RESEARCH ARTICLE SUMMARY

ARTIFICIAL INTELLIGENCE

Durably reducing conspiracy beliefs through dialogues with AI

Thomas H. Costello*, Gordon Pennycook, David G. Rand

INTRODUCTION: Widespread belief in unsubstantiated conspiracy theories is a major source of public concern and a focus of scholarly research. Despite often being quite implausible, many such conspiracies are widely believed. Prominent psychological theories propose that many people want to adopt conspiracy theories (to satisfy underlying psychic "needs" or motivations), and thus, believers cannot be convinced to abandon these unfounded and implausible beliefs using facts and counterevidence. Here, we question this conventional wisdom and ask whether it may be possible to talk people out of the conspiratorial "rabbit hole" with sufficiently compelling evidence.

RATIONALE: We hypothesized that interventions based on factual, corrective information may seem ineffective simply because they lack sufficient depth and personalization. To test this hypothesis, we leveraged advancements in large language models (LLMs), a form of artificial intelligence (AI) that has access to vast amounts of information and the ability to generate bespoke arguments. LLMs can thereby directly refute particular evidence each individual cites as supporting their conspiratorial beliefs.

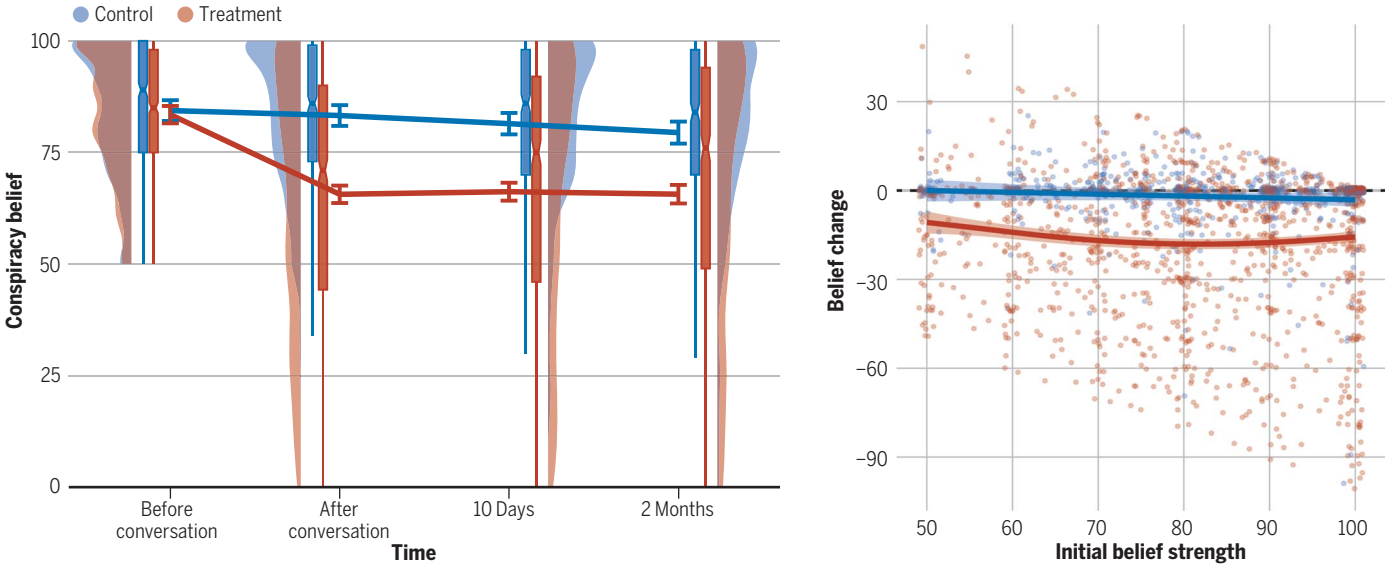
To do so, we developed a pipeline for conducting behavioral science research using real-time, personalized interactions between research subjects and AI. Across two experiments, 2190 Americans articulated—in their own words—a conspiracy theory in which they believe, along with the evidence they think supports this theory. They then engaged in a three-round conversation with the LLM GPT-4 Turbo, which we prompted to respond to this specific evidence while trying to reduce participants' belief in the conspiracy theory (or, as a control condition, to converse with the AI about an unrelated topic).

RESULTS: The treatment reduced participants' belief in their chosen conspiracy theory by 20% on average. This effect persisted undiminished for at least 2 months; was consistently observed across a wide range of conspiracy theories, from classic conspiracies involving the assassination of John F. Kennedy, aliens, and the illuminati, to those pertaining to topical events such as COVID-19 and the 2020 US presidential election; and occurred even for participants who conspiracy beliefs were deeply entrenched and important to their identities. Notably, the AI did not reduce belief in true conspiracies. Furthermore, when a professional fact-checker evaluated a sample of 128 claims made by the AI, 99.2% were true, 0.8% were misleading, and none were false. The debunking also spilled over to reduce beliefs in unrelated conspiracies, indicating a general decrease in conspiratorial worldview, and increased intentions to rebut other conspiracy believers.

CONCLUSION: Many people who strongly believe in seemingly fact-resistant conspiratorial beliefs can change their minds when presented with compelling evidence. From a theoretical perspective, this paints a surprisingly optimistic picture of human reasoning: Conspiratorial rabbit holes may indeed have an exit. Psychological needs and motivations do not inherently blind conspiracists to evidence—it simply takes the right evidence to reach them. Practically, by demonstrating the persuasive power of LLMs, our findings emphasize both the potential positive impacts of generative AI when deployed responsibly and the pressing importance of minimizing opportunities for this technology to be used irresponsibly. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: tcostello@american.edu
Cite this article as T. H. Costello *et al.*, *Science* **385**, eadq1814 (2024). DOI: 10.1126/science.adq1814

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.adq1814



Dialogues with AI durably reduce conspiracy beliefs even among strong believers. (Left) Average belief in participant's chosen conspiracy theory by condition (treatment, in which the AI attempted to refute the conspiracy theory, in red; control, in which the AI discussed an irrelevant topic, in blue) and time point for study 1. (Right) Change in belief in chosen conspiracy from before to after AI conversation, by condition and participant's pretreatment belief in the conspiracy.

RESEARCH ARTICLE

ARTIFICIAL INTELLIGENCE

Durably reducing conspiracy beliefs through dialogues with AI

Thomas H. Costello[1,2]*, Gordon Pennycook[3], David G. Rand[1]

Conspiracy theory beliefs are notoriously persistent. Influential hypotheses propose that they fulfill important psychological needs, thus resisting counterevidence. Yet previous failures in correcting conspiracy beliefs may be due to counterevidence being insufficiently compelling and tailored. To evaluate this possibility, we leveraged developments in generative artificial intelligence and engaged 2190 conspiracy believers in personalized evidence-based dialogues with GPT-4 Turbo. The intervention reduced conspiracy belief by ~20%. The effect remained 2 months later, generalized across a wide range of conspiracy theories, and occurred even among participants with deeply entrenched beliefs. Although the dialogues focused on a single conspiracy, they nonetheless diminished belief in unrelated conspiracies and shifted conspiracy-related behavioral intentions. These findings suggest that many conspiracy theory believers can revise their views if presented with sufficiently compelling evidence.

Widespread belief in unsubstantiated or false conspiracy theories is both a major source of public concern and a focus of scholarly research (*1*–*3*). Conspiracy theories—in which events are understood as being caused by secret, malevolent plots involving powerful conspirators—are often quite implausible. Yet a large fraction of the world has come to believe them, including as much as 50% of the US population by past estimates (*4*–*7*). Such prevalence is particularly concerning because conspiracy belief is often used as a paradigmatic example of resistance to evidence (*8*–*10*): There is little evidence of interventions that successfully debunk conspiracies among people who already believe them (*11*, *12*).

The apparent resilience of conspiracy theories in the face of clear counterevidence poses a powerful challenge to scientific theories that emphasize the role of reasoning in belief formation and revision (*13*, *14*). Instead, belief in conspiracies has primarily been explained through social-psychological processes thought to blunt rational decision-making and receptivity to evidence (*7*, *15*–*19*). Popular explanations propose that people adopt conspiracy theories to sate underlying psychic "needs" or motivations, such as the desire for control over one's environment and experiences (*15*), certainty and predictability (*20*), security and stability (*21*), and uniqueness (*22*). If these psychological needs are met by believing in conspiracy theories, the beliefs become more than just opinions; they become mechanisms for psychological equilibrium, and thus are argued to be highly resistant to counterevidence (*1*, *3*, *23*). Coupled with peoples' motivations to maintain their identity and/or group memberships, with which conspiracies also interface (*24*–*26*), believers may use specific forms of biased information processing (motivated reasoning) where counterevidence is selectively ignored (*27*–*29*).

These perspectives, which center the psychological drives of those who believe conspiracies, paint a grim picture for countering conspiratorial beliefs: Because conspiracy believers at some level "want" to believe, convincing them to abandon unfounded beliefs using facts should be virtually impossible (without more fundamentally altering their underlying psychology and identity commitments).

Here, we question this conventional wisdom about conspiracy theories and ask whether it may, in fact, be possible to talk people out of the conspiratorial "rabbit hole" with sufficiently compelling evidence. Leveraging recent advancements in large language models (LLMs), we shed light on whether counterevidence reduces belief in conspiracy theories. We hypothesize that fact-based interventions may appear to fall short because of a lack of depth and personalization of the corrective information. Entrenched conspiracy theorists are often quite knowledgeable about their conspiracy of interest, deploying prodigious (albeit often erroneous or misinterpreted) lists of evidence in support of the conspiracy that can leave skeptics outmatched in debates and arguments (*30*, *31*). Furthermore, people believe a wide range of conspiracies, and the specific evidence brought to bear in support of even a particular conspiracy theory may differ substantially from believer to believer. Canned debunking attempts that argue broadly against a given conspiracy theory may, therefore, be ineffective because they fail to address the specific evidence accepted by the believer—and thus fail to be convincing.

In contrast, we hypothesize that LLMs offer a promising solution to these challenges because they have two key capabilities: (i) access to a vast amount of information across diverse topics and (ii) the ability to tailor counterarguments to specific conspiracies, reasoning, and evidence the believer brings to bear (*32*). These capabilities allow LLMs to respond directly to—and refute—the particular evidence supporting an individual's conspiratorial beliefs. In so doing, LLMs can potentially overcome the heterogeneity in conspiracy beliefs and supporting evidence that we hypothesize have stymied previous debunking efforts.

