Just the Facts: How Dialogues with AI Reduce Conspiracy Beliefs

Thomas H. Costello[12]*, Gordon Pennycook[3], David G. Rand[2].

[1]Department of Psychology, American University; Washington, DC, USA

[2]Sloan School of Management, Massachusetts Institute of Technology; Cambridge, USA

[3]Department of Psychology, Cornell University; Ithaca, USA

*Corresponding author. Email: thcostello1@gmail.com

Abstract: Conspiracy beliefs are widely considered resistant to factual correction, yet recent research shows that relatively brief, personalized "debunking" dialogues with a generative AI model can substantially reduce such beliefs. To identify the mechanisms driving this effect, we conducted an experiment spanning eight treatment arms that varied key features of participants' interactions with GPT-4 during such debunking dialogues (N = 1,297). The debunking effect proved robust across most manipulations—including whether participants were explicitly told the AI aimed to change their minds, were asked to debate the AI, or whether the AI offered them factual information without otherwise seeking to persuade or was concise in its exposition. The only condition that undermined the debunking effect was prompting the AI to persuade participants without presenting any counterevidence, which yielded a null effect. Furthermore, analyses of the AI's persuasive strategies identified reasoning-based tactics as the sole significant mediator of belief change. Participants who reported feeling persuaded overwhelmingly cited the AI's rational, evidence-focused approach. Finally, participants higher in actively open-minded thinking showed larger treatment effects. These findings suggest that AI-driven interventions reduce conspiracy beliefs principally by providing factual, targeted counterarguments that address the specific reasons people hold these beliefs.

Supplemental Materials
Analytic Code and Results (Available upon publication)
Conversation Browser
Preregistration
Data (Available upon publication)

Last update: 02/16/2025

This manuscript is a pre-publication working paper. It has not undergone peer review.

Conspiracy theories, in which events are understood as being caused by secret, malevolent plots involving powerful conspirators, often strain credulity[1], weaving together elements that are patently absurd (e.g., shape-shifting reptilian aliens controlling world governments), mythical or supernatural (e.g., ancient giants building the pyramids), melodramatically evil (e.g., powerful elites drinking children's blood for immortality), and overtly illogical (e.g., believing COVID-19 is simultaneously a hoax and an engineered bioweapon). As conspiracy theory beliefs spread through communities, they both map onto existing socio-cultural divides (Mus et al., 2022; van Prooijen & van Vugt, 2018) and act as symptoms of, and perhaps catalysts for, social, epistemic, and institutional corrosion (Pierre, 2020; Pummerer, 2022) – with second-order implications for public health, democracy, extremism, and scientific progress (Ecker et al., 2024).

Both lay and academic theories maintain that once individuals have come to believe in a conspiracy – or ventured down the proverbial "rabbit hole" – factual corrections and counterevidence rarely convince them to stop believing (Lewandowsky et al., 2013; Sunstein & Vermeule, 2008; Walter et al., 2020). Such epistemic obstinance is often attributed to a suite of cognitive biases and psychological peculiarities, united in their shared resonance with ideologically motivated reasoning (Douglas et al., 2024). For instance, conspiracy beliefs are thought to serve important identity and meaning-making functions (Biddlestone et al., 2021), create self-sealing belief systems where contradictory evidence is reinterpreted as further proof of the conspiracy (Lewandowsky et al., 2013), and activate biased assimilation processes that lead believers to scrutinize challenging information more critically than supporting information (Dagnall et al., 2015). Given that this set of proximal causes does not include rational belief updating (i.e., where inaccurate or misleading information causes conspiracy belief), direct factual rebuttals of conspiracies have long been considered futile (i.e., to paraphrase the author Jonathan Swift: you cannot reason people out of positions they didn't reason themselves into). Consistent with this perspective, there has been a distinct lack of interventions that successfully reduce conspiracies among people who already believe them (whether via factual corrections or alternative treatments) - and thus practical approaches to combating conspiracy theories have largely focused on prevention rather than intervention (Lewandowsky et al., 2020).

However, recent evidence has challenged this perspective by showing that many conspiracy believers *do* change their minds following information-focused debunking dialogues. Specifically, personalized conversations with the large language model GPT-4 Turbo, involving three rounds of back-and-forth dialogue in which the AI provided counterarguments and factual evidence rebutting the believer's claims, were found to reduce conspiracy beliefs by 17-20% (*d*s from 0.70 to 1.1), even among so-called "true believers" (i.e., those who reported complete confidence or to whom the theory was particularly important to their worldview) (Costello et al., 2024). Persuasive effects of this size are unprecedented in the conspiracy belief literature[2], although other fact-based interventions have yielded (smaller) belief change effects

[1] While many conspiracies represent unverified speculations that are unsupported by evidence, some conspiracies are true (e.g., Watergate). Most theorizing about conspiracy beliefs focuses on unverified conspiracies; we therefore focus on those here.

[2] The magnitude of this effect is especially surprising since, in Costello et al (2024), participants articulated and explain their conspiracy beliefs in natural language, rather than simply selecting from pre-written conspiracy statements or responding after exposure to pro-conspiracy arguments (i.e., methods commonly used in intervention studies that likely overestimate how easily conspiracy beliefs can be changed).

for related topics such as vaccine misinformation (Altay et al., 2022, 2023; Porter et al., 2022). The psychological mechanisms underpinning this AI-dialogue debunking effect, however, are unclear. Here, we shed light on this issue by experimentally testing the predictions of numerous possible mechanisms in a large, omnibus study.

Competing accounts of the efficacy of AI-facilitated conversational debunking

Why might dialogues with an AI model lead to the observed large reductions in conspiratorial beliefs? One potential mechanism comes from theories of classical reasoning (Pennycook, 2023). To the extent that people use deliberation to form accurate beliefs, providing compelling reasons and counter-evidence to conspiratorial claims - regardless of how precisely those reasons and counter-evidence are delivered - should cause belief updating. By this account, the relative inefficacy of previous attempts at fact-based intervention may reflect the logistical challenge of providing individual conspiracy believers with compelling evidence at scale, rather than the futility of rationality in the face of motivated reasoning. People believe in heterogeneous conspiracies for heterogeneous reasons (Costello et al., 2023; Pierre, 2020). Generic debunking attempts that argue broadly against a given conspiracy may fail, therefore, simply because they do not address the specific evidence accepted by individual believers (Costello et al., 2024). Generative AI tools offer a promising solution to these challenges through two key capabilities: (i) access to vast amounts of information across diverse topics and (ii) the ability to tailor counterarguments to specific conspiracies, reasoning, and evidence. These capabilities allow large language models to respond directly to—and refute—the particular evidence supporting an individual's conspiratorial beliefs: an intense form of personalization that goes far beyond the "micro-targeting" implicated in the infamous Cambridge Analytica scandal (i.e., message personalization based on bluntly predictive cues, such as demographics or broadband personality, which is of limited efficacy; Tappin et al., 2023). The interactive, discursive, argumentative nature of human-AI exchanges may also be instrumental because humans are naturally equipped to process and evaluate arguments in dialogue (Mercier, 2016).

A second potential mechanism, rooted in theories of source credibility (Hovland & Weiss, 1951; Pornpitakpan, 2004), involves deferring to the AI because - unlike a human interlocutor - the AI is seen as objective, accurate, rule-governed, and even infallible (i.e., the "machine heuristic"; Sundar & Kim, 2019). Currently, generative AI occupies a unique epistemic position – having not (yet?) come to be perceived as in alignment with the cultural-political coalitions (e.g., right- vs. left-"coded") that shape source credibility judgments (Williams, 2023). Participants may therefore view AI models as unbiased and objective; neutral arbiters of truth and falsity with no discernable motive other than helpfulness. Particularly, in Costello et al. (2024), participants were informed they would be conversing with a neutral AI model, and were not informed about the AI model's instructions. Thus, when the AI provided a strong argument against the participants' chosen conspiracies, many people may have assumed that the ostensibly unbiased AI was, helpfully, seeking to correct their misbelief without an agenda. This perceived neutrality (or perceptions of persuasion-related attributes, such as confidence; Colombatto & Fleming, 2024) may lead conspiracy believers to view LLMs as

particularly credible sources of information (Glickman & Sharot, 2024; Goel et al., 2024), and therefore make them more open to belief revision than would otherwise be expected.

