիություն և substantive չափով**: *Science*-ը դեռ գնահատում է դրանք: Քանի դեռ այդ գնահատումը չի ավարտվել, ճշգրիտ թվերը պետք է դիտել որպես նախնական:

Այսպիսով սա միաժամանակ երկու պատմություն է. հետաքրքիր արդյունք այն մասին, թե փաստերը կարո՞ղ են շարժել ամուր համոզմունքները, և կենդանի օրինակ այն մասին, թե գիտական գրանցումը ինչպես է բաց ձևով ինքն իրեն ուղղում: Այս հոդվածի նպատակն այդ երկուսը իրարից չխառնելն է:

Belief in a chosen conspiracy, before and after an AI conversation

Study 1: mean belief among the 763 participants who believed their conspiracy, by condition



Reconstructed by The Clean Paper from the public OSF dataset (Costello, Pennycook & Rand, *Science* 385, eadq1814, 2024). Dataset under a June 2026 Editorial Expression of Concern; values are provisional, not exact published figures. Markers = group means; whiskers = 95% CI; shaded band = treatment-control gap.

Ուսումնասիրության մեջ երեք փուլով ԱԲ զրույցը, որը հակադարձում էր մասնակցի սեփական բերած «ապացույցներին», հաղորդվում էր որպես մոտ 20% միջին նվազում նրա ընտրած դավադրությունի նկատմամբ հավատում: Անկապ թեմայով վերահսկիչ զրույցի համեմատ այդ նվազումը դեռ չափելի էր 10 օր և 2 ամիս անց: Ազդեցությունը կայուն էր,

բայց մասնակի. մասնակիցների մեծ մասը հետաքայում էլ շարունակում էր հավատալ տեսությանը՝ պարզապես ավելի քիչ: Հոդվածը ներկայում *Science*-ի խմբագրական Expression of Concern-ի տակ է (2026 հունիս)՝ հաշվետվության անհամապատասխանությունների և հանրային տվյալների հավաքածուի սխալի պատճառով: Հեղինակները ասում են, որ ուղղված վերլուծությունը պահպանում է ազդեցության ուղղությունը, նշանակալիությունը և չափը, իսկ *Science*-ը շարունակում է գնահատումը: Այս chart-ը The Clean Paper-ը վերակառուցել է հանրային տվյալների հավաքածուից, ուստի թվերը նախնական են, ոչ հոդվածի վերջնական նկարները:

Original figure — The Clean Paper CC BY 4.0

Ինչ արեցին հեղինակները

- Անցկացրին երկու փորձ՝ 2,190 մասնակցով: Յուրաքանչյուր մարդ իր բառերով անվանեց այն դավադրություն տեսությունն, որին հավատում էր, և այն ապացույցները, որով կարծում էր, թե այն հիմնավորվում է:
- Յուրաքանչյուր մասնակից երեք փուլով խոսեց GPT-4 Turbo-ի հետ: **Treatment** պայմանում ԱԲ-ին հրահանգ էր տրված հակադարձելու հենց մասնակցի կոնկրետ ապացույցներին, իսկ **վերահսկիչ** պայմանում քննարկվում էր անկապ թեմա:
- Ընտրված դավադրությունի նկատմամբ հավատը չափվեց գրույցից առաջ և անմիջապես հետո, ապա մասնակիցներին նորից դիմեցին **10 օր և 2 ամիս** անց:
- Պրոֆեսիոնալ fact-checker-ը գնահատեց նմուշում ԱԲ-ի արած բոլոր 128 փաստական պնդումների ճշգրտությունը:
- Ստուգեցին՝ ազդեցությունը տարածվո՞ւմ է այլ, անկապ դավադրությունների վրա, և specificity թեստի համար՝ արդյոք այն նվազեցնո՞ւմ է նաև այն դավադրությունների նկատմամբ հավատը, որոնք պատահում է իրականում **ճիշտ լինեն**:

Ինչ գտան

- **Մոտ 20% միջին նվազում** բուժում խուճրում ընտրված դավադրությունի նկատմամբ հավատում՝ վերահսկիչ խմբի համեմատ (ուսումնասիրություն 1՝ 95% վստահության միջակայք՝ [13.8, 19.7] կետ 100-կետանոց սանդղակի վրա, $P < 0.001$):
- **Արդյունքը պահպանվեց**: Treatment խուճրում նվազումը չանհետացավ. հավատը 10 օր և երկու ամիս անց մնաց մոտավորապես անմիջապես գրույցից հետո եղած մակարդակին՝ չվերադառնալով սկզբնականին, մինչ վերահսկիչ խուճրը շարունակեց ավելի բարձր մնալ: Հոդվածը ազդեցությունը նկարագրում է որպես ընտրված դավադրությունի նկատմամբ առնվազն երկու ամիս առանց նկատելի թուլացման պահպանված փոփոխություն, ոչ կարճատև տատանում:
- **Արդյունքը generalize էր անում** մարդկանց նշած տարբեր դավադրությունների միջև և պահպանվում էր նաև սկզբում շատ ուժեղ հավատ ունեցող մասնակիցների մոտ:

- **Այս միջավայրում ԱԲի փաստական պնդումները շատ ճշգրիտ էին՝** 128-ից 127-ը (99.2%) ճիշտ, 1-ը (0.8%) misleading, ոչ մեկը սխալ:
- **Ազդեցությունը կոնկրետ էր,** ոչ ընդհանուր կասկածամտություն: Intervention-ը **չէր** նվազեցնում իրական դավադրությունների նկատմամբ հավատը, ինչը հուշում է, որ այն շարժում էր unsupported belief-երը, ոչ մարդկանց դարձնում ամեն ինչից ցինիկ:
- **Կար նաև modest spillover.** այլ, անկապ դավադրությունների նկատմամբ հավատը նվազեց մոտ 12%-ով, և մասնակիցներն ավելի մեծ պատրաստակամություն էին հայտնում հակադարձելու դավադրություն պնդումներին: Հիմնական ազդեցությունը պահպանվեց ուսումնասիրություն 2-ում և նույն ուղղությունն էր ցույց տալիս նաև առանց վերահսկիչ խումբի ավելի փոքր նմուշը՝ ավելի թույլ ապացույցներ, բայց համահունչ:

Ինչ սա չի ապացուցում

- Այս պահին արդյունքը **առանց հարցականների չի կանգնած:** Հոդվածը խմբագրական Expression of Concern-ի տակ է տվյալների մշակման և վերարտադրելիություն խնդիրների պատճառով, իսկ ուղղված թվերը դեռ *Science*-ի գնահատման փուլում են: Մինչ այդ գրախոսության ավարտը կոնկրետ թվերը նախնական են:
- Սա **չի** ցույց տալիս, որ ԱԲը «բուժում» է դավադրություն belief-ը: Մոտ 20% միջին նվազումը իմաստալից փոփոխություն է, ոչ վերացում: Մասնակիցների մեծ մասը հետո էլ հավատում էր տեսությանը՝ պարզապես պակաս:
- Սա **իրական աշխարհի կիրառում արդյունք չէ:** Մասնակիցները կամավոր մտել էին ուսումնասիրություն և բարեխղճորեն շփվում էին ԱԲի հետ: Իրական կյանքում դավադրությունին ամենաուժեղ կապված մարդիկ կարող են ամենաքիչը ցանկանալ դիմել chatbot-ի, որը հակադարձելու է իրենց:
- 99.2% ճշգրտությունը **կոնկրետ է այս առաջադրանքի, մոդելի և զգույշ prompting-ի համար:** Սա ընդհանուր երաշխիք չէ, որ լեզու մոդելները ճշմարտություն են ասում: Հեղինակները նույնիսկ շեշտում են, որ նույն personalized persuasive վիճակագրական հզորությունը կարող էր *սխալ* belief-եր ուժեղացնել, եթե այդ ուղղությամբ կիրառվեր:
- «կայուն» այստեղ նշանակում է **երկու ամիս:** Դա տպավորիչ է մեկ conversation-ի համար, բայց մշտական արդյունք չէ:

Որքա՞ն ուժեղ է ապացույցը

- **Դիզայնը իրական ուժեղ կողմ է:** Treatment-versus-վերահսկիչ վերահսկվող experiments, մաքուր պլացեբո-like վերահսկիչ, կրկնություն երկու ուսումնասիրությունում և անկախ նմուշում, durability հետագա դիտարկումներ և ԱԲի սեփական փաստական պնդումների explicit ճշգրտություն check. սա persuasion հետազոտությունի մեծ մասից զգույշ է, և ազդեցության ուղղությունը նույնն է այն բոլոր տեղերում,