To test whether LLMs can effectively refute conspiracy beliefs—or whether psychological needs and motivations render conspiracy believers impervious to counterevidence—we developed a pipeline for conducting behavioral science research using real-time, personalized interactions between research subjects and LLMs. In our experiments, participants articulated a conspiracy theory in which they believe—in their own words—along with the evidence they think supports the theory. They then engaged in a back-and-forth interaction with an artificial intelligence (AI) implemented using the LLM GPT-4 Turbo (*33*). In line with our theorizing around the distinctive capacities of LLMs for debunking conspiracies, we prompted the AI to use its store of knowledge to respond to the specific evidence raised by the participant and reduce the participant's belief in the conspiracy theory (or, in a control condition, participants conversed with AI about an unrelated topic). The AI was specifically instructed to "very effectively persuade" users against belief in their chosen conspiracy, allowing it to flexibly adapt its strategy to the participant's specific arguments and evidence. To further enhance this tailored approach, we provided the AI with each participant's written conspiracy rationale as the conversation's opening message, along with the participant's initial rating of their belief in the conspiracy. This design choice directed the AI's attention to refuting specific claims, while simulating a more natural dialogue wherein the participant had already articulated their viewpoint. For the full prompts given to the model, see table S2. The conversation lasted 8.4 min on average and comprised three rounds of back-and-forth interaction (not counting the initial elicitation of reasons for belief from the participant), a length chosen to balance the need for substantive dialogue with pragmatic concerns around study length and participant engagement.

This design allowed us to test whether tailored persuasive communication is indeed able to reduce already-held conspiracy beliefs; how

[1]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Department of Psychology, American University, Washington, DC, USA. [3]Department of Psychology, Cornell University, Ithaca, NY, USA.
*Corresponding author. Email: tcostello@american.edu

the effectiveness of such communication varies on the basis of factors such as the intensity of the preexisting belief, the participant's subscription to a more general conspiratorial mindset, the importance of the conspiracy to the participant's life, and the content of the specific conspiracy theory articulated by the participants; and whether any such persuasion spills over into other related beliefs and behaviors. Finally, our design produced rich textual data from thousands of conversations between the AI and the human participants (see our web-based Conversation Browser that displays verbatim interactions sorted by topic and effect size: https://8cz637-thc.shinyapps.io/ConspiracyDebunkingConversations), which we analyzed to gain insight into what the participants believe and how the LLM engages in persuasion.

Can conspiracy beliefs be refuted?

In study 1, participants rated their belief in 15 popular conspiracy theories [taken from the Belief in Conspiracy Theories Inventory (BCTI)], completed a distractor task, and were then asked to identify and describe a particular conspiracy theory they believed in (not necessarily one of the 15 rated earlier) as well as providing details about evidence or experiences supporting their belief. In real time, the AI created a summary statement of each participant's free-text conspiratorial belief description, and each participant was then asked to indicate their belief in the AI summary of their conspiracy statement—providing a pretreatment measure of belief. This open-ended measurement approach avoids a long-standing criticism of discrete conspiracism measures, such as the BCTI, for failing to representatively sample from the universe of possible conspiracies (*34*).

Out of N = 1055 American participants (quota-matched to the US census on age, gender, race, and ethnicity) who completed the pretreatment measures, 72.2% indicated belief in a conspiracy theory and were included in our subsequent analyses, whereas 20.6% said they did not believe any conspiracy theories or described a belief that the AI classified as not actually conspiratorial [for coding validation, see supplementary materials (SM), supplementary text section 1 and table S4], 3.5% described a conspiracy theory but had belief below the scale midpoint, and 3.6% described a conspiracy theory that was inaccurately summarized by the AI.

To assess whether the AI could reduce conspiracy beliefs, participants were then randomly assigned to either have a three-round conversation with the AI about their favored conspiracy belief (treatment group, 60% of the sample) or to participate in a similarly structured conversation about a neutral topic (control group, 40% of the sample). Although past work has typically found that people are less receptive to corrections (*35*), advice (*36*), and persuasion (*37*) labeled as coming from AI, we opted to avoid deception and explicitly informed participants that they were interacting with an AI.

For each participant, the AI was (i) provided with that participant's specific open-ended response, including their stated rationale for believing the conspiracy theory and their degree of endorsement, and (ii) prompted to use simple language to persuade the user that their conspiracy theory is unsubstantiated and change their beliefs to be less conspiratorial. After the conversations, all participants rerated belief in their stated conspiracy theory and the BCTI items (see Fig. 1 for key methodological steps and a sample conversation).

Was conversing with an AI able to successfully reduce participants' conspiratorial beliefs? Yes, as the treatment reduced participants' belief in their stated (i.e., focal) conspiracy by 16.8 points more than the control (linear regression with robust standard errors controlling for pretreatment belief, 95% confidence interval (CI) [13.8, 19.7], $P < 0.001$, d = 1.15; Fig. 2A and SM supplementary text section 2). This translates to a 21.43% decrease in belief among those in treatment (versus 1.04% in the control). Furthermore, more than a quarter (27.4%) of participants in the treatment became uncertain of their conspiracy belief (i.e., belief below the scale midpoint) after the conversation, compared with only 2.4% in the control. We also found a significant effect when using the pre- and posttreatment BCTI ratings, as opposed to the pre- and posttreatment evaluations of the AI summary, among the subset of participants (n = 303) who provided a focal conspiracy that strongly resembled a BCTI item (b = −12.04, 95% CI [−16.63, −7.46], $P < 0.001$, d = 0.70; see SM supplementary text section 2.1), indicating the robustness of the results to our measurement approach.

To assess the persistence of this effect, we recontacted participants 10 days and 2 months later for a short follow-up in which they once again completed the outcome measures. We found no significant change in belief in the focal conspiracy theory from immediately after the AI conversation to either 10 days or 2 months later in a mixed-effects model with fixed effects for experimental condition and time point and random intercepts for participants ($b_{\Delta\text{ImmediatelyPost–10Days}}$ = 0.63, 95% CI [−2.72, 1.46], P = 0.56; $b_{\Delta\text{ImmediatelyPost–2Months}}$ = 0.03, 95% CI [−2.24, 2.31], P = 0.98; Fig. 2A and table S9). We continue to observe a significant treatment effect under the conservative assumption that the 14% of participants who did not complete the follow-up reverted to their initial pretreatment belief levels (b = 12.70, 95% CI [9.47, 15.93], $P < 0.001$). Thus, the change in beliefs we observe is very persistent.

However, in study 1, the proportion of participants who endorsed a conspiracy in free-text response was somewhat higher than prior estimates of the American public (*4*). Given that participants in study 1 completed the BCTI before supplying their conspiracy theory, it is possible that exposure to the BCTI items increased the salience of particular conspiracy theories and thereby increased reported belief.

We explored this possibility, as well as the replicability of our results and robustness to minor design changes, in study 2, where N = 2286 Americans completed an extremely similar procedure without the BCTI. We also changed the wording for the conspiracy elicitation prompt such that, instead of directly asking participants which conspiracy theories they believed in, we provided a definition of what a conspiracy theory is and asked participants whether they found any such theories compelling. Finally, we disabled copy-and-paste functionality to guard against participants themselves using LLMs to complete the study (*38*). Here, 64.6% of participants indicated belief in a conspiracy theory (see table S3). Most importantly, we replicated the experimental results of study 1. Participants in the treatment in study 2 reduced belief in their focal conspiracy by 12.3 points more than participants in the control (95% CI [10.07, 14.72], $P < 0.001$, d = 0.79; Fig. 2B and table S8), which translates to a 19.41% average decrease in belief (versus a 2.94% decrease in the control).

Further demonstrating the robustness of our results, we also replicated our findings in a supplemental study conducted using a sample recruited through the participant supplier Lucid (b = −10.99, 95% CI [−16.09, −5.88], $P < 0.001$, d = 0.53; see SM supplementary text section 8 and fig. S13), which provides relatively inattentive respondents who mostly do nonacademic surveys (*39*). Thus, the effect is not specific to attentive and engaged participants from CloudResearch Connect.

Robustness across topics and people

Next, we examined the robustness of the AI conversation treatment effect. We began by investigating whether the treatment size varies across the specific focal conspiracy theories articulated by the participants. To do so, we used a multistep natural language processing and clustering approach to classify each focal conspiracy theory according to its contents (see SM supplementary text section 3). We found that the treatment effect did not differ significantly across conspiracy type in an omnibus test ($F_{12,1971}$ = 1.30, P = 0.21) and that the treatment significantly decreased belief across all but one of the 12 different types of conspiracy theory identified with >1% prevalence in the sample (Fig. 2C). Notably, the treatment worked even for highly salient—and likely deeply entrenched—political conspiracies such as those involving fraud in the

