Durably reducing conspiracy beliefs through dialogues with AI

Preprint · April 2024

DOI: 10.31234/osf.io/xcwdn

Thomas H Costello
Massachusetts Institute of Technology

Gordon Pennycook
Cornell University

Durably reducing conspiracy beliefs through dialogues with AI

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

[1]Sloan School of Management, Massachusetts Institute of Technology; Cambridge, MA, USA
[2]Department of Psychology, Cornell University; Ithaca, NY, USA
*Corresponding Author. Email: thcost@mit.edu

Abstract: Conspiracy theories are a paradigmatic example of beliefs that, once adopted, are extremely difficult to dispel. Influential psychological theories propose that conspiracy beliefs are uniquely resistant to counterevidence because they satisfy important needs and motivations. Here, we raise the possibility that previous attempts to correct conspiracy beliefs have been unsuccessful merely because they failed to deliver counterevidence that was sufficiently compelling and tailored to each believer's specific conspiracy theory (which vary dramatically from believer to believer). To evaluate this possibility, we leverage recent developments in generative artificial intelligence (AI) to deliver well-argued, person-specific debunks to a total of *N* = 2,190 conspiracy theory believers. Participants in our experiments provided detailed, open-ended explanations of a conspiracy theory they believed, and then engaged in a 3 round dialogue with a frontier generative AI model (GPT-4 Turbo) which was instructed to reduce each participant's belief in their conspiracy theory (or discuss a banal topic in a control condition). Across two experiments, we find robust evidence that the debunking conversation with the AI reduced belief in conspiracy theories by roughly 20%. This effect did not decay over 2 months time, was consistently observed across a wide range of different conspiracy theories, and occurred even for participants whose conspiracy beliefs were deeply entrenched and of great importance to their identities. Furthermore, although the dialogues were focused on a single conspiracy theory, the intervention spilled over to reduce beliefs in unrelated conspiracies, indicating a general decrease in conspiratorial worldview, as well as increasing intentions to challenge others who espouse their chosen conspiracy. These findings highlight that even many people who strongly believe in seemingly fact-resistant conspiratorial beliefs can change their minds in the face of sufficient evidence.

Note: This is a working paper, a preliminary version of research that is shared with the community for feedback and discussion. It has not yet been peer reviewed. Readers should keep this in mind when interpreting our findings and conclusions. We will make all the code, data, and materials associated with this research publicly available.

Last update: Apr 3, 2024

Widespread belief in unsubstantiated or false conspiracy theories is both a major source of public concern and focus of scholarly research (*1*–*3*). Conspiracy theories – in which events are understood as being caused by secret, malevolent plots involving powerful conspirators – have gained considerable traction across a wide range of topics with substantial societal relevance, from terrorist attacks to the COVID-19 pandemic, wars to elections, and aliens to assassinations. Despite the often quite implausible nature of many conspiracy theories, 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*).

This prevalence is particularly concerning, given that conspiracy beliefs are notoriously difficult to displace once adopted. To date, attempts to undermine entrenched conspiracy beliefs have proven largely unsuccessful (*8*–*10*). Indeed, conspiracy belief is often used as a paradigmatic example of resistance to evidence (*11*–*14*). That conspiracy theories seemingly thrive in the face of stark epistemic challenges has been a major driver of scholarly interest – if conspiracy theories persist despite being obviously wrong, this stands as a powerful challenge to scientific theories that center the importance of reasoning for belief formation and revision.

Accordingly, the belief in conspiracies - and believers' resistance to debunking attempts - have primarily been explained via social-psychological processes thought to blunt rational decision-making and receptivity to evidence. Popular explanations propose that people adopt conspiracy theories in order to sate underlying psychic "needs" or motivations, such as the need for control over one's environment and experiences (*3*, *15*). For instance, conspiracy theories may provide a sense of control by offering explanations that suggest that chaotic societal events, such as terrorist attacks or pandemics, are not random but are instead the result of deliberate actions by powerful groups (*16*). Other "needs" theorized to promote conspiracism include certainty and predictabliity (*17*), security and stability (*18*), and uniqueness (*19*). If these psychological needs are met by believing in conspiracy theories, the beliefs become more than just opinions; they become mechanisms for psychological comfort and stability - and thus are argued to be highly resistant to counterevidence (*3*). Conspiracies are also thought to become intertwined with individuals' overarching identity and worldview, which represent another psychological dimension that is argued to insulate conspiracy beliefs from counterevidence. Given that people have strong motivations to maintain their identity and/or group memberships (*20*–*22*), believers may use specific forms of biased information processing (motivated reasoning) where counterevidence is selectively ignored (*23*–*25*).

These perspectives, which center the needs and motivations of conspiracy theorists, paint a grim picture for countering conspiratorial beliefs: Because conspiracy believers at some level *want* to believe, convincing them to abandon unfounded beliefs should be virtually impossible without more fundamentally altering their underlying psychology and identity commitments.