որտեղ ստուգվել է:

- **Բայց Expression of Concern-ը ծանոթագրություն չէ:** Հրապարակված տվյալներից ու կողից հաղորդված թիվները նորից ստանալ կարողանալը արդյունքի վստահելիության մաս է, և հենց դա է այստեղ խախտվել՝ գտման չափանիշների անհամապատասխանությունների ու կողի միավորման սխալից առաջացած կրկնվող տողի-երի պատճառով: Հեղինակների պնդումը, որ ուղղված վերլուծական շղթան պահպանում է արդյունքը, հուսադրող ու հավանական է, բայց դեռ հեղինակների սեփական հաշվետվությունն է, ոչ անկախ verdict: Սպասելու բանն այստեղ *Science*-ի գնահատումն է:
- **Ազնիվ կարգավիճակը «flagged» է, ոչ «debunked» կամ «հաստատված»:** Լավ կառուցված ու հետաքրքիր ուսումնասիրություն, որի ճշգրիտ թվերը պաշտոնական գրախոսության տակ են: Սա գուցե անհարմար վայր է լավ պատմությունը թողնելու համար, բայց հենց հիմա ամենաճշգրիտ ձևակերպումն է:

Ինչու է սա կարևոր

Տարիներ շարունակ ազդեցիկ տեսակետ կար, որ դավադրություն believer-ներին շարժում են psychological need-երն ու identity-ն, ուստի փաստերով նրանց համոզմունքները փոխել հնարավոր չէ: Եթե այս արդյունքը պահպանվի, ապացույցներ-based և կայուն նվազումը կարևոր հակակշիռ կլինի այդ pessimism-ին: Այն հուշում է, որ խնդրի մի մասը կարող էր լինել այն, որ generic debunking-ը երբեք չէր զբաղվում *այն կոնկրետ ապացույցներով*, որը տվյալ մարդը իրականում է մեջբերում, մինչ LLM-ը կարող է դա անել մասշտաբով:

Արդյունքը կարևոր է նաև որպես օրինակ՝ գիտությունը ինչպես է *պետք* աշխատի: Expression of Concern-ի դասը այն չէ, որ հայտնի արդյունքները անարժեք են: Դասն այն է, որ սխալները դուրս են գալիս, տվյալները նորից է ուսումնասիրվում, պնդումները հանրային ձևով կրկին են ստուգվում: Այդ մեխանիզմի աշխատելը համակարգի առավելություն է, ոչ սկանդալ, և հենց դրա պատճառով այս արդյունքի նկատմամբ ճիշտ դիրքը համբերությունն է՝ ոչ անմիջական ծափահարություն, ոչ անմիջական մերժում:

Եվ ԱԲԻ համար պատմությունը երկսայր է: Նույն կարողությունը, որը անհատականացված փաստարկով կարող է թուլացնել սխալ belief-ը, նույնքան կարող է ուժեղացնել այն: Technique-ը չեզոք է, նպատակը՝ ոչ:

Կարճ ամփոփում

2024 թվականի ուսումնասիրությունը գտել էր, որ GPT-4 Turbo-ի հետ երեք փուլով զրույցը մարդկանց մոտ մոտ 20%-ով նվազեցնում է իրենց հավատացած դավադրություն

տեսությունի նկատմամբ համոզվածությունը: Ազդեցությունը պահպանվել էր երկու ամիս, երևացել էր տարբեր դավադրությունների տարբեր տեսակների դեպքում, իսկ ԱԲի փաստական պնդումները հիմնականում գնահատվել էին ճիշտ: 2026 թվականի հունիսին հոդվածը տվյալների մշակման և վերաբաշխման խնդիրների պատճառով հայտնվեց խմբագրական Expression of Concernի տակ: Հեղինակները հայտնում են, որ ուղղված վերլուծությունը պահպանում է արդյունքի ուղղությունը, նշանակալիությունը և չափը, իսկ *Science*-ը դեռ գնահատում է այն: Ճիշտ է այն կարդալ որպես իսկապես հետաքրքիր, զգույշ կառուցված արդյունք, որը ներկայում գրախոսության տակ է՝ ոչ հաստատված վերջնական փաստ, ոչ հերքված պատմություն: Ամենաազնիվ եզրակացությունը հենց գիտական գործընթացն է թելադրում. նորից ստուգել:

Աղբյուրներ

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>