PERSPECTIVES

PSYCHOLOGY

Generative AI as a tool for truth

Conversation with a trained chatbot can reduce conspiratorial beliefs

By **Bence Bago**[1] *and* **Jean-François Bonnefon**[2]

Generative artificial intelligence (AI) has received broad criticism for its role in spreading misinformation (*1*–*5*). In its 2024 Global Risks Report, the World Economic Forum ranked AI-amplified misinformation as one of the most severe risks that the world currently faces (*6*). In this context, evidence for the potential positive impacts of AI is particularly welcome. On page 1183 of this issue, Costello *et al.* (*7*) report such evidence. The authors recruited more than 2000 conspiracy believers and showed that a brief but personalized conversation with an AI-driven chatbot could durably reduce research subjects' misinformed beliefs by 20% on average. Notably, this effect persisted for at least 2 months after the intervention and was observed across a wide range of conspiracy theories. The results challenge conventional wisdom about conspiratorial beliefs and demonstrate that it is possible to counter even deeply entrenched views with sufficiently compelling evidence.

The size, robustness, and persistence of the reduction in conspiracy beliefs reported by Costello *et al.* suggest that a scalable intervention to recalibrate misinformed beliefs may be within reach. The findings also raise questions about the range of potential applications that may be amenable to this approach. Popular psychological theories posit that people adopt conspiracy beliefs to fulfill underlying psychological needs, which renders the believers impervious to counterevidence. In addition, entrenched conspiracy theorists are often quite knowledgeable about their chosen conspiracy, which makes it difficult for nonbelievers to flexibly marshal sufficient facts and arguments to counter them. As described by Costello *et al.*, AI programs known as large language models offer a promising solution to this challenge because these models can draw from an extensive body of information across diverse topics and have the ability to tailor counterarguments to specific conspiracies and lines of argument.

A corollary of this approach, however, is that the AI dialogue technique used by Costello *et al.* may only work for conspiratorial beliefs resulting from thorough elaboration and may be less effective on more superficial misinformed beliefs with little justification. Ideally, a scalable AI dialogue intervention should have a broad range of applications and be able to help recalibrate misinformed beliefs in domains as varied as pseudoscience, health myths, climate skepticism, or political extremism. However, it is unclear whether these types of misinformed beliefs resist correction for the same reasons as conspiracy beliefs and thus can be rebutted in the same way. As a heuristic, one might expect that any misinformed beliefs that respond well to counterevidence, such as climate beliefs (*8*), may react even better to the flexible, thorough counterevidence provided by generative AI. Finally, more research is needed to assess how feasible it is for generative AI to quickly respond to emerging conspiracy theories—for which no specific training data may be available—at times when speed is crucial, such as during the early days of a pandemic or after an assassination attempt on an elected official.

> **"...a scalable intervention to recalibrate misinformed beliefs may be within reach."**

Another avenue for future work involves understanding how long and frequent AI dialogues should be to be effective. Costello *et al.* reported a 16-point drop in conspiracy beliefs on a 100-point scale after only three rounds of back-and-forth conversation between research subjects and a trained large language model. This effect size is substantial, especially when compared with other interventions encouraging reflective, analytical thinking, which yielded only a one- to six-point drop in reducing conspiracy beliefs (*9*, *10*).

The persistence of the effect reported by Costello *et al.* over time is also noteworthy, albeit insufficient to completely eliminate misinformed conspiracy beliefs. More data are needed to determine whether stronger doses (e.g., longer conversations) and repeated interventions might yield even stronger results. By systematically varying the duration and frequency of AI-driven dialogues, there can be better understanding of how to maximize their impact and move closer to a comprehensive solution for mitigating misinformed beliefs.

The AI dialogue technique is so powerful because it automates the generation of specific and thorough counterevidence to the intricate arguments of conspiracy believers and therefore could be deployed to provide accurate, corrective information at scale. An important limitation to realizing this potential lies in delivery—namely, how to get individuals with entrenched conspiracy beliefs to engage with a properly trained AI program to begin with. The Costello *et al.* study relied on online survey respondents who chose to participate in their scientific study. However, conspiracy believers are likely to distrust scientific institutions (*11*), and it may be difficult to persuade them to opt into AI dialogues designed to challenge their beliefs (*12*) in everyday settings. Overcoming or bypassing this resistance will require innovative delivery strategies.

One option described by Costello *et al.* is to match internet search terms related to conspiracy theories with AI-generated summaries of accurate information. There is certainly value in intercepting misinformation at the point of search and providing users with immediate, relevant counterevidence. However, short summaries based on generic search terms would forego the two main strengths of the AI dialogue technique: the length and specificity of the AI responses. Another option put forward by the authors involves AI-powered social media accounts that could reply to users who share inaccurate conspiracy-related content. This approach

[1]Department of Social Psychology, Tilburg University, Tilburg, Netherlands. [2]Toulouse School of Economics, CNRS (TSM-R), Toulouse, France. Email: b.bago@tilburguniversity.edu; jean-francois.bonnefon@tse-fr.eu

would allow for specificity, and perhaps for length on some platforms, but may still be insufficient because users might deem these responses intrusive or untrustworthy and therefore might ignore or block the source AI accounts.

A more personal approach to introducing AI may also be effective. Conspiracy believers often have friends or relatives who are desperate for a way to debunk misinformed beliefs. These connections could be leveraged by encouraging these friends and relatives to coax the believers into engaging in AI dialogue. Friends and relatives themselves could also use AI for inspiration when debating with their conspiracy-believing contacts. Such inspiration could come both from the facts provided by AI and from the dispassionate way that AI provides them. In both scenarios, ensuring that only properly trained AI programs are used is essential to maintaining the integrity and effectiveness of the intervention. Indeed, absent this training, it is possible that AI programs could also convince people to adopt dubious beliefs.