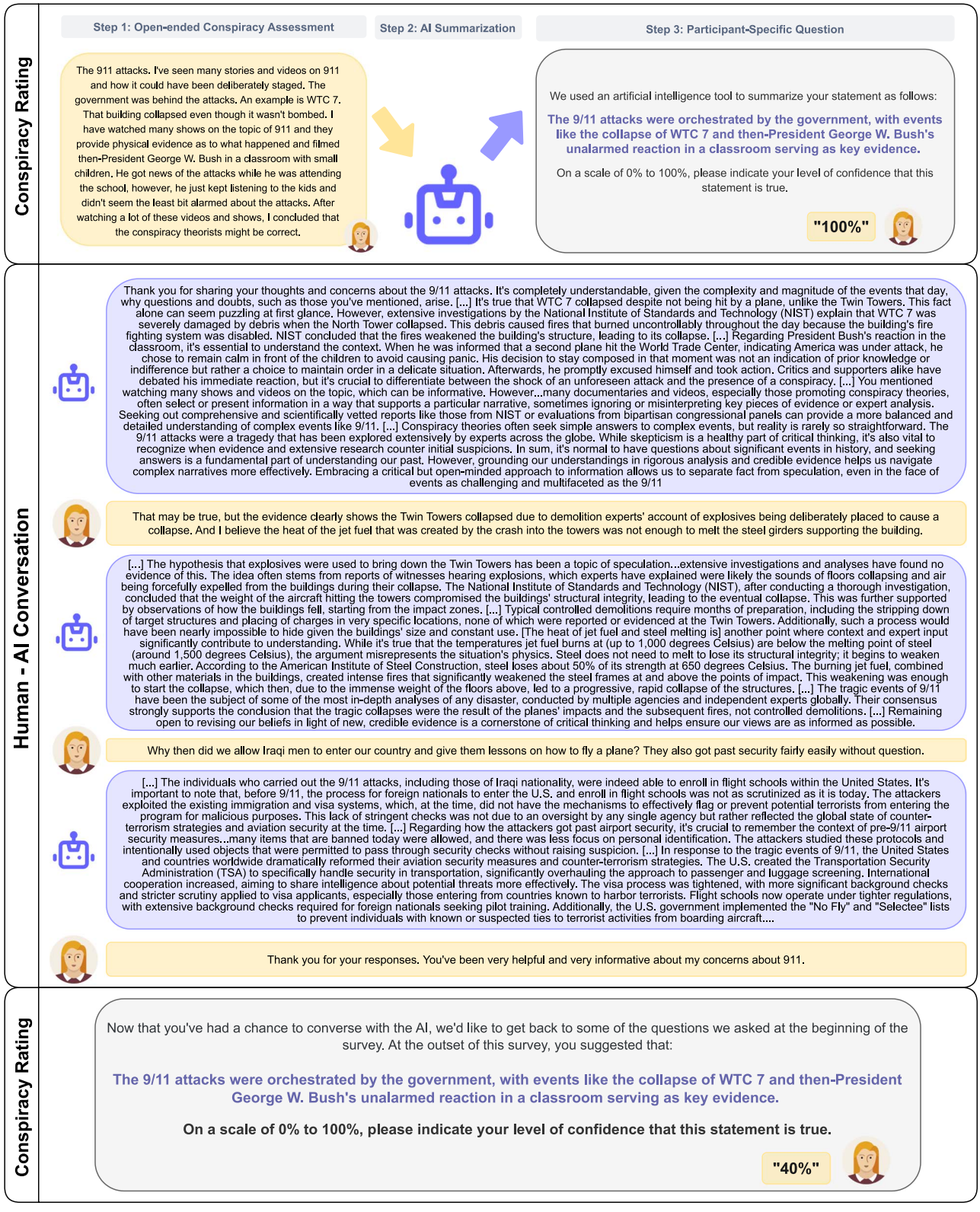


Fig. 1. Design and flow of the human-AI dialogues. Respondents (yellow) described a conspiracy theory they believed in, along with the evidence they thought supported it. Each response was fed-forward to a query instructing the AI model (GPT-4 Turbo, shown in purple) to generate a brief, relatively standardized statement of that conspiracy. Participants then rated their belief in the summary statement, yielding our pretreatment measure (0–100 scale, with 0 being "definitely false," 50 being "uncertain," and 100 being "definitely true"). All respondents then entered into a conversation with the AI model (treatment argued against the conspiracy theory's veracity, control discussed irrelevant topics). After three rounds of dialogue, respondents once again rated their belief in the summarized conspiracy statement, serving as our posttreatment measure. Shown is an example treatment dialogue that led to a substantial reduction in the participant's belief in a conspiracy.

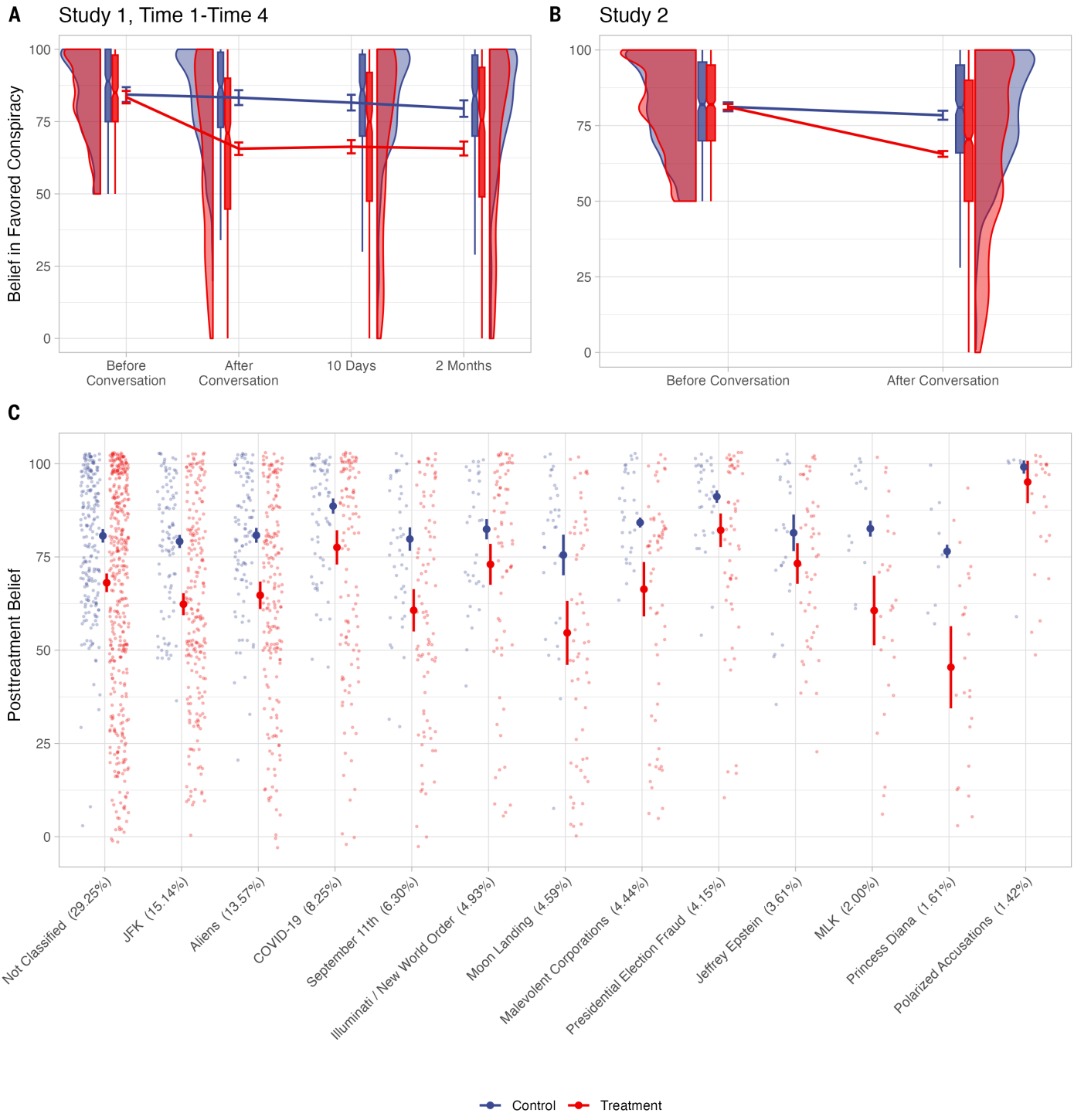


Fig. 2. A brief conversation with an AI model durably reduces belief in conspiracy theories. (**A** and **B**) Average belief in each participant's focal conspiracy theory by condition (treatment, in which the AI attempted to refute the conspiracy theory, in red; control, in which the AI discussed an irrelevant topic, in blue) and time point for study 1 (A) and study 2 (B). Before-conversation belief was greater than 50 for all participants because participants with initial belief less than 50 were excluded from the study. (**C**) Belief immediately after the AI conversation by condition and topic of the participant's focal conspiracy theory; see SM supplementary text section 3 for details on topic detection. Error bars indicate 95% confidence intervals.

2020 US presidential election (b = 10.61 [5.54, 15.67], P < 0.001, d = 0.82) and the COVID-19 pandemic (b = 11.79 [6.98, 16.60], P < 0.001, d = 0.73). In addition to allowing us to test for the robustness of our treatment, this classification based on the participants' open-ended responses also provides descriptive insight into which particular conspiracy theories Americans subscribe to. Alternative specifications of this clustering solution yielded highly similar patterns (see SM supplementary text section 3.1; figs. S6 and S7).

We then turned to variation in effect sizes across individuals. In particular, we sought to determine whether the treatment is effective

even among participants likely to have particularly entrenched beliefs. We used generalized additive models to analyze how the treatment effect varies in a nonlinear manner based on several measures relevant to entrenchment. First, we examined participants' level of pretreatment belief in the focal conspiracy and found that it does significantly moderate the treatment effect, resulting in a u-shaped curve [ΔAIC (Akaike information criterion) = −3.25, ΔR^2 (coefficient of determination) = 0.002, P = 0.022; Fig. 3A and table S14]. Next, we examined how important participants indicated that the conspiracy theory is to their worldview (Fig. 3B and table S15), which does significantly decrease the size of the treatment effect (ΔAIC = 3.12, ΔR^2 = 0.003, P = 0.025). Critically, however, the effect was significant even among those who indicated the highest level of importance (b = 5.84 [0.33, 11.35], P = 0.038, d = 0.53). Lastly, we examined participants' level of general conspiratorial ideation (i.e., the intensity with which they believed BCTI conspiracies), which showed nonsignificant moderation of the treatment effect (ΔAIC = 0.77, ΔR^2 = 0.002, P = 0.108; Fig. 3C and table S16). Participants at or above the 90th percentile of conspiratorial ideation in our sample (i.e., endorsing virtually all of the 15 diverse conspiracy statements) still displayed a substantial average treatment effect of b = 9.07 (95% CI [2.73, 15.44], P = 0.006, d = 0.53).

We also examined moderation by demographic characteristics (age, race, gender, education) and other individual difference variables (political orientation, political extremism, religiosity, familiarity with generative AI, usage of generative AI, trust in generative AI, and institutional trust). In a single linear regression model including all candidate moderators and their interaction with the experimental condition, as well as a control for conspiracy type and its interaction with the experimental condition, only (i) trust in generative AI and (ii) institutional trust consistently moderated the treatment effect, such that those higher in both kinds of trust showed larger treatment effects (see tables S17 and S18). We conducted a post hoc analysis using the causal forest method (*40*) to further clarify and identify heterogeneous effects of the intervention across all moderators (including conspiracy type, pretreatment beliefs, and importance) (see SM supplementary text section 4.3). Variable importance analyses indicated that, for experiment 1, the predominant determinants of treatment effect heterogeneity (in order) were the participant's age, trust in generative AI, and BCTI scores; in experiment 2, these were institutional trust (which was not measured in experiment 1), trust in generative AI, age, and conspiracy importance. While there were heterogeneous treatment effects across subgroups (t = 4.97, P < 0.001), the conditional average treatment effects (CATE) across covariate profile subgroups ranged from −20.54 to −6.56—implying that the treatment reduced belief for all subgroups. For example, the CATE ranged from −17.7 to −4.5 (median = −9.7) for individuals who rated their focal conspiracy belief as "extremely important" to their personal beliefs; from −13.6 to −6.7 (median = −9.8) for individuals with minimal trust in AI; and from −18.2 to −10.0 (median = −15.4) for individuals with BCTI scores in the 95th percentile and above.