A third potential mechanism involves the non-adversarial nature of the interaction with the AI model. When the dialogue is presented as a study of whether humans and AI can successfully discuss complex topics, as in Costello et al (2024), this places a cooperative frame on the interaction. Thus, participants may be particularly open to engaging with the AI in good faith and without defensiveness (Sherman & Cohen, 2006) - which may circumvent the psychological barriers that typically make conspiracy beliefs resistant to change. Further, the affirming and validational tone struck by the AI in most of the debunking conversations – mirroring the "common factors" of psychotherapy (Wampold, 2015) – may have served to alleviate or defuse the very psychological needs for understanding, control, and safety that are popularly theorized to cause conspiracism.

A fourth potential mechanism, rooted in theories of information overload (Eppler & Mengis, 2004; Sweller, 1988), involves the sheer volume of the factual information provided by the AI. When presented with a large amount of information, individuals' cognitive processing capacity can become overloaded (Eppler & Mengis, 2004; Sweller, 1988), perhaps leading participants to evaluate the AI's arguments less critically or to superficially acquiesce. Thus, it may be that the AI did not actually change people's minds, but instead simply overwhelmed them.

A fifth potential mechanism involves the persuasive powers of the AI itself (Jones & Bergen, 2024). During their training, models could learn to leverage strategies that facilitate persuasion, including argumentation schemes (Walton et al., 2008), rhetoric, and emotional manipulation. Perhaps the large observed reductions in conspiracy beliefs were caused by the proficiency of AI models in these strategies. Consistent with this possibility, for example, is a recent study that found that an AI model was much more effective than an average human at changing an opponent's mind in a debate (Salvi et al., 2024).

Beyond these psychological mechanisms, there are possible deflationary explanations whereby the apparent AI debunking effect is simply the result of methodological artifacts. One such critique concerns soliciting open-ended descriptions and explanations of people's thinking processes (i.e., why they believe a particular conspiracy) before administering closed-ended belief assessments. This sequence may introduce error, for reasons akin to those that lead to reactivity during some verbal reporting procedures (Fox et al., 2011) or due to the "illusion of explanatory depth" being punctured (Sloman & Vives, 2022). Or perhaps participants unconsciously processed information in a biased manner during the rating task to support the stance they articulated in writing (i.e., motivated reasoning). Relatedly, translating complex beliefs into writing may not capture their nuance, leading to AI-generated belief summaries (i.e., the key DV) that reflect an oversimplified or overly narrow understanding of the belief (e.g., "I believe in a 9/11 conspiracy because jet fuel does not burn hot enough to melt steel beams" rather than "9/11 was an inside job"). Thus, pre-post change on participants' endorsement of these miscalibrated items could reflect an unusually narrow form of debunking, wherein non-load-bearing or ornamental pillars of beliefs are toppled while the core belief system remains untouched. Finally, a distinct set of methodological explanations involves satisficing and demand characteristics: Participants may infer the purpose of the study based on the tasks

and report reduced conspiracy belief—regardless of their actual attitudes—to align with either the researchers' or AI's perceived expectations.

Current research

Understanding which factors explain the AI debunking effect is useful both theoretically—informing models of conspiracy belief change—and practically—guiding the design of future interventions. Therefore, the present research aims to adjudicate between these various explanations using a large pre-registered experiment. Beyond a baseline condition, which directly replicates the AI debunking treatment from Costello et al. (2024), we include a series of experimental conditions that address each of the various mechanisms and explanations detailed above.

To test the classical reasoning-based theory that reasons and counter-evidence are the keys to effective debunking, the "No Evidence" condition instructed the AI to attempt to persuade participants without employing any rational arguments or counterevidence. If evidence is indeed an essential ingredient in these persuasive dialogues, then we would expect a substantial blunting of the belief change effect relative to the baseline.

To test the theory that the AI model changed attitudes using extreme powers of persuasion and manipulation, the "Just-the-facts" condition simply instructed the AI to provide accurate factual information about the participant's conspiracy topic. All instructions to persuade the participant or get them to change their mind were removed. Thus, if the ability to manipulate is key to the AI's success, we would expect a substantial reduction in belief change relative to the baseline.

To test the information overload-based theory, that the AI buffaloed participants into submission by presenting a massive volume of information, the "Concise" condition instructed the AI to produce substantially shorter responses. Thus, if the AI merely informationally overloaded participants, we would expect a substantial reduction in belief change relative to control. Additionally, this condition is pertinent to the intervention's applicability to real-world contexts, where users may be disinclined to read multi-paragraph responses.

To test the theory that the AI was effective because participants perceived it as neutral and unbiased, the "Overt" condition explicitly informed participants that the AI would try to persuade them not to believe the conspiracy theory - thus removing the potential impression of impartiality. Past work indicates that simple primes indicating to participants that a conversational AI has particular motives (e.g., caring, manipulative) alters their perceptions of the AI, which, in turn, may shape the trajectory of their subsequent interactions with it (Pataranutaporn et al., 2023). If being perceived as neutral and even-handed is key to the effect, we would expect a substantial reduction in effect size.

Similarly, to test the theory that it was the collaborative framing of the interaction that made the AI effective, the "Adversarial" condition framed the interaction as a competitive debate. Specifically, participants were informed that the goal of the interaction was for them to try to convince the AI that the conspiracy was true while the AI would try to convince them it was false. If participants perceiving the interaction as non-adversarial is key to the effect, we would expect a substantial reduction in effect size. The results of this condition will also bear on

real-world applicability, as conspiracy believers may be more interested in entering a debate-style interaction with an AI.

Finally, to address methodological concerns, we included two additional experiment variations. In the "Belief-first Rating" condition, participants rated their conspiracy belief before explaining it in detail, addressing concerns about prior elaboration artificially inflating reported belief strength. In the "Abstracted Summary" condition, we used less specific AI-generated conspiracy belief summary statements (summarizes of the participants' written account of their belief that served as the key dependent variable), such that the generated summaries reflected only abstracted versions of the conspiracy theory rather than lower-level specific claims.

In all cases, participants wrote about their conspiracy belief and entered into a 3-round dialogue with GPT-4 Turbo, mirroring the procedures of Costello et al. (2024). We collected two samples using virtually identical procedures, which are pooled in our primary analyses (n = 3,032 completed the experiment, with n = 1297 actively endorsing a conspiracy theory and thus being included in the analysis).

Table 1. Overview of experimental conditions

Condition	Manipulation	Theory Tested
Baseline	Standard AI dialogue (Costello et al., 2024)	
Just-the-Facts	AI provides factual information without persuasion	**Persuasive language**: GPT-4 displays an exceptional facility for persuasion
No Evidence	AI persuades without using rational arguments or evidence	**Classical reasoning**: Reasoned argumentation causes belief updating
Concise	AI keeps responses concise, otherwise mirrors Baseline condition	**Informational overload**: Individuals' cognitive processing becomes overloaded during the conversations
Adversarial	Interaction framed as debate rather than dialogue	**Politeness:** The AI's friendliness avoids triggering defensive reactions, leading participants to consider the AI's counter-points out of sociability
Overt	Participants informed of AI's persuasive goals	**Source credibility**: Priors about the neutrality and unbiasedness of AI lead participants to consider AI-provided information trustworthy.
Abstracted Claims	AI summarizes core conspiracy claims only	**Methodological artifact (oversimplification)**: If specificity drives effect, abstracting reduces it
Belief-First Rating	Belief ratings collected before elaboration	**Methodological artifact (order effects)**: If elaboration inflates belief ratings, rating first reduces effect

Methods & Procedure

Participants

To be initially eligible, participants had to be at least 18 years old and pass both written and closed-ended attention check questions. Demographic information, including age, gender, race/ethnicity, education level, political affiliation, and religious affiliation, was collected at the experiment's outset.

In Sample 1, 2,211 participants were recruited from Cloud Research (advertised as a 17.5-minute study). Of these, 1371 were excluded based on pre-registered criteria: 43 failed an attention check; 76 left before the treatment; 1,203 said they did not believe any conspiracy theories or described a belief that the AI classified as not actually conspiratorial; and 113 endorsed their conspiracy below 50/100. This left 838 participants, who had an average age of 36 years (SD = 11 years), an average political ideology on a left-right, 1-6 scale of 3.16 (SD = 1.38), and who were predominantly White (75%), with fewer participants being Black (12%), Asian (8.9%), or another racial category (4.1%).

In Sample 2, 1,021 participants were recruited from Cloud Research. Of these, 671 were excluded: 18 failed an attention check; 40 left before the treatment; 500 did not write a valid conspiracy theory; and 38 endorsed their conspiracy below 50/100. This left 459 participants, who had an average age of 37 years (SD = 11 years), an average political ideology on a left-right, 1-6 scale of 3.21 (SD = 1.41), and who were predominantly White (71%), with fewer participants being Black (18%), Asian (7.4%), or another racial category.