For better or worse, AI is set to profoundly change our culture (*13*, *14*). Although widely criticized as a force multiplier for misinformation, the study by Costello *et al.* demonstrates a potential positive application of generative AI's persuasive power. The findings also underscore the ongoing importance of thorough follow-up research and appropriate guardrails to ensure that this transformative technology is deployed responsibly. ■

REFERENCES AND NOTES

1. V. Capraro et al., *Proc. Natl. Acad. Sci. U.S.A. Nexus* **3**, pgae191 (2024).
2. G. Spitale, N. Biller-Andorno, F. Germani, *Sci. Adv.* **9**, eadh1850 (2023).
3. C. Kidd, A. Birhane, *Science* **380**, 1222 (2023).
4. S. Y. Shin, J. Lee, *Digit. Journal.* **10**, 412 (2022).
5. A. Simchon, M. Edwards, S. Lewandowsky, *Proc. Natl. Acad. Sci. U.S.A. Nexus* **3**, pgae035 (2024).
6. "Global Risks Report 2024" (World Economic Forum, 2024); https://www.weforum.org/reports/global-risks-report-2024.
7. T. H. Costello, G. Pennycook, D. G. Rand, *Science* **385**, eadq1814 (2024).
8. B. Bago, D. G. Rand, G. Pennycook, *Proc. Natl. Acad. Sci. U.S.A. Nexus* **2**, pgad100 (2023).
9. V. Swami, M. Voracek, S. Stieger, U. S. Tran, A. Furnham, *Cognition* **133**, 572 (2014).
10. B. Bago, D. G. Rand, G. Pennycook, *J. Exp. Soc. Psychol.* **103**, 104395 (2022).
11. B. T. Rutjens, B. Većkalov, *Curr. Opin. Psychol.* **46**, 101392 (2022).
12. J.-F. Bonnefon, A. Shariff, I. Rahwan, *Ethics of Artificial Intelligence*, S. M. Liao, Ed. (Oxford Univ. Press, 2020).
13. L. Brinkmann *et al.*, *Nat. Hum. Behav.* **7**, 1855 (2023).
14. J.-F. Bonnefon, I. Rahwan, A. Shariff, *Annu. Rev. Psychol.* **75**, 653 (2024).

ACKNOWLEDGMENTS

J.-F.B. acknowledges support from grant ANR-19-PI3A-0004, grant ANR-17-EURE- 0010, and the Toulouse School of Economics (TSE)–Partnerships Foundation.

10.1126/science.ads0433

GRAPHIC: N. BURGESS/*SCIENCE*

CHEMISTRY

A crystallized view of acid-base chemistry

The structural relationship between Lewis adduct isomers is resolved

By **Andrew R. Jupp**

Interactions between electron pair acceptors (Lewis acids) and electron pair donors (Lewis bases) to create bimolecular Lewis adducts are fundamental to understanding and designing a multitude of chemical reactions. The eponymous acids and bases were codified by Gilbert Lewis in 1923 (*1*), and in 1952, Robert Mulliken posited that Lewis acids and bases associate through "inner" and "outer" forms (*2*). Yet, the molecular structures of these two forms of the Lewis adduct have been elusive. On page 1184 of this issue, Liu and Gabbaï (*3*) report the crystal structures of the inner and outer forms of a Lewis adduct. This work provides a long-awaited opportunity to compare and study these fundamental compounds.

The inner form of a Lewis adduct is the "classical" version that features a covalent bond between the donor and acceptor of the electron pair. For a conventional Lewis acid such as trivalent boranes (BR_3; R is a substituent atom or group), this association is accompanied by a change in geometry around the boron center, from trigonal planar to tetrahedral. This distortion usually comes with an energetic penalty, but this is compensated by the formation of the dative bond. There are some acid-base systems where this is not the case. For example, whereas BCl_3 is typically a stronger Lewis acid than BF_3 toward strong Lewis bases, the opposite is true for very weak Lewis bases such as carbon monoxide (CO). This is because in the OC–BF_3 and OC–BCl_3 adducts, the C–B bonds are very weak, with long C–B bond distances and minimal distortion of the Lewis acids (*4*). This weak complex is the outer form and is usually observed when the strength of the bond being formed does not outweigh the distortion energy. For systems where the inner form is favored, it is postulated that the inner adduct is accessed by passing through a higher-energy outer adduct.

Liu and Gabbaï devised a system that balances the energies of the inner and outer forms of a Lewis adduct, rendering both accessible to crystallographic characterization. Instead of the traditional boron Lewis acid, an isoelectronic and isolobal carbenium ($^{+}CR_3$) center was used, and a phosphine oxide (R_3PO) was used as the Lewis base. These were tethered onto an acenaphthene ($C_{12}H_{10}$) framework to ensure proximity in space. The outer isomer was characterized with a C–O distance of 2.653(3) Å between the acid and base centers, whereas the inner form showed a closer contact of 1.534(4) Å. The two isomers were examined with a wide range of techniques, including infrared spectroscopy, nuclear magnetic resonance spectroscopy, and computational modeling. Interestingly, the modeling showed that the outer isomer was energetically favored (see the figure), but only by 5.4 kJ/mol. The authors also demonstrated that the adduct could function as a photoredox catalyst for a small

School of Chemistry, University of Birmingham, Edgbaston, UK. Email: a.jupp@bham.ac.uk

Lewis acid and base adducts

A reaction profile for the formation of the outer and inner adducts of a Lewis acid and base is shown in which the outer version is more stable.