Spillover effects and behavioral implications

Next, we examined treatment effects on outcomes beyond belief in the focal conspiracy. First, we asked whether the treatment affected individuals' beliefs in conspiracy theories that were not targeted by the conversation with the AI model (see SM supplementary text sections 2 and 7). We did so by analyzing respondents' belief in 15 widespread conspiracy theories from the BCTI (which was assessed by both pretreatment and posttreatment in study 1). We used a linear mixed model with fixed effects for experimental conditions and time point (pre, post, 10 days, and 2 months) and random intercepts for participant. Postintervention, there was a 3.05-point decrease in general conspiracy beliefs in the active condition (95% CI [−3.90, −2.20], P < 0.001, 8.2% decrease; Fig. 4A and table S10) compared with a 1.64-point increase in the control (d = 0.21). This effect was still evident at the 2-month follow-up, with a 2.46-point decrease from pretreatment (95% CI [−3.44, −1.49], P < 0.001). When only analyzing belief in BCTI conspiracy theories that a given participant believed pretreatment (i.e., endorsed above the scale midpoint), the impact was more pronounced: a 9.39-point reduction immediately postintervention (95%

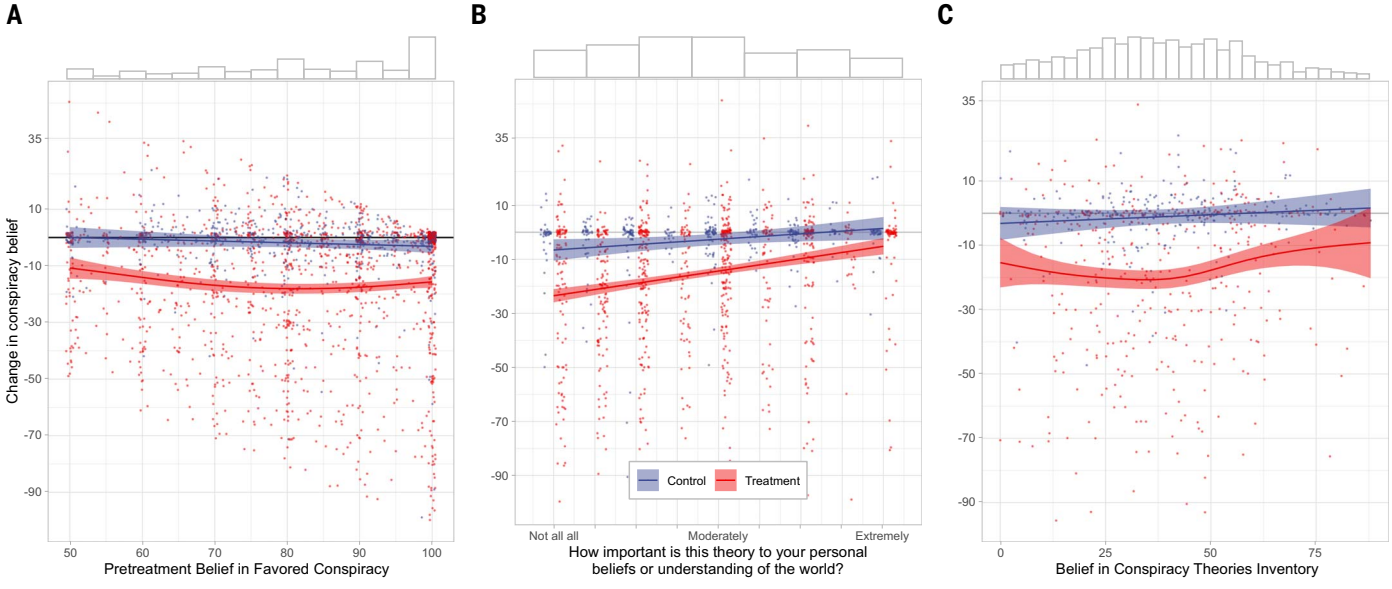


Fig. 3. The treatment is effective even for those who are strongly attached to their conspiracy beliefs. (**A** to **C**) Shown is the change in belief in the focal conspiracy from before AI conversation to after AI conversation, for the treatment (red) and control (blue) conditions. Data are pooled across studies to maximize power. Individual observations are plotted along with fit lines and 95% confidence intervals generated using generalized additive models. We conducted separate analyses for predictors of participant's pretreatment level of belief in the focal conspiracy (A), rating of how important the focal conspiracy is to their personal beliefs or understanding of the world (B), and general conspiratorial mindset as measured by average belief in 15 conspiracies from the BCTI completed before treatment (C).

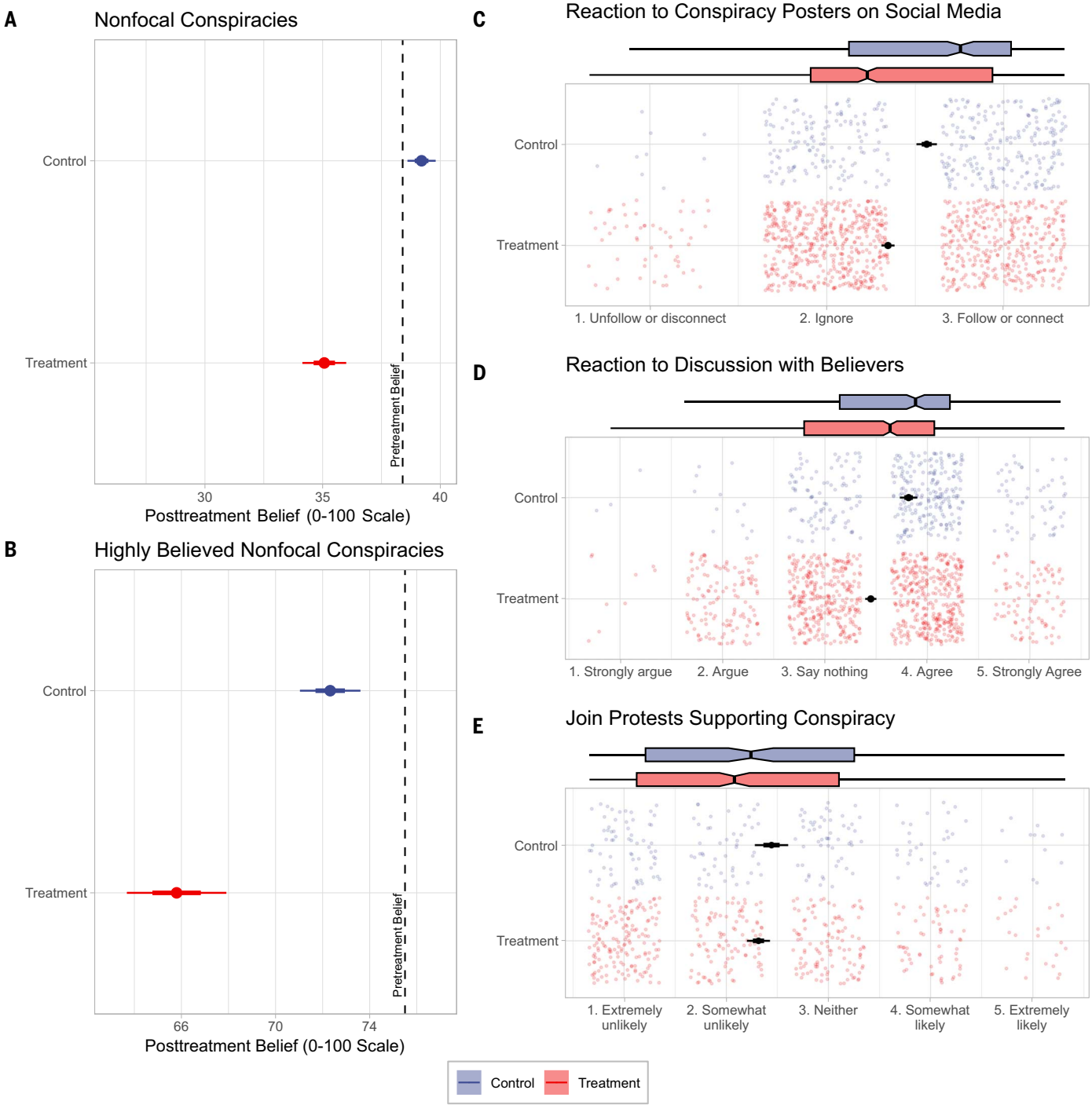


Fig. 4. The treatment also affects belief in other conspiracy theories and behavioral intentions. (**A** and **B**) Postconversation average belief in the 15 conspiracies from the BCTI (excluding the focal conspiracy, if it was one of those 15) by condition, for all conspiracies (A) and for only the subset of conspiracies that the participant indicated they believed in pretreatment (B). Vertical dashed line indicates average pretreatment belief. (**C** to **E**) Postconversation behavioral intentions by condition. Shown are participants' intentions regarding how they would respond to social media users who espouse their focal conspiracy (C), how they would behave in conversation with someone who believes the focal conspiracy (D), and how likely they would be to participate in a protest in support of the focal conspiracy (E). Thick error bars indicate 66% confidence intervals, and thin error bars indicate 95% confidence intervals. Boxplots narrow at the median.

CI [−11.06, −7.72], $P < 0.001$, 12% decrease; Fig. 4B and table S11), compared with a 3.32-point reduction in the control ($d = 0.53$). This difference between treatment and control persisted at the 2-month follow-up ($b_{\Delta\text{Treatment–Control}} = -5.34$, 95% CI [−8.40, −2.29], $P < 0.001$).

In study 2, we investigated the treatment's influence on participants' behavioral intentions (see SM supplementary text section 5). We found

that the treatment significantly increased intentions to ignore or unfollow social media accounts espousing the focal conspiracy (β = 0.39 [0.27, 0.50], $P < 0.001$; Fig. 4C and table S19) and significantly increased willingness to ignore or argue against people who believe the focal conspiracy (β = 0.42 [0.31, 0.54], $P < 0.001$; Fig. 4D and table S20). There was a directional but nonsignificant decrease in intentions to join pro-conspiracy protests (β = −0.12 [−0.27, 0.03], $P = 0.12$; Fig. 4E and table S21)—intentions that were low at baseline, potentially creating a floor effect.

How accurate is the AI?