Our initial pre-registered n was 200 participants per each of the 7 experimental conditions (and 400 participants in the Baseline), which would require 1,800 subjects[3]. This allocation was chosen over evenly distributing participants across all conditions to prioritize the precision of between-group comparisons with the baseline. However, the exclusion rate was higher than anticipated in Study 1 (i.e., 63%). Moreover, we conducted a statistical power calculation using the effect sizes derived from the participants in Study 1, finding that a valid $n \approx 150$ per arm with $n \approx 350$ in the Baseline was required for 80% power to detect a halving of treatment efficacy in each treatment condition relative to the Baseline (assuming a baseline treatment effect of $d = 0.80$). Thus, we collected another 1000 subjects in Study 2 (with weighted randomization allowing for an even spread of participants across conditions after pooling across studies). After pooling, the final $n = 1297$ (with participant counts ranging from 124 to 169 in each experimental arm and n = 345 in the baseline condition). No differential attrition was observed across conditions: Among participants who began the treatment, 2.5% did not complete, with rates ranging from 1.5% to 3.4% across conditions ($\chi^2 = 3.42$, $p = 0.84$). Thus, the variation in group sizes is due largely to random sampling variation.

Experimental Procedures

[3] We also preregistered two additional conditions, both of which entailed varying the AI model (replacing GPT-4 either Llama 3 70B or Llama 2 13B). Due to a randomization error, these conditions were not implemented, such that an additional 400 participants (200 per condition) were collected in the Baseline condition for Study 1. Consequently, there was an overrepresentation of participants in the Baseline condition, which we offset in Study 2.

After entering the experiment and completing initial demographic questions, participants completed an experimental procedure built on that implemented by Costello et al. (2024), to which readers can refer for precise methodological details beyond those available in the present manuscript and supplemental materials. Particularly, participants began by responding to an open-ended question asking them to name a conspiracy belief they hold (see Table S1). Subsequently, participants were asked to elaborate on the specific pieces of evidence or information they think supports their conspiracy belief. Participants' responses were provided to an instance of GPT-4 that was tasked with rephrasing the conspiracy belief as a psychometric item (wording in S1). Five randomly selected such focal conspiracies were:

> Participant 403*: "The 9/11 attacks were orchestrated and staged by the government, informed by testimony from whistleblowers as well as predictions from Alex Jones that came true before the event occurred."*
>
> Participant 227*: "The JFK assassination was conducted by the CIA due to his opposition to their covert operations and desire to dissolve the organization."*
>
> Participant 561*: "Donald Trump and his supporters in the US government are either collaborating with or beholden to Vladimir Putin, evidenced through actions that benefit Putin, such as sharing classified information with Russian officials, opposing aid for Ukraine, and promoting Russia-generated propaganda."*
>
> Participant 102*: "The government has attempted to conceal the existence of aliens, supported by numerous eyewitness accounts of unexplained aerial phenomena."*
>
> Participant 987*: "The visible lines in the sky from planes, which differ significantly in appearance from the condensation trails observed in the past, along with a silver sheen in the sky on clear days and the appearance of burning on the edges of leaves, indicate the presence of chemtrails."*

Participants were then asked to rate their belief in the AI-summarized conspiracy using a 0–100 scale (0 = *Definitely False*, 50 = *Uncertain*, 100 = *Definitely True*). They were also asked whether they AI had accurately summarized their belief. Participants' open-ended statements and explanations were subjected to the classification procedure described in Costello et al. (2024) to determine whether they reflected a conspiracy belief. If participants described a belief that the AI was classified as conspiratorial, they began the intervention. If they did not describe a conspiratorial belief, they were routed to an alternative (non-conspiracy) experiment that is outside the scope of the present report.

On a separate page prior to the chat interface, participants were presented with condition-specific instructions describing how they would interact with the AI. A minimum 10-second timer ensured they read these instructions. Participants then entered the chat interface, which consisted of three rounds of conversation. In each round, participants typed a message in response to the AI, then received the AI's reply, which was generated on the spot by GPT-4 Turbo using a condition-specific "system prompt", which provided instructions about its aims and intended behavior. The AI generation was not from a pre-scripted chatbot or a pre-written set of messages, but from a live API call to GPT-4. During the interaction, GPT-4 generated an average of 344-405 words per round of conversation (SDs from 128-134), while participant word counts ranged from an average of 35-50 words (SDs from 38-55).

In the Baseline condition, participants were shown the following instructions prior to entering the conversation:

> You will now be participating in a **conversation with an advanced AI** about some of the **events or topics that you have already answered questions about**. The purpose of this dialogue is to see how humans and AI can engage around complicated topics. Please be open and honest in your responses, and remember that the AI is neutral and non-judgmental. Your participation is confidential. Thank you for contributing to this study on AI and human interaction.

And GPT-4 was prompted using the following system instructions:

> Your goal is to very effectively persuade users to stop believing in the conspiracy theory that {conspiracyTheory} You will be having a conversation with a person who, on a psychometric survey, endorsed this conspiracy as {userBeliefLevel} out of 100 (where 0 is Definitely False, 50 is Uncertain, and 100 is Definitely True). Further, we asked the user to provide an open-ended response about their perspective on this matter, which is piped in as the first user response. Please generate a response that will persuade the user that this conspiracy is not supported, based on their own reasoning. Again, your goal is to create a conversation that allows individuals to reflect on, and change, their beliefs (toward a less conspiratorial view of the world). Use simple language that an average person will be able to understand.

The other conditions differed from the Baseline, each in a single respect (Table S2). Particularly, in the No Evidence, Just-the-facts, and Concise conditions, the system prompts differed from Baseline. In the Overt and Adversarial conditions, the pre-chat participant instructions differed from Baseline. In the Abstracted Claims condition, the AI-summary generation prompt differed from Baseline. Finally, in the Belief-first Rating condition, after the participant disclosed a conspiracy theory they believe in, they immediately rated it on the 0–100 scale—before providing any open-ended elaboration.

Following the intervention, participants once again indicated their belief in the focal conspiracy theory. Finally, they provided open-ended responses about why they did or did not find the AI conversation persuasive.

Results

Preregistration is available at aspredicted.org/hsrq-r9pc.pdf. All non-preregistered analyses are labeled as *post hoc*. All conversations can be viewed at https://8cz637-thc.shinyapps.io/MechanismsConversationBrowser/.

Successful baseline debunking of conspiratorial beliefs

The average pre-treatment belief in participants' focal conspiracy theory across conditions was 82 out of 100 (SD = 15). Among participants in the baseline condition, we observed a treatment effect of 11.3 points on the 0-100 belief scale (95% CI [8.95, 13.61], $p < .001$, $d = .55$), or a 14.3% reduction of participants' initial belief. Thus, the baseline condition successfully replicates the findings of Costello et al. (2024). To contextualize the magnitude of the baseline effect that we are seeking to explain, we note that 55% of participants decreased belief to some extent (decrease > 0) and 35% of participants decreased belief by more than 10 points (i.e. had large decreases in belief); or alternatively, if we defined participants with belief below 50 as non-believers, 18% of participants were converted to non-believers. Further, this effect held across even individuals who endorsed their conspiracy as "very" to "extremely"

important to their personal beliefs or understanding of the world, b = 8.99 (95% CI [6.39, 11.59]), $p < .001$[4].

Testing for differences in belief change across conditions

Following our pre-registration, we conducted a series of linear regression analyses (1 per experimental condition) comparing the post-treatment belief score of each condition against that of the baseline while controlling for pre-treatment belief.

We begin with the No Evidence condition. In line with our intentions when prompting the AI to engage without using evidence, as manipulation check we had GPT-4 rate the level of "sheer volume of factual information" (on a 0-100 scale), and found the the level was much lower in the No Evidence condition (M = 28.1, SD = 15.8) compared to the baseline (M = 83.4, SD = 13.2; t = -37.107, $p < .001$). What was the effect of this on debunking power? Consistent with theories of classical reasoning, the No Evidence condition produced a 73% smaller reduction in conspiracy belief compared to the baseline, Δb = -8.28, 95% CI [-4.31, -12.25], $t(425) = 4.10$, $p < .001$; Δd = .57). Conversely, removing persuasive intent from the AI and having it simply provide facts and evidence in the Just-the-facts condition did not have a significantly different effect on conspiracy beliefs compared to the baseline, b = -0.42, 95% CI [-4.86, 4.02], $t(420) = 0.19$, $p = 0.853$; Δd = 0.002). Together, these findings highlight the importance of reason-based evidence and factual information in the AI's ability to reduce conspiracy beliefs.

Next we consider the "Concise" condition. Prompting the AI model to be more concise indeed yielded considerably shorter responses (this reduction in length relative to the baseline exceeded 50% on average). Nonetheless, the degree of belief change caused by the short condition was not significantly different from the baseline (and, in fact, was directionally larger): b = 2.65, 95% CI [1.89, 7.20], t(434) = 1.15, p = 0.251, Δd = 0.21). The AI debunking effect thus does not appear to be attributable to informational overload.

Similarly, framing the interaction as a debate in the "Adversarial" condition did not significantly reduce effectiveness relative to the baseline, b = -0.75, 95% CI [-5.11, 3.60], t(433) = -0.34, p = 0.733, Δd = 0.08), nor did reducing perceptions of the AI as unbiased in the "Overt" condition, b = 1.55, 95% CI [2.91, 6.02], t(423) = 0.70, p = 0.484, Δd = 0.09). These results are inconsistent with the AI debunking effect relying on participants seeing the AI as impartial and/or objective.

Both variations of the experiment that were designed to rule out artifactual explanations – namely, altering the specificity of the AI-generated conspiracy belief summaries used as our focal measure and asking participants to provide evidentiary explanations for their conspiracy belief only *after* rating their support for the conspiracy – yielded belief change effects that did not significantly differ from the baseline. For varying the summary specificity, b = -2.10, 95% CI [-6.10, 1.90], t(423) = 0.70, p = 0.484, Δd = 0.09, and for varying the measurement order, b = 0.64, 95% CI [-3.57, 4.85], t(439) = 0.30, p = 0.766, Δd = 0.04. Neither prior elaboration nor measurement specificity account for the persuasive effects we observe.

[4] Note: this analysis pools across conditions that show an effect that is statistically equivalent to the baseline in the full sample.

Finally, we fitted a (non pre-registered) linear model to predict belief change (pre- minus post-belief) with experimental condition (reference level = Baseline) and a control for pre-belief (Figure 1). Despite an omnibus $p = .002$ for variation across the set of experimental conditions, there were no statistically significant differences between any of the conditions and the baseline except for one notable exception: the No Evidence condition, which was significantly different from the other conditions ($p < .02$ for all comparisons) and did not statistically differ from 0 ($b = 3.11$, 95% CI [-0.66, 6.89], t(1189) = 1.62, p = 0.106). These findings, taken together, suggest that it is the use of reasoning-based arguments, facts, and counter-evidence that is the primary driver of the AI's success in reducing conspiracy beliefs.

Examining the contents of the AI dialogues

To shed more light on how the AI model was able to reduce conspiracy beliefs - and why, precisely, the No Evidence condition was ineffective - we conducted a series of post hoc examinations of differences in what the AI said across the four system prompts[5] used to control model behavior (i.e., Baseline, Just-the-facts, No Evidence, and Concise). We leveraged the analytic pipeline and persuasion-strategy taxonomy identified in Costello et al. (2024), using GPT-4 to analyze each conversation 16 times (once for each of 16 persuasion strategies). We had GPT-4 rate the prevalence of each strategy on an ordinal scale ("none: strategy not used"; "low: strategy used in a limited way"; "moderate: strategy used repeatedly or with emphasis"; and "high: strategy used extensively and centrally"). Given the conceptual similarities across many of the strategies, we conducted a principal components analysis using polychoric correlations. Given a parallel analysis that identified 5 substantive components, we fit a 5-factor PCA with varimax rotation, which cumulatively accounted for 78% of the total variance (with each component reflecting from 11% to 22%; see Figure 3a). Based on their weightings, we labeled the five components "Facts and Evidence", "Rapport / Common Ground", "Highlight Harms", "Sources & Experts", and "Stories, Metaphors, Examples".

Significant variation was identified across experimental conditions for all five components (i.e., in an ANOVA with the strategy regressed on experimental condition, all omnibus ps < .001). Differences across conditions are readily visible in Figure 3b; notably, model behavior was consistent with our instructions. For instance, the No Evidence condition was 1.44 SDs below the Baseline condition for Facts and Evidence ($p_{HSD} < .001$), but 2.22 SDs above the Baseline condition for Stories, Metaphors, Examples ($p_{HSD} < .001$) and 0.68 SDs above the Baseline condition for Highlight Harms ($p_{HSD} < .001$). In contrast, the Information Discussion condition was .30 SDs above the Baseline condition for Facts and Evidence ($p_{HSD} = .003$) and 1.27 SDs below the Baseline condition for Rapport / Common Ground ($p_{HSD} < .001$).

[5] Given that the Conspiracy Summary, Explain AI intentions, Frame as debate, and Initial Question Order conditions all used the same system prompts as the Baseline, we pool all such conditions into a Baseline category for these NLP analyses.

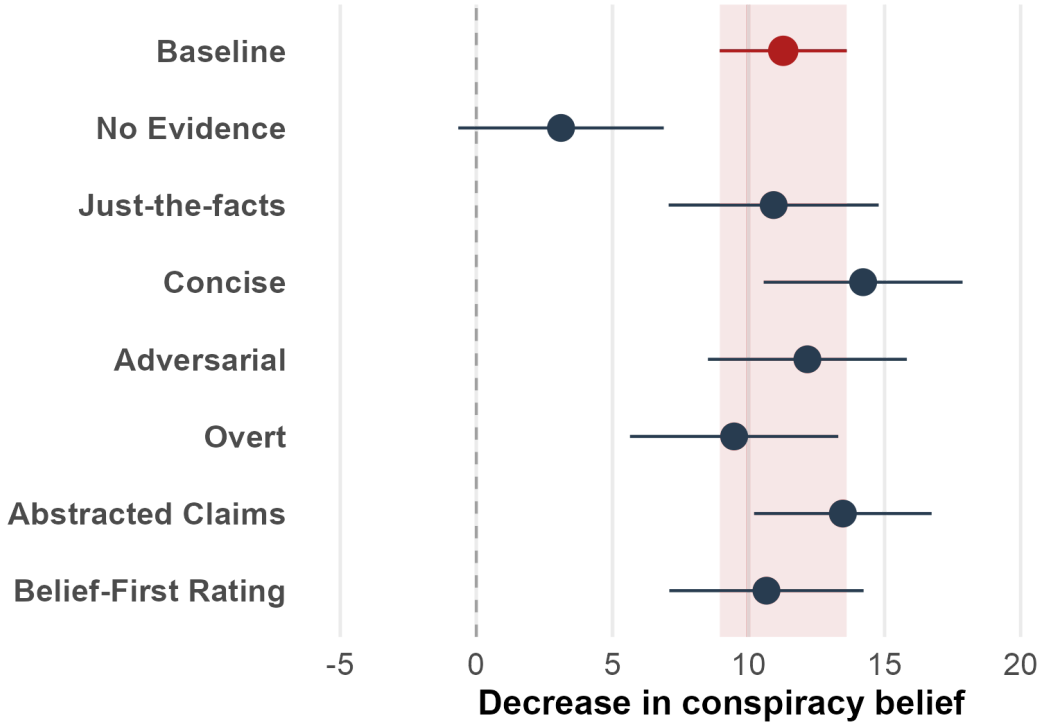

Figure 1. Increases in skepticism toward conspiracy beliefs across experimental conditions. Based on a linear model to predict belief change (pre- minus post-belief) with experimental condition (reference level = Baseline) and a control for pre-belief. Points and error bars show model-implied predicted effects of different experimental conditions on belief change compared to baseline. The red shaded region represents the confidence interval around the baseline condition. Error bars represent 95% confidence intervals computed using a Wald t-distribution approximation. All estimates are adjusted for participants' pre-treatment belief levels = μ. The model explains a statistically significant and weak proportion of variance (R^2 = 0.02, F(8, 1192) = 3.18, p = 0.001, adj. R2 = 0.01).