Although it was not possible for us to ensure that all the claims produced by the AI in our experiment were accurate, we hired a professional fact-checker to evaluate the veracity and potential bias of all 128 claims made by GPT-4 Turbo across representative example conversations from each of the 11 major conspiracy clusters generated by participants in our experiments. Of these claims, 127 (99.2%) were rated as "true," 1 (0.8%) as "misleading," and 0 as "false"; and none of the claims were found to contain liberal or conservative bias. Together with a recent benchmarking study that found that only 2.5% of the claims produced by GPT-4 Turbo when summarizing text were hallucinations (*41*), these findings give us reason to believe that the information provided by the AI in our studies was largely accurate.

Furthermore, in 1.2% of cases, the participant named a focal conspiracy that was unambiguously true (e.g., MK Ultra). In these cases, the treatment effect was nonsignificant and directionally positive ($b = 6.51$, 95% CI [−39.42, 52.45], $P = 0.76$, $d = 0.43$) and significantly different from the effect for the other conspiracies ($b_{\Delta\text{True–False Conspiracies}} = -20.57$, 95% CI [−33.14, −8.00], $P = 0.001$; see SM supplementary text section 2.3).

Discussion

Although conspiracy theories are widely seen as a paradigmatic example of beliefs that rarely change in response to evidence (*8*–*10*), we hypothesized that dialogues with LLMs—which can use facts and evidence to rebut the specific claims made by any given conspiracy believer—would be efficacious in debunking conspiracy beliefs. Our findings confirmed this prediction: A brief interaction with a pretrained LLM substantially reduced belief in a wide range of conspiracy theories. The robustness of this effect is particularly noteworthy: (i) It occurred for both conspiracies that participants articulated in their own words and a broader conspiratorial worldview, (ii) it was evident even among participants with strong commitment to their chosen conspiracy, and (iii) its impact persisted, virtually undiminished, for (at least) 2 months after the intervention. Dialogues with the AI produced a meaningful and enduring shift in beliefs among a meaningful proportion of committed conspiracy believers in our study.

Theoretical, practical, and methodological advances

Our findings fundamentally challenge the view that evidence and arguments are of little use once someone has "gone down the rabbit hole" and come to believe a conspiracy theory. They also call into question social-psychological theories that center psychological "needs" and motivations as primary drivers of conspiratorial belief (*1*, *15*, *42*). Instead, our results align more closely with an alternative theoretical perspective that posits a central role for analytic thinking in protecting against epistemically suspect beliefs and behaviors (*14*), such as superstitions and paranormal beliefs (*43*), misinformation (*44*), and pseudo-profound bullshit (*45*). This viewpoint suggests that reasoning is not unduly constrained by identity needs and non-accuracy motivations; rather, people are generally willing to update their beliefs when presented with compelling evidence (*46*). Our study supports this perspective in several ways. Most straightforwardly, many conspiracists—including those strongly committed to their beliefs—updated their views when confronted with an AI that argued compellingly against their positions. Furthermore, the AI primarily provided alternative, nonconspiratorial explanations and evidence while encouraging critical thinking, rather than attempting to satisfy psychological needs (see SM supplementary text section 6). The durability of our findings across 2 months, along with the intervention's spillover effects on unrelated conspiracies and behavioral intentions, also suggests that participants seriously considered and internalized the AI's arguments—consistent with the "central route" to persuasion (*47*), which is known to promote durable belief change (and in contrast to the "peripheral route," which leverages superficial identity cues or emotional appeals and produces more ephemeral changes). Of course, our results do not wholly rule out some role for needs and motivations in the formation and maintenance of conspiracy beliefs, but they do indicate an important (perhaps countervailing, in some cases) role for evidence-based deliberation—especially in challenging and changing these beliefs once they are established. It is important to note that our goal—refuting existing conspiratorial beliefs—is distinct from other related challenges that have received more attention in the literature. These include presenting and then debunking pro-conspiracy arguments (*48*, *49*), attempting to increase resistance to conspiracy theories in general (*50*, *51*), and presenting arguments against a specific conspiracy theory to randomly selected crowd workers (*52*–*55*), which have been targeted with modest success by past work (e.g., meta-analytic $g = 0.16$ across 273 effect sizes) (*12*).

Our findings also have practical implications. Most broadly—in contrast to notions of a "post-truth" world in which facts no longer matter—arguments and evidence should not be abandoned by those seeking to reduce belief in dubious conspiracy theories. More specifically, AI models are powerful, flexible tools for reducing epistemically suspect beliefs and have the potential to be deployed to provide accurate information at scale. For example, internet search terms related to conspiracies could be met with AI-generated summaries of accurate information—tailored to the precise search—that solicit the user's response and engagement. Similarly, AI-powered social media accounts could reply to users who share inaccurate conspiracy-related content (providing corrective information for the potential benefit of both the poster and observers). Consistent with the potential for uptake of AI dialogues, some conspiracy-believing respondents in our sample expressed excitement and appreciation in their conversations with the AI (e.g., "Now this is the very first time I have gotten a response that made real, logical, sense. I must admit this really shifted my imagination when it comes to the subject of Illuminati. I think it was extremely helpful in my conclusion of rather the Illuminati is actually real."). However, it is quite unlikely that all, or even many, entrenched believers will choose to engage with AI chatbots. Exploring a variety of short- and long-term strategies to encourage engagement—such as gamification, transparency efforts (e.g., disclosing the AI model prompt and fine-tuning; clearly labeling sources), incentive programs, anonymous interaction options, and the integration of AI-assisted critical thinking exercises into school curricula—is an important direction for future applied work.

The effectiveness of AI persuasion demonstrated in our studies also relates to ongoing debates regarding the promise versus peril of generative AI (*56*, *57*). In our experiments, we sought to use AI to increase the accuracy of people's beliefs by debunking conspiracy theories. Absent appropriate guardrails, however, it is entirely possible that such models could also convince people to adopt epistemically suspect beliefs (*58*)—or be used as tools of large-scale persuasion more generally (*59*). Thus, our findings emphasize both the potential positive impacts of generative AI when deployed responsibly and the pressing importance of minimizing opportunities for this technology to be used irresponsibly. One especially key outstanding question, with far-reaching implications for AI's impact on the

global information ecosystem, is the degree of (a)symmetry in the efficacy of AI-based persuasion for true versus false content.

Finally, the experimental paradigm presented in this paper represents a substantial methodological advancement in behavioral science. Traditional survey experiments typically rely on static, predetermined stimuli and questions, which limits their ability to probe and respond to individuals' beliefs (*60*). In contrast, the real-time use of LLMs embedded in a survey enables the researcher to elicit open-ended statements of belief (or anything else) and translate them into quantitative outcomes (*61*). As we have seen, AI can engage in back-and-forth dialogues with participants, adapting its responses on the basis of the specific information provided by each individual (as opposed to, for example, using LLMs to pregenerate static stimuli, as in past work) (*62*–*66*). This personalized approach is particularly valuable when studying complex phenomena such as conspiracy beliefs, where a one-size-fits-all intervention may be less effective (*11*, *12*, *55*, *67*). The open-ended nature of the human-AI conversations also produces rich textual data that can be analyzed using natural language processing or qualitative techniques (*68*), which allows researchers to gain deeper insights into the content and structure of participants' beliefs, as well as the strategies used by the AI to challenge those beliefs. Integrating human-LLM interactions into behavioral science has the potential to meaningfully enhance our understanding of complex psychological phenomena.

Limitations and future directions

Although our results are promising, there are important limitations to highlight. Our study primarily relied on American online survey respondents who chose to participate in studies for material compensation, which raises questions about generalizability. Future work should test whether our findings extend to conspiracy believers who do not typically participate in survey studies, as well as to populations from countries and cultures beyond the United States. Although many participants in our study expressed maximal confidence in their conspiracy beliefs, it remains to be seen whether AI dialogues would effectively change the beliefs of even more entrenched conspiracy adherents, such as those actively participating in conspiracy-related groups or events. Moreover, our use of GPT-4 Turbo, a frontier, closed-source, pretrained, and fine-tuned language model, presents challenges related to interpretability and replicability (*69*–*71*). Although GPT-4 demonstrated both high accuracy and persuasiveness, serving as a proof of concept for AI-driven debunking, it remains unknown whether other models would perform similarly along either or both dimensions (*72*). This uncertainty extends to the potentially interactive relationship between accuracy and persuasive capacity: Hallucinations or lies may afford more compelling arguments, allowing models with less restrictive guardrails to outcompete heavily moderated models such as GPT-4 on persuasion. Finally, the causal mechanisms underpinning our results remain unformalized. While our study demonstrates the effectiveness of AI-facilitated dialogues in changing conspiracy beliefs, the specific cognitive or psychological processes through which this change occurs are unusually difficult to confirm—each conversation was unique and contained an admixture of rational argumentation and social cues. Both qualitative examination of the conversations and a structured, natural language processing–based analysis of the persuasive strategies used by the AI (see SM supplementary text section 6) suggest that fact-based argumentation was the focal point of each interaction; future research should examine this in greater detail.

Conclusions

It has become almost a truism that people "down the rabbit hole" of conspiracy belief are virtually impossible to reach. In contrast to this pessimistic view, we have shown that a relatively brief conversation with a generative AI model can produce a large and lasting decrease in conspiracy beliefs, even among people whose beliefs are deeply entrenched. It may be that it has proven so difficult to dissuade people from their conspiracy beliefs because they simply have not been given sufficiently good counterevidence. This paints a picture of human reasoning that is surprisingly optimistic: Even the deepest of rabbit holes may have an exit. Conspiracists are not necessarily blinded by psychological needs and motivations—it just takes a genuinely strong argument to reach them.

Materials and methods

Informed consent was obtained from participants before each study began. The studies did not involve deception, and after the studies were completed, all participants were debriefed and informed about the limitations and constraints of generative AI models. All studies were deemed minimal risk and exempt by the MIT Committee on the Use of Humans as Experimental Subjects (protocol E-5539).

We excluded participants for inattentiveness (both before they entered the study, using an open-ended text response, and early on in the study before random assignment using an attention check item). All studies were preregistered (see aspredicted.org/RPG_RY9, aspredicted.org/HSD_41Q, and aspredicted.org/KSN_PNL). Any non-preregistered analyses are labeled "post hoc," and any deviations from the preregistrations are reported. Conversational data from all participants, including those removed from our analyses, is available via web application. All GPT-4 model prompts used during the experiment are provided in table S2.