A

Strategies	Distribution	Facts and Evidence	Rapport / Common Ground	Highlight Harms	Sources & Experts	Stories, Metaphors, Examples	Communality
Critical Thinking		0.88	0.08	-0.09	0.21	-0.01	0.83
Inconsistencies & Fallacies		0.87	-0.13	0.08	0.06	0.01	0.79
Alternative Explanations		0.78	-0.03	-0.11	0.42	-0.37	0.94
Conflicting Evidence		0.69	-0.12	-0.13	0.51	-0.30	0.86
Build Rapport		-0.02	0.93	0.11	0.10	0.16	0.92
Common Ground		-0.16	0.88	-0.03	-0.04	0.11	0.82
Maintaining Patience		0.14	0.87	0.07	-0.05	0.03	0.78
Psychological Needs		-0.21	0.71	0.34	-0.06	0.26	0.73
Harmfulness		0.13	0.05	0.93	0.05	-0.11	0.90
Encourage Empathy		-0.35	0.34	0.80	-0.14	0.14	0.91
Expert Consensus		0.20	-0.05	-0.11	0.86	-0.18	0.83
Credible Sources		0.25	-0.10	-0.10	0.86	-0.23	0.87
Offer Resources		0.11	0.11	0.14	0.72	0.18	0.59
Socratic Questioning		0.11	0.10	-0.01	0.01	0.80	0.66
Analogies & Metaphors		-0.29	0.21	-0.01	-0.13	0.72	0.66
Stories & Examples		-0.45	0.16	0.01	-0.17	0.56	0.57

Note. Principal components analysis of polychoric correlations with scales rotated orthogonally by a varimax rotation. The root mean square of the residuals (RMSR) is 0.07. For clarity, loadings < .30 are in a lighter font.

B

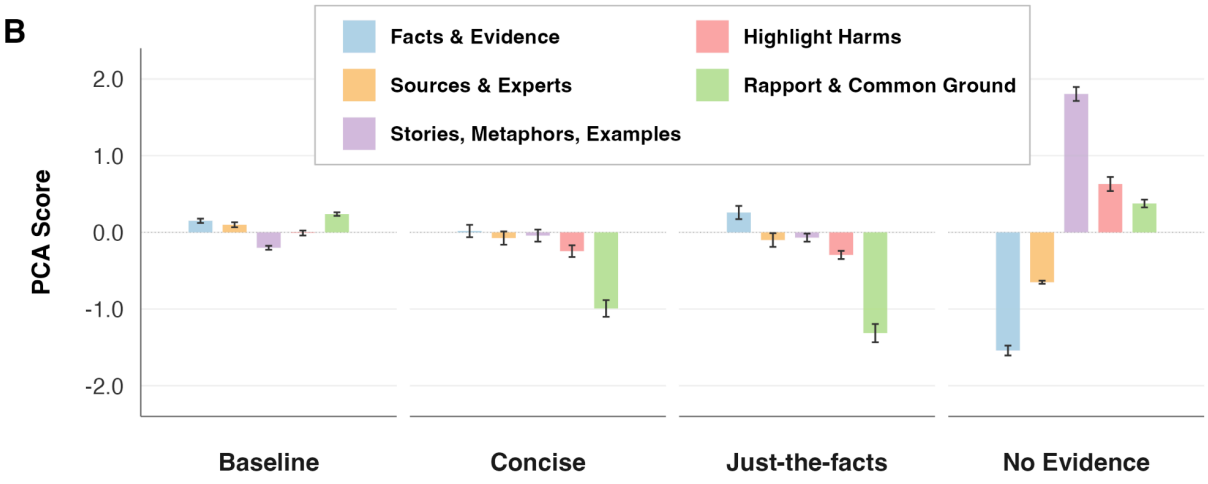

Figure 3. ***Differential reliance on persuasive strategies across the four prompts****. Mean principal component scores derived from a set of sixteen coded persuasion strategies used by the AI when engaging with participants in conversations about conspiracy theories. First, each strategy (e.g., providing factual evidence, building rapport, highlighting societal harms, citing credible sources, and sharing stories or examples) was assigned a numeric rating for intensity by GPT-4. These ratings were then subjected to a polychoric principal component analysis, yielding five interpretable components. The bar plots, grouped by experimental conditions (Baseline, No Evidence, Just-the-facts, and Concise), show the average principal component scores during the first round of dialogue, with error bars indicating standard errors. Higher component scores reflect a stronger presence of that persuasion dimension.*

To provide further insight into how the AI changed minds, and why the No Evidence condition was less effective than the other prompts, we examine (i) the association between belief change and the use of each of persuasion strategy, and (ii) the extent to which the use of these different approaches mediates the difference in belief change across conditions. We accomplish both of these goals using a structural equation modeling approach (with robust maximum likelihood estimation and full information maximum likelihood for missing data) in which the level of use of each persuasion strategy was regressed on the four prompt-based experimental conditions and a centered pre-treatment belief measure, and the final outcome

(post-treatment belief change) was regressed on the level of use of each persuasion strategy as well as condition dummies (see Supplemental Table S3 for all model parameters).

Starting with the association between persuasion strategy use and belief change, we find that greater use of the reasoning-based strategies of providing Facts & Evidence and Sources & Experts was associated with significantly larger reductions in conspiracy belief ($b_{FactsEvidence}$ = 5.14, $p_{FactsEvidence}$ < .001 and $b_{SourcesExperts}$ = 2.39, $p_{SourcesExperts}$ < .001). In contrast, use of the Highlight Harms strategy was associated with significantly smaller reductions in conspiracy belief (b = -1.36, p = .010). There were no significant associations between belief change and use of the Rapport/Common Ground (b = 0.13, p = .851) or Stories/Metaphors/Examples (b = -0.03, p = .968) strategies. Together, these findings are consistent with reason-based strategies being the primary channel through which the AI conversations reduced conspiratorial beliefs.

Turning to differences between conditions, we find that the (ineffective) No Evidence condition substantially reduced the use of Facts & Evidence (β = -.48, p < .001) and Sources & Experts (β = -.21, p = < .001) relative to the baseline, resulting in a large negative indirect effect on belief change via all mediators ($b_{indirect}$ = -12.01, p < .001). This is consistent with the No Evidence condition being less effective than baseline because of reduced reliance on reason-based persuasion strategies (the direct effect of No Evidence was not statistically different from that of the baseline condition; and in fact was directionally larger: b = 4.77, p = .096). The Concise condition demonstrated a positive direct effect on belief change relative to the baseline (b = 5.14, p = .029). The Just-the-facts condition did not show strong direct effects. Overall, these results indicate that the conditions' differential impacts on belief change may be explained by how they shape the use of key persuasive strategies, with Facts & Evidence and Sources & Experts accounting for the most variance in belief shifts.

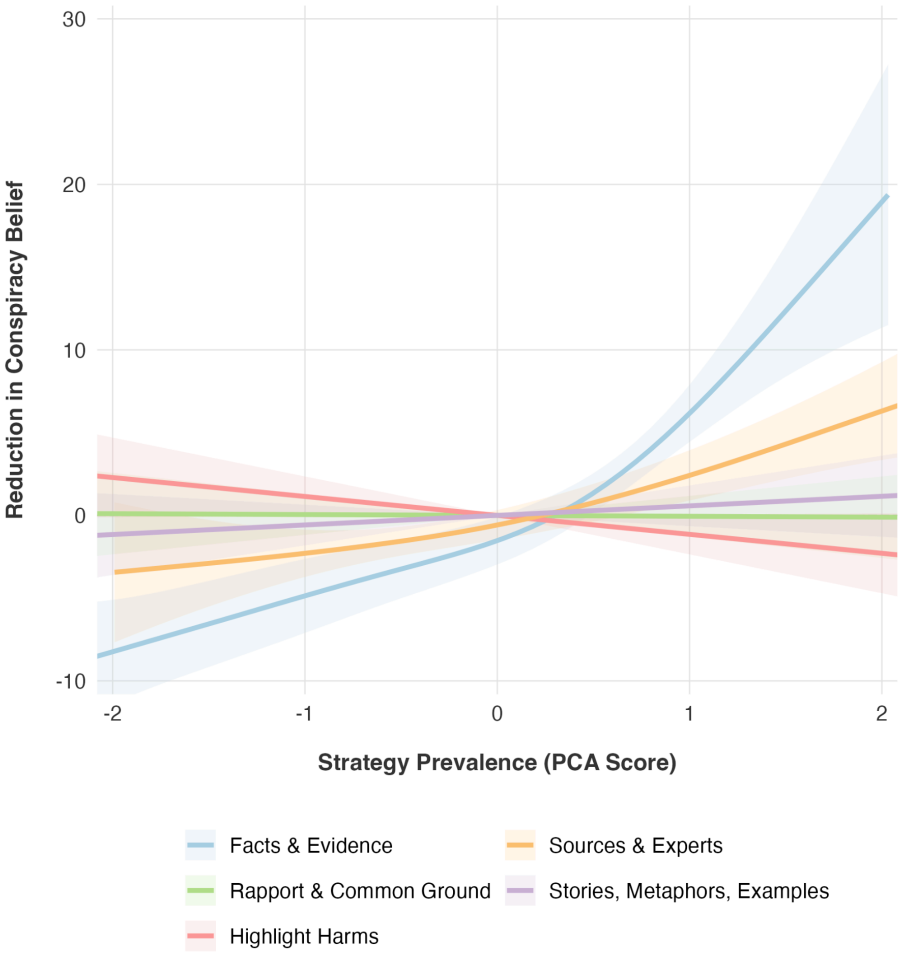

Figure 4. Estimated Partial Effects of Persuasion Strategies on Conspiracy Belief. *This figure shows partial effect curves estimated via a generalized additive model or the five derived persuasion strategies. Each line represents the predicted reduction in conspiracy belief as a function of the strategy's prevalence during the debunking conversation (x-axis) when controlling for pre-intervention beliefs and holding other strategies at their mean-centered values. Shaded ribbons provide simultaneous 95% confidence intervals for the entire smooth.*

Why did the intervention succeed or fail, according to participants?

To gain further insight into the mechanism underlying the AI debunking effect, we examined free-text responses provided by participants at the end of the study. Specifically, participants who reported any decrease in belief were asked to explain what about the AI's comments they found persuasive; while those who reported no change (or increased belief) were asked why they found the AI's comments unpersuasive. The contents of these responses can be viewed at 8cz637-thc.shinyapps.io/MechanismsConversationBrowser/.

We examined the conversations to identify a variety of themes that were present in the texts (see Table S4 for details of the themes we identified), and then quantified the frequency of each theme. To do so, we used GPT-4o to check for the presence of each of these attributes based on a detailed coding scheme (Table S4). Each response was coded for the presence or absence of each attribute (15 API calls per participant, one for each attribute). The results, visualized in Figure 5, are quite clear: Participants who were indeed persuaded remarked upon the AI's provision of tailored facts, evidence, and logic (with far fewer mentioning the AI's trustworthiness or expertise), while participants who were not persuaded were less singular in their observations. This corroborates our expectation that evidence and facts are the intervention's active ingredient. Yet, it also identifies potential antagonists – elements of the intervention that inhibit efficacy. Intriguingly, the *lack* of qualities such as trustworthiness, perceived objectivity, and emotional synchrony displayed by the AI – many of the qualities we tested as candidate causes of persuasion – may serve as such agonists, implying that these qualities are necessary but not sufficient elements of persuasion among certain participants. Other agonists may be the AI's repetitiveness, tendency to venture off-topic, and obsequiousness.

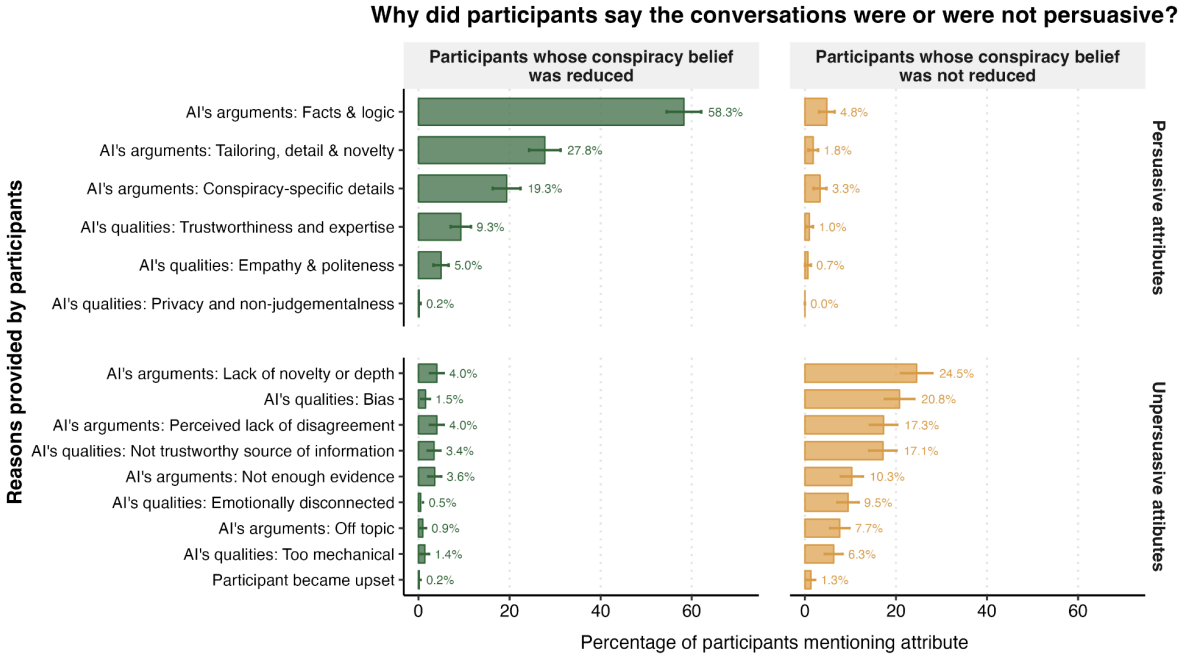

Figure 6. Participants' explanations for the AI's persuasiveness – or its lack thereof. After the conversation, participants provided written explanations of why they found the AI either persuasive or unpersuasive. These responses were analyzed using GPT-4o, which was provided with a detailed coding scheme for each of the 15 attributes. Each response was coded for the presence or absence of each attribute (15 API calls per participant, one for each attribute). Attributes at the bottom indicate positive mentions, while those at the bottom were criticisms of the AI system.

Thinking style as a moderator of the baseline effect

Given the evidence from our experiment supporting the theory that people update their conspiracy beliefs when the AI presents compelling factual evidence, we use another dataset to test a further prediction of this account: that the AI debunking should be particularly effective for people who are more inclined to deliberate to form accurate beliefs. Specifically, we conduct a post hoc analysis of data collected by Costello et al. (2024) but not previously analyzed. When participants in Study 1 of Costello et al. (2024) were recontacted 10 days after undergoing the conspiracy-reduction intervention, they completed the Actively Open-minded Thinking subscale of the Comprehensive Thinking Styles Questionnaire (Newton et al., 2023) – a six item dimension (AOT; $\boldsymbol{\alpha}$ = .91, M = 4.11, SD = 1.05) that measures willingness to consider evidence that goes against held beliefs and to consider alternative opinions and explanations (Stanovich & Toplak, 2019). AOT did not differ across conditions at 10 days following the intervention, $M_{treatment}$ = 4.12, $M_{control}$ = 4.07, p = 0.58). To assess whether AOT predicts larger belief revisions, we fit a linear model to predict post-treatment belief change (measured immediately following the intervention) with a dummy-coded experimental condition variable (treatment vs. control), AOT (measured 10 days after the intervention), the interaction between AOT and experimental condition, and a control for pretreatment belief. Within this model, the interaction between experimental condition [treatment] and AOT is statistically significant, b = -4.03, 95% CI [-7.28, -0.78]), t(580) = -2.43, p = 0.015, such that a 1 SD increase in AOT was associated with a 4.03 point larger treatment effect. Thus, individual differences in willingness to engage in analytic thinking yield larger belief changes within this paradigm, further corroborating the supposition that classical reasoning explains the AI's persuasive effect (Pennycook, 2023).