Study 1

Participants

We preregistered a target sample of 1000 responses from CloudResearch's Connect participant pool. In total, 1214 individuals began the survey (this includes 75 participants from a pilot conducted before the preregistration; for completeness, we include these participants in our analyses, but excluding them does not qualitatively change the results). An initial (pretreatment) screener only allowed participants who passed a writing quality and coherence screener to continue and complete the survey. The purpose of this screening criterion was to ensure that participants were not using automated survey completion programs, were capable of reading and writing in English, and were willing to answer the sort of open-ended questions on which the intervention relies. Of the participants who entered the survey, 70 failed this writing screener. A further 14 participants failed pretreatment attention checks and were removed from the survey; 90 discontinued before reaching the treatment. Further, using preregistered criteria, we excluded 156 participants who did not supply a genuine conspiracy theory (e.g., by noting that they do not believe any conspiracy theories in the open-ended response), 56 participants who provided a genuine conspiracy theory but endorsed it at below 50% veracity, and 54 participants for whom the AI provided an inaccurate summary (see SM supplementary text section 1 and fig. S2). Thus, 774 participants were included in our analyses (although all who passed the writing screener were allowed to complete the experiment). The overall attrition rate was 1.8%. Using a logistic regression model predicting whether a person attrited, we found no evidence of differential rates of attrition in treatment versus control ($b = -0.53$, $P = 0.37$). The treatment sample [mean age = 45.7, mean ideology = 3.04 on a scale from 1 (liberal) to 6 (conservative)] included 383 males, 384 females, and 7 participants who selected another gender option. A balance check found that our sample was balanced on pretreatment covariates (see table S1). This study was run on 19 to 22 January 2024 and took 30.98 min on average to complete.

Pretreatment measures

Participants completed a battery of self-report measures concerning their endorsement of a diverse set of 15 conspiracy beliefs, their attitudes concerning AI, and demographic items including beliefs about politics and religion. Conspiracy beliefs were assessed using a modified version of the Belief in Conspiracy Theories

Inventory (α = 0.90; example item: "Government agencies in the UK are involved in the distribution of illegal drugs to ethnic minorities") (*73*), which updated several items to reflect contemporary versions of the original (e.g., "SARS" was swapped with "COVID-19"). The scale labels ranged from "0 (Definitely False)" to "25 (Probably False)" to "50 (Uncertain)" to "75 (Probably True)" to "100 (Definitely True)," with the mean score in the treatment sample being 38.6% (SD = 20.0%). In addition to the 15 false conspiracy theories comprising the BCTI, we included three true conspiracy theories (pertaining to Project MK Ultra, Operation Northwoods, and the tobacco industry). Attitudes concerning AI were measured using items adapted from a Pew survey (*74*).

Subsequently, participants responded to an open-ended question concerning a conspiracy theory that they support (which we refer to as the "focal conspiracy")

> "What is a significant conspiracy theory that you find credible and compelling? Could you please describe this theory and share why it resonates with you?"

They then were asked to elaborate on the next page

> "On the previous question, you wrote [RESPONSE]. Can you describe in detail the specific evidence or events that initially led you to believe in this conspiracy theory? How do you interpret this evidence in relation to commonly accepted explanations for the same events?"

This information was fed forward to an instance of GPT-4 Turbo, which was tasked with summarizing the conspiratorial belief into a single sentence (see table S4 for the exact wording of this API query). Participants were then asked to rate their belief in the summarized conspiracy's veracity ("Please indicate your level of confidence that this statement is true") using a scale that ranged from 0 ("Definitely False") to 25 ("Probably False") to 50 ("Uncertain") to 75 ("Probably True") to 100 ("Definitely True").

Human-AI dialogues

After these pretreatment measures, participants were informed that they would be conversing with an advanced AI. To facilitate this real-time interaction within the Qualtrics survey platform, we used JavaScript to call OpenAI's Chat Completions API, dynamically inject participant-specific information into the model's instructions, and display the model's responses. Several details of this approach are worth mentioning. First, we used the latest available GPT-4 model, which was gpt-4-1106-preview for study 1 and gpt-4-0125-preview for study 2. Second, conversations were formatted to begin with a system message, followed by alternating user and AI messages. The system message, in our case, included details about the context, the goal of refuting a conspiracy belief, and instructions for how the model should behave (all of which were invariant across each model call), as well as the participant's specific conspiracy theory, the participant's stated reasons for believing that theory, and the participant's level of belief in the conspiracy (which varied across participants). Otherwise, our model instructions were simple and did not involve hidden reasoning steps (e.g., chain-of-thought) or access to external tools (e.g., internet browsing) beyond those provided by default. To facilitate a continuous conversation, for rounds 2 and 3, the previous AI and human messages were included in the prompt as conversation histories. Third, no token limit was placed on the AI's responses, which frequently comprised hundreds of words (fig. S3), multiple paragraphs, and markdown formatting (e.g., lists and section headings). Thus, although each dialogue only lasted three rounds, the dialogues (i) represented 8.4 min of AI-human engagement on average and (ii) were information dense yet comparatively easy to read and parse. Fourth, the AI's messages were sent to participants after the full response was constructed (rather than streamed word by word), necessitating idle time between each round of dialogue during which a loading screen was shown.

In the treatment condition, the AI was instructed to argue persuasively against the participant's conspiracy theory. In the control conditions, the AI was instructed to either (i) discuss the American medical system, (ii) debate with participants about whether they prefer dogs or cats, or (iii) discuss participants' past experiences with firefighters. We used a 60/40 split when randomizing participants into the treatment or control conditions, and participants assigned to the control were further randomized to one of the three control conditions, such that roughly 13 to 14% of the sample was assigned to each control condition. No significant differences were identified across the control groups, so we pooled them for all subsequent analyses.

Posttreatment measures

After the conversations, participants rerated their belief in the focal conspiracy and then again completed the modified BCTI (α = 0.92). Given that, in many cases, participants' focal conspiracies resembled at least one item on the BCTI (the items were chosen to reflect the most popular conspiracy theories), we computed three versions of pre- and posttreatment BCTI scores. The first version was the mean response on all 15 BCTI items, which we used to identify participants with a highly conspiratorial worldview. In the second version, we dropped items that matched the participants' focal conspiracy theory. Overlap was identified using an instance of GPT-4 that was supplied with each participant's conspiracy and each BCTI item and queried concerning which of the BCTI items reflected an affirmative belief in the participant's conspiracy using a binary judgment (see SM supplementary text section 7), yielding overlap-adjusted BCTI scores for pretreatment (α = 0.90) and posttreatment (α = 0.92). Thirdly, we further filtered the BCTI item pool by retaining nonoverlapping items that participants initially rated above 50% (more belief than "uncertain"), which allowed for pretreatment (α = 0.90) and posttreatment (α = 0.90) overlap-adjusted BCTI scores for conspiracy theories that each participant actively endorsed. We also administered the three true conspiracy items.

Recontacting at 10 days and 2 months

The participants from study 1 were recontacted twice. The first recontact occurred 10 days after completing the intervention (T3; n = 631, dropout rate = 15.7 and 15.6% for the treatment and control groups, respectively). Participants in the treatment condition who completed the T3 follow-up did not significantly differ from those who did not return for either pretreatment belief in their chosen conspiracy (t[454] = 0.61, P = 0.544) or on the pretreatment BCTI (t[454] = −0.71, P = 0.475). Participants completed the same dependent variables as in study 1 (i.e., endorsement of their chosen conspiracy theory and the BCTI). The second recontact occurred 2 months (T4) after completing the intervention (n = 529, dropout rate = 32.1 and 31.1% for the treatment and control groups, respectively). As with T3, participants in the treatment who remained did not differ from those who dropped out for either pretreatment belief in their chosen conspiracy (t[450] = 0.02, P = 0.977) or on the pretreatment BCTI (t[450] = −1.33, P = 0.183).

Study 2

For Study 2, two additional samples (study 2a and 2b) were fielded from CloudResearch Connect to corroborate, replicate, and extend our experimental findings. Although most materials were identical across studies 2a and 2b, we describe them separately because (i) we preregistered separate rounds of data collection, (ii) we used different phrasings for the behavioral outcome items, and (iii) the data were collected several weeks apart. Particularly, we carried out study 2b because of imprecise wording used in certain behavioral outcome items in study 2a, as noted below. In the main

text, results are pooled across studies 2a and 2b, except for those pertaining to the behavioral outcomes that were modified between 2a and 2b.

Participants

In study 2a, we preregistered a target sample of 1000 complete responses from CloudResearch's Connect participant pool, using quota-based sampling for age, race, ethnicity, and gender. A total of 1427 individuals entered the survey, of whom 312 were redirected for using a cell phone, 30 failed the initial pretreatment writing screener, 14 failed an attention check, and 104 discontinued before treatment, leaving 967. Of these participants, 237 did not provide a genuine conspiracy theory, and 81 endorsed their conspiracy statement at below 50% certainty—such that the final sample analyzed sample size was n = 649. Similarly, in study 2b, we recruited 1555 demographically representative participants using the Connect pool, of whom 27 were redirected for using a cell phone, 30 failed the writing screen, 27 failed an attention check, and 152 discontinued before treatment, leaving 1319. Of these participants, 325 did not provide a genuine conspiracy theory and another 128 did not endorse their conspiracy above 50%, leaving a treatment sample of 866.

Thus, the full sample size across both rounds of study 2 was n = 1515 (mean age = 41.9, mean ideology = 3.09), which included 726 males, 744 females, and 15 participants who selected another gender option (see fig. S4 and table S3). These studies were run on 25 to 28 February and 4 to 9 March 2024 and took 24.4 and 27.85 min on average, respectively, to complete. The overall attrition rate was 3.7%. Using a logistic regression model predicting whether or not a person attrited, we found no evidence of differential rates of attrition in treatment versus control (b = 0.02, P = 0.97).

Pretreatment measures

For all open-ended responses, including those in the human-AI dialogues, the "paste" functionality was disabled to prevent automated responding. As in study 1, participants began the experiment by answering a simple, writing-intensive question designed to gauge their willingness and ability to take part in a written conversation. Those whose responses were determined by GPT-4 Turbo to be low-effort or incoherent were redirected from the survey. Subsequently, participants completed self-report items about their AI attitudes and demographic characteristics (mirroring those from study 1). We did not administer the BCTI in study 2, and instead proceeded directly to the person-specific conspiracy assessment.

The wording of the person-specific instructions was modified slightly from study 1 to (i) explicitly define the theories to be described and (ii) only indirectly classify the theories as "conspiracies." The first question's wording was as follows

> "Throughout history, various theories have emerged that suggest certain significant events or situations are the result of secret plans by individuals or groups. These theories often offer alternative explanations for events than those that are widely accepted by the public or presented by official sources. Some people call these 'conspiracy theories.' Reflecting on this, are there any specific such theories that you find particularly credible or compelling? Please describe one below and share your reasons for finding it compelling."