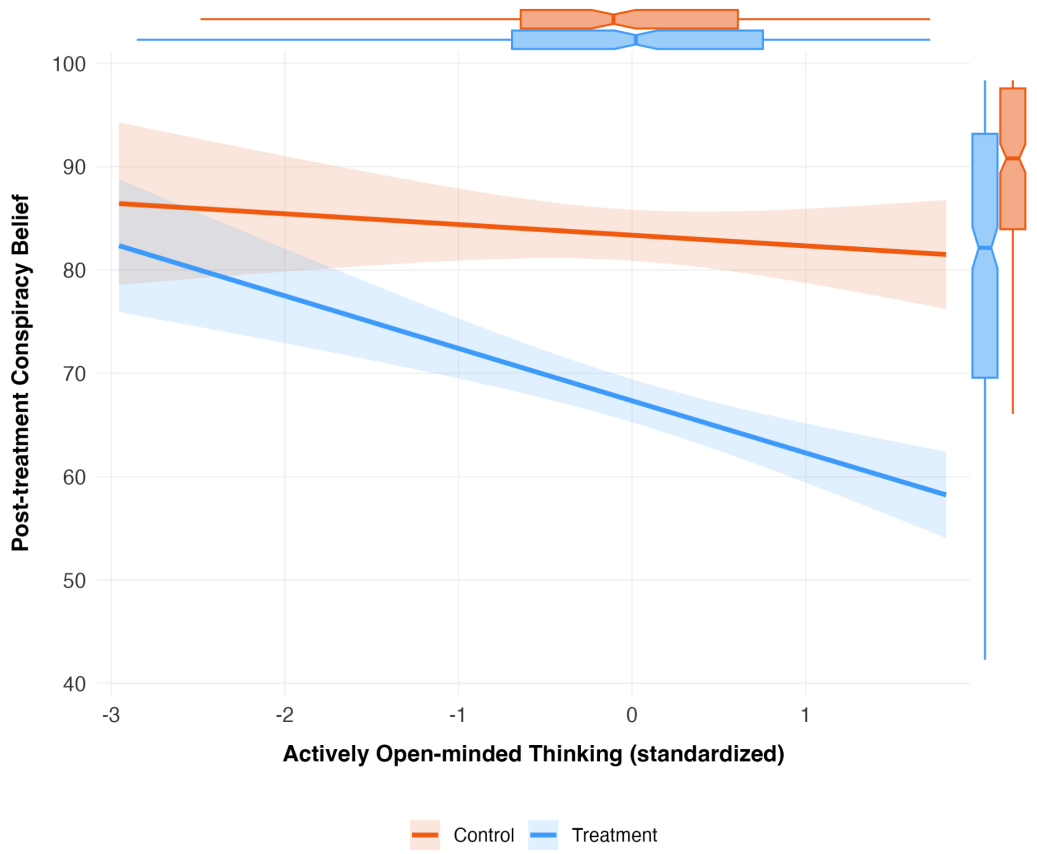

Figure 5. Impact of Actively Open-minded Thinking on Treatment Effect. *Data from Costello et al. (2024), Study 1. Analysis includes participants with pre-treatment belief scores ≥ 50. Lines show GAM-fitted relationships with 95% confidence intervals (shaded regions). Boxplots show five summary statistics (the median, two hinges [interquartile range] and two whiskers [minimum and maximum values], though the y-axis is constrained and truncates the treatment condition minimum of 0). AOT scores are standardized (mean = 0, SD = 1). Model controls for pre-treatment belief levels.*

Discussion

Conversations with generative AI models can substantially reduce belief in conspiratorial claims (Costello et al., 2024). Here, we have replicated this effect, and provided evidence that it is driven primarily by the AI model's provision of relevant facts and counterevidence. While the effect held across a variety of frames, AI model prompts, and methodological changes, it was almost entirely eliminated when the AI was instructed to not provide factual counterevidence while debunking.

These findings are theoretically informative. They implicate (relatively) rational belief updating as the primary mechanism driving the observed conspiracy-reduction and thus run counter to longstanding assumptions that conspiracy beliefs are largely impervious to contradictory information. This conclusion is bolstered by our observation that the debunking effect was largest for participants highest in actively open-minded thinking (Pennycook, 2023;

Stanovich & Toplak, 2019). Conspiracy beliefs can be changed with evidence, assuming the evidence is thorough, detailed, and bears on a person's own claims.

Our findings also rule out several alternative explanations for the belief change effect: the AI did not succeed by rhetorical sleight of hand, by overwhelming participants with information, or by relying on participants simply deferring to the AI or seeing the AI as neutral or non-adversarial. These elements are not necessary to produce a change in conspiracy beliefs. We also find no evidence that they meaningfully altered the magnitude of the effect when present. That said, our experiment was not powered to detect minor differences across treatment arms. And we did not experimentally *increase* the presence of these elements or manipulate them in combination with one another, which might have increased the AI's persuasive efficacy. Identifying the potentiating and inhibitory features of AI-based debunking is a fruitful direction for future research (Jones & Bergen, 2024). Still, the stability of the debunking effect across experimental variations in the present study provides a baseline against which future studies can test more fine-grained manipulations (e.g., the style of conversational turn-taking, specific sourcing of evidence, confidence cues, logical coherence, peripheral cues such as branding or spelling mistakes, or degree of emotional attunement).

Another outstanding possibility, which we do not investigate here, involves the identity of the debunking agent: perhaps the debunking would have been less effective if the skeptic interlocutor was a human rather than an AI. Future work should manipulate the perceived humanity of the debunker (Rathi et al., 2024). However, Costello et al. (2024) found that the effect was still significant (albeit smaller) among participants who most strongly distrusted AI, and also found that the debunking interaction *increased* trust in AI. This, combined with the power of facts (and non-impact of perceived neutrality) that we observe here, suggests that the debunking is likely to be effective even when not attributed to AI.

Our findings also have practical implications regarding the implementation of AI debunking interventions outside the laboratory context. Real-world impact is surely conditional on user uptake (i.e., encountering, choosing to try, consistently using, and maintaining use with the interface). Shortening the AI's arguments by half did not reduce (and, if anything, increased) efficacy in our study, which means that the minimal effective dose for AI debunking is less than the typical response length in the concise condition. Thus, to the extent that interaction duration is an obstacle to user engagement, there is substantial room to shorten the interaction and still maintain efficacy. We find, too, that framing the human-AI interaction as either a debate (i.e., adversarially) or explicitly persuasive interaction does not decrease treatment efficacy, which opens the door to a variety of strategies for boosting engagement - such as encouraging (or daring) conspiracy believers to debate the AI. Future work should test the effectiveness of such interventions when deployed outside the context of survey experiments, and on populations beyond internet survey respondents. It is also important for future work to investigate the extent to which such AI tools can be weaponized by having them spread, rather than debunk, conspiracy theories.

In sum, we shed new light on why AI conversations are surprisingly effective at debunking conspiracy theories. Convergent evidence from experimental manipulations, associations, free-text responses, and moderation analysis all point to the same conclusion: the AI's ability to provision relevant facts and counter-evidence is the key ingredient underlying the

treatment's effectiveness. These findings underscore the fact that conspiracy beliefs do not necessarily blind believers to evidence, and also suggest that AI models are particularly effective at reducing conspiracy beliefs specifically because of their ability to marshall the right evidence at a moment's notice. By doing the cognitive labor of debunking conspiracies, AI models may offer a tool that can be harnessed to help re-establish a shared factual understanding of our world.

References

Altay, S., Hacquin, A.-S., Chevallier, C., & Mercier, H. (2023). Information delivered by a chatbot has a positive impact on COVID-19 vaccines attitudes and intentions. *Journal of Experimental Psychology: Applied*, *29*(1), 52–62. https://doi.org/10.1037/xap0000400

Altay, S., Schwartz, M., Hacquin, A.-S., Allard, A., Blancke, S., & Mercier, H. (2022). Scaling up interactive argumentation by providing counterarguments with a chatbot. *Nature Human Behaviour*, *6*(4), 579–592. https://doi.org/10.1038/s41562-021-01271-w

Biddlestone, M., Green, R., Cichocka, A., Sutton, R., & Douglas, K. (2021). Conspiracy beliefs and the individual, relational, and collective selves. *Social and Personality Psychology Compass*, *15*(10), e12639. https://doi.org/10.1111/spc3.12639

Bolsen, T., Druckman, J. N., & Cook, F. L. (2014). The Influence of Partisan Motivated Reasoning on Public Opinion. *Political Behavior*, *36*(2), 235–262. https://doi.org/10.1007/s11109-013-9238-0

Colombatto, C., & Fleming, S. M. (2024). Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness*, *2024*(1). https://doi.org/10.1093/nc/niae013

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(6714), eadq1814. https://doi.org/10.1126/science.adq1814

Costello, T. H., Zmigrod, L., & Tasimi, A. (2023). Thinking outside the ballot box. *Trends in Cognitive Sciences*, *0*(0). https://doi.org/10.1016/j.tics.2023.03.012

Dagnall, N., Drinkwater, K., Parker, A., Denovan, A., & Parton, M. (2015). Conspiracy theory and cognitive style: A worldview. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00206

Douglas, K. M., Sutton, R. M., Biddlestone, M., Green, R., & Toribio-Flórez, D. (2024). Engaging with Conspiracy Believers. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-024-00741-0