And the follow-up question, presented on a separate page, was as follows

> "On the previous question, you wrote: "[conspiracy]". Could you share more about what led you to find this theory compelling? For instance, are there specific pieces of evidence, events, sources of information, or personal experiences that have particularly influenced your perspective? Please describe these in as much detail as you feel comfortable."

As in study 1, this information was fed forward to an instance of GPT-4 Turbo, which was tasked with summarizing the conspiratorial belief into a single sentence. Participants then provided a rating reflecting their confidence in the summarized statement's truth. The vast majority (90.6%) reported that the AI model accurately summarized their perspective; participants who received inaccurate summaries were excluded from subsequent analysis (note that this is a pretreatment exclusion). Before proceeding to the treatment, participants reported how important the conspiracy was to them ("How important is this theory to your personal beliefs or understanding of the world?") on a scale from 0 ("Not all all important to my beliefs and worldview") to 8 ("Extremely important to my beliefs and worldview").

Posttreatment measures

After the conversations, participants rerated the focal conspiracy's veracity and then completed a set of measures related to conspiracy-relevant behavior and trust. In both studies, we assessed (i) intentions to ignore or unfollow social media accounts espousing the focal conspiracy and (ii) willingness to ignore or argue against people who believe the focal conspiracy; in our analyses of these items, we pooled data across studies 2a and 2b. Study 2a also asked about (iii) willingness to engage in collective actions opposing the focal conspiracy and (iv) intentions to join protests related to the focal conspiracy theory. After data collection, however, we noticed problems in the wording of these items that made them uninterpretable, and thus we did not analyze these items. Item (iii), concerning collective actions, was both counterdirectionally worded (relative to the other items) and used a response scale containing negative and positive options that was not counterdirectionally worded, potentially resulting in a confused pattern of results. Item (iv), reflecting protest intentions, did not specify whether the protests supported or opposed the focal conspiracy, making responses to that item uninterpretable. In study 2b, we attempted to rectify these issues by dropping item (iii) and changing the wording of item (iv) to remove the ambiguity (i.e., "If people you knew were going to engage in a protest or action in support of the theory you described, how likely would you be to join in?"), as well as visually highlighting words indicating item directionality and having response-option direction randomized between participants and standardized within participants.

Finally, in study 2b, we asked GPT-4 Turbo to generate petitions opposing the participants' focal conspiracy theory, which we then asked participants if they wanted to sign. Unfortunately, inspecting these petitions indicated that many participants were not actually in opposition to the focal conspiracy theory (e.g., for a participant who thought the government was concealing the existence of aliens, GPT-4 Turbo asked whether they wanted to sign a petition calling for greater government transparency about aliens—which plays into the conspiracy theory, rather than opposing it). To determine how serious of a problem this was, we conducted a post hoc analysis in which 670 crowd workers each rated a random subset of three petitions as either "opposing" or "not opposing" its corresponding conspiracy theory after completing a brief training exercise. Of the 404 petitions rated at least twice, only 199 (49.3%) were rated as actually opposing the focal conspiracy in more than half of responses; and only 118 (29.2%) were unanimously rated as opposing the conspiracy. This makes participants' choice of whether to sign the petition not useful for determining the effect of the intervention, and thus we did not include analysis of it.

In both study 2a and study 2b, participants then completed measures of general trust (one item), personal trust (one item), and institutional trust (five items), which were adapted from the Organisation for Economic Cooperation and Development (OECD) Guidelines on Measuring Trust (*75*).