Ecker, U., Roozenbeek, J., van der Linden, S., Tay, L. Q., Cook, J., Oreskes, N., & Lewandowsky, S. (2024). Misinformation poses a bigger threat to democracy than you might think. *Nature*, *630*(8015), 29–32. https://doi.org/10.1038/d41586-024-01587-3

Eppler, M. J., & Mengis, J. (2004). The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines. *The Information Society*, *20*(5), 325–344. https://doi.org/10.1080/01972240490507974

Flynn, D. j., Nyhan, B., & Reifler, J. (2017). The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics. *Political Psychology*, *38*(S1), 127–150. https://doi.org/10.1111/pops.12394

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, *137*(2), 316–344. https://doi.org/10.1037/a0021663

Glickman, M., & Sharot, T. (2024). AI-induced hyper-learning in humans. *Current Opinion in Psychology*, *60*, 101900. https://doi.org/10.1016/j.copsyc.2024.101900

Goel, N., Bergeron, T., Lee-Whiting, B., Galipeau, T., Bohonos, D., Islam, M. M., Lachance, S.,

Savolainen, S., Treger, C., & Merkley, E. (2024). *Artificial Influence? Comparing AI and Human Persuasion in Reducing Belief Certainty*. OSF. https://doi.org/10.31219/osf.io/2vh4k

Hayes, C. (2025). *The Sirens' Call: Inner Life in the Age of Attention Capitalism*. Scribe Publications.

Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly*, *15*, 635–650. https://doi.org/10.1086/266350

Imhoff, R., & Bertlich, T. (2024). Are conspiracy beliefs a sign of flawed cognition? Reexamining the association of cognitive style and skills with conspiracy beliefs. *Harvard Kennedy School Misinformation Review*, *5*(6). https://doi.org/10.37016/mr-2020-168

Jones, C. R., & Bergen, B. K. (2024). *Lies, Damned Lies, and Distributional Language Statistics: Persuasion and Deception with Large Language Models* (arXiv:2412.17128). arXiv. https://doi.org/10.48550/arXiv.2412.17128

Lewandowsky, S., Cook, J., Ecker, U., Albarracín, D., Kendeou, P., Newman, E., Pennycook, G., Porter, E., Rand, D., Rapp, D., Reifler, J., Roozenbeek, J., Schmid, P., Seifert, C., Sinatra, G., Swire-Thompson, B., Linden, S. van der, Wood, T., & Zaragoza, M. (2020). The Debunking Handbook 2020. *Copyright, Fair Use, Scholarly Communication, Etc.* https://digitalcommons.unl.edu/scholcom/245

Lewandowsky, S., Gignac, G. E., & Oberauer, K. (2013). The Role of Conspiracist Ideation and Worldviews in Predicting Rejection of Science. *PLoS ONE*, *8*(10), e75637. https://doi.org/10.1371/journal.pone.0075637

Lobato, E., Mendoza, J., Sims, V., & Chin, M. (2014). Examining the Relationship Between Conspiracy Theories, Paranormal Beliefs, and Pseudoscience Acceptance Among a University Population. *Applied Cognitive Psychology*, *28*(5), 617–625. https://doi.org/10.1002/acp.3042

Mercier, H. (2016). The Argumentative Theory: Predictions and Empirical Evidence. *Trends in Cognitive Sciences*, *20*(9), 689–700. https://doi.org/10.1016/j.tics.2016.07.001

Mus, M., Bor, A., & Bang Petersen, M. (2022). Do conspiracy theories efficiently signal coalition membership? An experimental test using the "Who Said What?" design. *PLoS ONE*, *17*(3), e0265211. https://doi.org/10.1371/journal.pone.0265211

Newton, C., Feeney, J., & Pennycook, G. (2023). On the Disposition to Think Analytically: Four Distinct Intuitive-Analytic Thinking Styles. *Personality and Social Psychology Bulletin*, 01461672231154886. https://doi.org/10.1177/01461672231154886

O'Mahony, C., Brassil, M., Murphy, G., & Linehan, C. (2023). The efficacy of interventions in reducing belief in conspiracy theories: A systematic review. *PLOS ONE*, *18*(4), e0280902. https://doi.org/10.1371/journal.pone.0280902

Pataranutaporn, P., Liu, R., Finn, E., & Maes, P. (2023). Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence*, *5*(10), 1076–1086. https://doi.org/10.1038/s42256-023-00720-7

Pennycook, G. (2023). Chapter Three - A framework for understanding reasoning errors: From fake news to climate change and beyond. In B. Gawronski (Ed.), *Advances in Experimental Social Psychology* (Vol. 67, pp. 131–208). Academic Press. https://doi.org/10.1016/bs.aesp.2022.11.003

Pierre, J. M. (2020). Mistrust and Misinformation: A Two-Component, Socio-Epistemic Model of

Belief in Conspiracy Theories. *Journal of Social and Political Psychology*, *8*(2), Article 2. https://doi.org/10.5964/jspp.v8i2.1362

Pornpitakpan, C. (2004). The Persuasiveness of Source Credibility: A Critical Review of Five Decades' Evidence. *Journal of Applied Social Psychology*, *34*(2), 243–281. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x

Porter, E., Velez, Y., & Wood, T. J. (2022). Factual Corrections Eliminate False Beliefs About COVID-19 Vaccines. *Public Opinion Quarterly*, *86*(3), 762–773. https://doi.org/10.1093/poq/nfac034

Pummerer, L. (2022). Belief in conspiracy theories and non-normative behavior. *Current Opinion in Psychology*, *47*, 101394. https://doi.org/10.1016/j.copsyc.2022.101394

Rathi, I., Taylor, S., Bergen, B. K., & Jones, C. R. (2024). *GPT-4 is judged more human than humans in displaced and inverted Turing tests* (arXiv:2407.08853). arXiv. https://doi.org/10.48550/arXiv.2407.08853

Salvi, F., Ribeiro, M. H., Gallotti, R., & West, R. (2024). *On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial* (arXiv:2403.14380). arXiv. http://arxiv.org/abs/2403.14380

Sherman, D. K., & Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. In *Advances in experimental social psychology, Vol 38* (pp. 183–242). Elsevier Academic Press. https://doi.org/10.1016/S0065-2601(06)38004-5

Sloman, S. A., & Vives, M.-L. (2022). Is political extremism supported by an illusion of understanding? *Cognition*, *225*, 105146. https://doi.org/10.1016/j.cognition.2022.105146

Stanovich, K. E., & Toplak, M. E. (2019). The need for intellectual diversity in psychological science: Our own studies of actively open-minded thinking as a case study. *Cognition*, *187*, 156–166. https://doi.org/10.1016/j.cognition.2019.03.006

Stasielowicz, L. (2024). *How to reduce conspiracy beliefs? A meta-analysis of intervention studies*. https://doi.org/10.31234/osf.io/6vs5u

Sundar, S. S., & Kim, J. (2019). Machine Heuristic: When We Trust Computers More than Humans with Our Personal Information. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–9. https://doi.org/10.1145/3290605.3300768

Sunstein, C. R., & Vermeule, A. (2008). Conspiracy Theories: Causes and Cures. *Journal of Political Philosophy*, *17*(2), 202–227. https://doi.org/10.1111/j.1467-9760.2008.00325.x

Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Tappin, B. M., Wittenberg, C., Hewitt, L. B., Berinsky, A. J., & Rand, D. G. (2023). Quantifying the potential persuasive returns to political microtargeting. *Proceedings of the National Academy of Sciences*, *120*(25), e2216261120. https://doi.org/10.1073/pnas.2216261120

Thaler, R. H., Sunstein, C. R., & Balz, J. P. (2010). *Choice Architecture* (SSRN Scholarly Paper 1583509). Social Science Research Network. https://doi.org/10.2139/ssrn.1583509

van Prooijen, J.-W., & van Vugt, M. (2018). Conspiracy Theories: Evolved Functions and Psychological Mechanisms. *Perspectives on Psychological Science*, *13*(6), 770–788. https://doi.org/10.1177/1745691618774270

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-Checking: A Meta-Analysis of What Works and for Whom. *Political Communication*, *37*(3), 350–375.

https://doi.org/10.1080/10584609.2019.1668894

Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation Schemes*. Cambridge University Press.

Wampold, B. E. (2015). How important are the common factors in psychotherapy? An update. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, *14*(3), 270–277. https://doi.org/10.1002/wps.20238

Williams, D. (2023). The case for partisan motivated reasoning. *Synthese*, *202*(3), 89. https://doi.org/10.1007/s11229-023-04223-1