REFERENCES AND NOTES

1. S. M. Bowes, T. H. Costello, A. Tasimi, The conspiratorial mind: A meta-analytic review of motivational and personological correlates. *Psychol. Bull.* **149**, 259–293 (2023). doi: 10.1037/bul0000392; pmid: 37358543
2. M. Butter, P. Knight, Eds., *Routledge Handbook of Conspiracy Theories* (Routledge, 2020). doi: 10.4324/9780429452734
3. K. M. Douglas, R. M. Sutton, What are conspiracy theories? A definitional approach to their correlates, consequences, and communication. *Annu. Rev. Psychol.* **74**, 271–298 (2023). doi: 10.1146/annurev-psych-032420-031329; pmid: 36170672
4. J. E. Oliver, T. J. Wood, Conspiracy theories and the paranoid style(s) of mass opinion. *Am. J. Pol. Sci.* **58**, 952–966 (2014). doi: 10.1111/ajps.12084
5. H. G. West, T. Sanders, Eds., *Transparency and Conspiracy: Ethnographies of Suspicion in the New World Order* (Duke Univ. Press, 2003).
6. J. E. Uscinski, J. M. Parent, *American Conspiracy Theories* (Oxford Univ. Press, 2014). doi: 10.1093/acprof:oso/9780199351800.001.0001
7. J.-W. van Prooijen, K. M. Douglas, Belief in conspiracy theories: Basic principles of an emerging research domain. *Eur. J. Soc. Psychol.* **48**, 897–908 (2018). doi: 10.1002/ejsp.2530; pmid: 30555188
8. S. Lewandowsky, G. E. Gignac, K. Oberauer, The role of conspiracist ideation and worldviews in predicting rejection of science. *PLOS ONE* **8**, e75637 (2013). doi: 10.1371/journal.pone.0075637; pmid: 24098391
9. M. G. Napolitano, "Conspiracy theories and resistance to evidence," thesis, University of California, Irvine (2022).
10. C. R. Sunstein, A. Vermeule, Conspiracy theories: Causes and cures. *J. Polit. Philos.* **17**, 202–227 (2009). doi: 10.1111/j.1467-9760.2008.00325.x
11. C. O'Mahony, M. Brassil, G. Murphy, C. Linehan, The efficacy of interventions in reducing belief in conspiracy theories: A systematic review. *PLOS ONE* **18**, e0280902 (2023). doi: 10.1371/journal.pone.0280902; pmid: 37018172
12. L. Stasielowicz, The effectiveness of interventions addressing conspiracy beliefs: A meta-analysis. PsyArXiv [Preprint] (2024); https://doi.org/10.31234/osf.io/6vs5u.
13. J. K. Madsen *et al.*, Behavioral science should start by assuming people are reasonable. *Trends Cogn. Sci.* **28**, 583–585 (2024). doi: 10.1016/j.tics.2024.04.010; pmid: 38763803
14. G. Pennycook, Chapter Three - A framework for understanding reasoning errors: From fake news to climate change and beyond. *Adv. Exp. Soc. Psychol.* **67**, 131–208 (2023). doi: 10.1016/bs.aesp.2022.11.003
15. K. M. Douglas, R. M. Sutton, A. Cichocka, The psychology of conspiracy theories. *Curr. Dir. Psychol. Sci.* **26**, 538–542 (2017). doi: 10.1177/0963721417718261; pmid: 29276345
16. J. T. Jost, A. Ledgerwood, C. D. Hardin, Shared reality, system justification, and the relational basis of ideological beliefs. *Soc. Personal. Psychol. Compass* **2**, 171–186 (2008). doi: 10.1111/j.1751-9004.2007.00056.x
17. R. Hofstadter, *The Paranoid Style in American Politics* (Knopf Doubleday Publishing Group, 1964).
18. J. A. Whitson, A. D. Galinsky, Lacking control increases illusory pattern perception. *Science* **322**, 115–117 (2008). doi: 10.1126/science.1159845; pmid: 18832647
19. S. Lewandowsky *et al.*, Recurrent fury: Conspiratorial discourse in the blogosphere triggered by research on the role of conspiracist ideation in climate denial. *J. Soc. Polit. Psych.* **3**, 142–178 (2015). doi: 10.5964/jspp.v3i1.443
20. J.-W. van Prooijen, N. B. Jostmann, Belief in conspiracy theories: The influence of uncertainty and perceived morality. *Eur. J. Soc. Psychol.* **43**, 109–115 (2013). doi: 10.1002/ejsp.1922
21. J.-W. van Prooijen, An existential threat model of conspiracy theories. *Eur. Psychol.* **25**, 16–25 (2020). doi: 10.1027/1016-9040/a000381
22. A. Lantian, D. Muller, C. Nurra, K. M. Douglas, I know things they don't know! *Soc. Psychol. (Gott.)* **48**, 160–173 (2017). doi: 10.1027/1864-9335/a000306
23. M. Biddlestone *et al.*, Reasons to believe: A systematic review and meta-analytic synthesis of the motives associated with conspiracy beliefs. PsyArXiv [Preprint] (2022); https://doi.org/10.31234/osf.io/rxjqc.
24. M. Biddlestone, R. Green, A. Cichocka, R. Sutton, K. Douglas, Conspiracy beliefs and the individual, relational, and collective selves. *Soc. Personal. Psychol. Compass* **15**, e12639 (2021). doi: 10.1111/spc3.12639
25. A. Cichocka, M. Marchlewska, A. Golec de Zavala, M. Olechowski, 'They will not control us': Ingroup positivity and belief in intergroup conspiracies. *Br. J. Psychol.* **107**, 556–576 (2016). doi: 10.1111/bjop.12158; pmid: 26511288
26. A. Sternisko, A. Cichocka, A. Cislak, J. J. Van Bavel, National narcissism predicts the belief in and the dissemination of conspiracy theories during the COVID-19 pandemic: Evidence From 56 countries. *Pers. Soc. Psychol. Bull.* **49**, 48–65 (2023). doi: 10.1177/01461672211054947; pmid: 34872399
27. R. Brotherton, *Suspicious Minds: Why We Believe Conspiracy Theories* (Bloomsbury Publishing, 2015). doi: 10.5040/9781472944528
28. R. K. Garrett, B. E. Weeks, Epistemic beliefs' role in promoting misperceptions and conspiracist ideation. *PLOS ONE* **12**, e0184733 (2017). doi: 10.1371/journal.pone.0184733; pmid: 28922387
29. N. Dagnall, K. Drinkwater, A. Parker, A. Denovan, M. Parton, Conspiracy theory and cognitive style: A worldview. *Front. Psychol.* **6**, 206 (2015). doi: 10.3389/fpsyg.2015.00206; pmid: 25762969
30. S. Novella, *The Skeptics' Guide to the Universe: How to Know What's Really Real in a World Increasingly Full of Fake* (Hachette UK, 2018).
31. P. M. Fernbach, J. E. Bogard, Conspiracy theory as individual and group behavior: Observations from the Flat Earth International Conference. *Top. Cogn. Sci.* **16**, 187–205 (2024). doi: 10.1111/tops.12662; pmid: 37202921
32. H. Wang, J. Li, H. Wu, E. Hovy, Y. Sun, Pre-trained language models and their applications. *Engineering (Beijing)* **25**, 51–65 (2023). doi: 10.1016/j.eng.2022.04.024
33. OpenAI, GPT-4 technical report. arXiv:2303.08774 [cs.CL] (2024).
34. K. Arceneaux, B. N. Bakker, N. Fasching, Y. Lelkes, A critical evaluation and research agenda for the study of psychological dispositions and political attitudes. *Polit. Psychol.* 10.1111/pops.12958 (2024). doi: 10.1111/pops.12958
35. W. Yaqub, O. Kakhidze, M. L. Brockman, N. Memon, S. Patil, "Effects of credibility indicators on social media news sharing intent" in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (ACM, 2020), pp. 1–14.
36. R. Böhm, M. Jörling, L. Reiter, C. Fuchs, People devalue generative AI's competence but not its advice in addressing societal and personal challenges. *Community Psychol.* **1**, 32 (2023). doi: 10.1038/s44271-023-00032-x
37. Y. Zhang, R. Gosline, Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation. *Judgm. Decis. Mak.* **18**, e41 (2023). doi: 10.1017/jdm.2023.37
38. V. Veselovsky *et al.*, Prevalence and prevention of large language model use in crowd work. arXiv:2310.15683 [cs.CL] (2023).
39. M. N. Stagnaro, D. G. Rand, "The coevolution of religious belief and intuitive cognitive style via individual-level selection" in *The Oxford Handbook of Evolutionary Psychology and Religion*, J. R. Liddle, T. K. Shackelford, Eds. (Oxford Univ. Press, 2016), pp. 153–173. doi: 10.1093/oxfordhb/9780199397747.013.10
40. S. Athey, J. Tibshirani, S. Wager, Generalized random forests. *Ann. Stat.* **47**, 1148–1178 (2019). doi: 10.1214/18-AOS1709
41. S. Hughes, M. Bae, M. Li, Vectara Hallucination Leaderboard [Data set], GitHub (2023); https://github.com/vectara/hallucination-leaderboard.
42. M. J. Hornsey, K. Bierwiaczonek, K. Sassenberg, K. M. Douglas, Individual, intergroup and nation-level influences on belief in conspiracy theories. *Nat. Rev. Psychol.* **2**, 85–97 (2023). doi: 10.1038/s44159-022-00133-0; pmid: 36467717
43. G. Pennycook, J. A. Cheyne, P. Seli, D. J. Koehler, J. A. Fugelsang, Analytic cognitive style predicts religious and paranormal belief. *Cognition* **123**, 335–346 (2012). doi: 10.1016/j.cognition.2012.03.003; pmid: 22481051
44. G. Pennycook, D. G. Rand, Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* **188**, 39–50 (2019). doi: 10.1016/j.cognition.2018.06.011; pmid: 29935897
45. G. Pennycook, J. A. Cheyne, N. Barr, D. J. Koehler, J. A. Fugelsang, On the reception and detection of pseudo-profound bullshit. *Judgm. Decis. Mak.* **10**, 549–563 (2015). doi: 10.1017/S1930297500006999
46. B. M. Tappin, A. J. Berinsky, D. G. Rand, Partisans' receptivity to persuasive messaging is undiminished by countervailing party leader cues. *Nat. Hum. Behav.* **7**, 568–582 (2023). doi: 10.1038/s41562-023-01551-7; pmid: 36864137
47. R. E. Petty, J. T. Cacioppo, "The elaboration likelihood model of persuasion" in *Communication and Persuasion: Central and Peripheral Routes to Attitude Change* (Springer, 1986), pp. 1–24.
48. E. Porter, Y. Velez, T. J. Wood, Factual corrections eliminate false beliefs about COVID-19 vaccines. *Public Opin. Q.* **86**, 762–773 (2022). doi: 10.1093/poq/nfac034
49. G. Orosz *et al.*, Changing conspiracy beliefs through rationality and ridiculing. *Front. Psychol.* **7**, 1525 (2016). doi: 10.3389/fpsyg.2016.01525; pmid: 27790164
50. J. A. Banas, G. Miller, Inducing resistance to conspiracy theory propaganda: Testing inoculation and metainoculation strategies. *Hum. Commun. Res.* **39**, 184–207 (2013). doi: 10.1111/hcre.12000
51. E. Bonetto, J. Troïan, F. Varet, G. Lo Monaco, F. Girandola, Priming resistance to persuasion decreases adherence to conspiracy theories. *Soc. Influence* **13**, 125–136 (2018). doi: 10.1080/15534510.2018.1471415
52. V. Swami *et al.*, Lunar lies: The impact of informational framing and individual differences in shaping conspiracist beliefs about the moon landings. *Appl. Cogn. Psychol.* **27**, 71–80 (2013). doi: 10.1002/acp.2873
53. D. Jolley, K. M. Douglas, Prevention is better than cure: Addressing anti-vaccine conspiracy theories. *J. Appl. Soc. Psychol.* **47**, 459–469 (2017). doi: 10.1111/jasp.12453
54. S. Altay, A.-S. Hacquin, C. Chevallier, H. Mercier, Information delivered by a chatbot has a positive impact on COVID-19 vaccines attitudes and intentions. *J. Exp. Psychol. Appl.* **29**, 52–62 (2023). doi: 10.1037/xap0000400; pmid: 34726454
55. S. Altay *et al.*, Scaling up interactive argumentation by providing counterarguments with a chatbot. *Nat. Hum. Behav.* **6**, 579–592 (2022). doi: 10.1038/s41562-021-01271-w; pmid: 35165435
56. E. Klein, "This changes everything," *The New York Times*, 12 March 2023; https://www.nytimes.com/2023/03/12/opinion/chatbots-artificial-intelligence-future-weirdness.html.
57. D. Allen, E. G. Weyl, The real dangers of generative AI. *J. Democracy* **35**, 147–162 (2024). doi: 10.1353/jod.2024.a915355
58. M. Phuong *et al.*, Evaluating frontier models for dangerous capabilities. arXiv:2403.13793 [cs.LG] (2024).
59. M. Burtell, T. Woodside, Artificial influence: An analysis of AI-driven persuasion. arXiv:2303.08721 [cs.CY] (2023).
60. Y. Velez, Crowdsourced adaptive surveys. arXiv:2401.12986 [cs.CL] (2024).
61. Y. R. Velez, P. Liu, Confronting core issues: A critical assessment of attitude polarization. *Am. Polit. Sci. Rev.* 10.1017/S0003055424000819 (2024). doi: 10.1017/S0003055424000819
62. H. Bai, J. Voelkel, J. Eichstaedt, R. Willer, Artificial intelligence can persuade humans on political issues. OSF Preprints (2023); https://doi.org/10.31219/osf.io/stakv.
63. E. Karinshak, S. X. Liu, J. S. Park, J. T. Hancock, Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proc. ACM Hum. Comput. Interact.* **7**, 116 (2023). doi: 10.1145/3579592
64. K. Hackenburg, H. Margetts, Evaluating the persuasive influence of political microtargeting with large language models. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2403116121 (2024). doi: 10.1073/pnas.2403116121; pmid: 38848300
65. S. C. Matz *et al.*, The potential of generative AI for personalized persuasion at scale. *Sci. Rep.* **14**, 4692 (2024). doi: 10.1038/s41598-024-53755-0; pmid: 38409168
66. E. Durmus, L. Lovitt, A. Tamkin, S. Ritchie, J. Clark, D. Ganguli, "Measuring the persuasiveness of language models," Anthropic, 9 April 2024; https://www.anthropic.com/news/measuring-model-persuasiveness.
67. M. N. Williams *et al.*, People do change their beliefs about conspiracy theories-but not often. *Sci. Rep.* **14**, 3836 (2024). doi: 10.1038/s41598-024-51653-z; pmid: 38360799
68. C. Olah, A. Jermyn, "Reflections on qualitative research," Transformer Circuits Thread (2024); https://transformer-circuits.pub/2024/qualitative-essay/index.html.
69. Z. Hussain, M. Binz, R. Mata, D. U. Wulff, A tutorial on open-source large language models for behavioral science. PsyArXiv [Preprint] (2023); https://doi.org/10.31234/osf.io/f7stn.
70. A. Spirling, Why open-source generative AI models are an ethical way forward for science. *Nature* **616**, 413 (2023). doi: 10.1038/d41586-023-01295-4; pmid: 37072520
71. I. Grossmann *et al.*, AI and the transformation of social science research. *Science* **380**, 1108–1109 (2023). doi: 10.1126/science.adi1778; pmid: 37319216

72. K. Hackenburg *et al*., Evidence of a log scaling law for political persuasion with large language models. arXiv:2406.14508 [cs.CL] (2024).
73. V. Swami, T. Chamorro-Premuzic, A. Furnham, Unanswered questions: A preliminary investigation of personality and individual difference predictors of 9/11 conspiracist beliefs. *Appl. Cogn. Psychol.* **24**, 749–761 (2010). doi: 10.1002/acp.1583
74. M. Faverio, A. Tyson, "What the data says about Americans' views of artificial intelligence," Pew Research Center, 21 November 2023; https://www.pewresearch.org/short-reads/2023/11/21/what-the-data-says-about-americans-views-of-artificial-intelligence/.
75. OECD, *OECD Guidelines on Measuring Trust* (Organisation for Economic Cooperation and Development, 2017); https://www.oecd-ilibrary.org/governance/oecd-guidelines-on-measuring-trust_9789264278219-en.
76. T. Costello, G. Pennycook, D. Rand, Durably reducing conspiracy beliefs through dialogues with AI [Dataset], Dryad (2024); https://doi.org/10.5061/dryad.v6wwpzh4h.

ACKNOWLEDGMENTS

Funding: MIT Generative AI Initiative (D.G.R.) and John Templeton Foundation Grant 61779 (G.P.). **Author contributions:** Conceptualization: T.H.C., G.P., and D.R. Methodology: T.H.C., G.P., and D.R. Investigation: T.H.C., G.P., and D.R. Visualization: T.H.C., G.P., and D.R. Funding acquisition: G.P. and D.R. Project administration: T.H.C. and D.R. Supervision: G.P. and D.R. Writing – original draft: T.H.C., G.P., and D.R. Writing – review & editing: T.H.C., G.P., and D.R. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Relevant data, analytic code, study materials, and preregistration documents are accessible in Dryad (76).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adq1814
Supplementary Text
Figs. S1 to S15
Tables S1 to S21
References (*77*–*89*)

Submitted 1 May 2024; accepted 18 July 2024
10.1126/science.adq1814