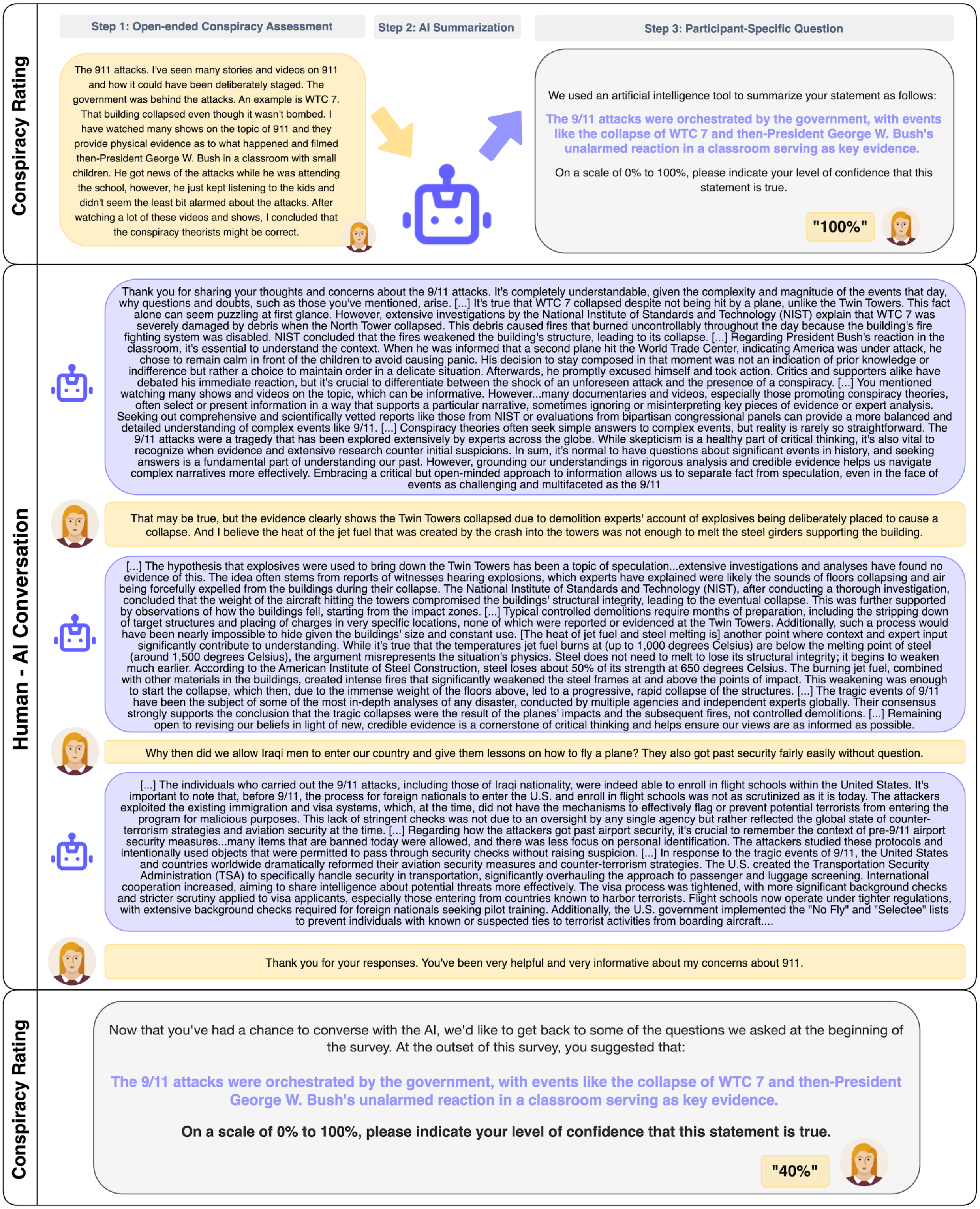
Here, we question the conventional wisdom about conspiracy theorists and ask whether it may, in fact, be possible to talk people out of the conspiratorial "rabbit hole" with sufficiently compelling evidence. By this line of reasoning, prior attempts at fact-based intervention may have failed simply due to a lack of depth and personalization of the corrective information. Entrenched conspiracy theorists are often quite knowledgeable about their conspiracy of interest, deploying prodigious (but typically erroneous or misinterpreted) lists of evidence in support of the conspiracy that can leave skeptics outmatched in debates and arguments (*26*, *27*). 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. Thus, canned persuasion attempts that argue broadly against a given conspiracy theory may not successfully address the specific evidence held by the believer – and thus may fail to be convincing.

This highly variable set of conspiratorial beliefs and supporting evidence presents a major challenge to would-be debunkers: Effectively refuting conspiracy theories using evidence and arguments is likely to require (a) vast stores of knowledge across a wide array of topics and (b) the ability to personalize counterarguments to match the specific conspiracies and evidence the believer brings to bear. In the absence of such capacities, it is difficult to assess whether psychological motives and identity commitments indeed render believers resistant to corrective information.

In the present work, we leverage recent developments in Large Language Models (LLMs) to shed new light on these issues (*28*). Powerful LLMs offer precisely the capacities described above: they have been trained on immense corpora of information, and can have real-time personalized conversations in which their responses are specifically tailored to the statements of their conversational partner (*29*). Thus, we ask whether LLMs can effectively refute conspiracy beliefs.

To do so, we develop a novel pipeline for conducting behavioral science research using real-time, personalized interactions between research subjects and LLMs. In our experiments, participants articulate a conspiracy theory in which they believe – in their own words – along with the evidence they think supports the theory. They then engage in a back-and-forth conversation with an artificial intelligence (AI) implemented using the LLM GPT-4 Turbo (*30*), in which the AI uses its store of knowledge to try to respond to the specific evidence raised by the participant and reduce the participant's belief in the particular conspiracy theory (or a control condition in which they converse with AI about an unrelated topic).

This design allows us to test whether tailored persuasive communication is indeed able to reduce already-held conspiracy beliefs; how the effectiveness of such communication varies based on 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 specific conspiracy theory articulated by the participants; and whether any such persuasion spills over to other related beliefs and behaviors. Finally, our design produces rich textual data from the conversations between the AI and the human participants, which we analyze to gain insight into what the humans believe and how the LLM engages in persuasion.



The 911 attacks. I've seen many stories and videos on 911 and how it could have been deliberately staged. The government was behind the attacks. An example is WTC 7. That building collapsed even though it wasn't bombed. I have watched many shows on the topic of 911 and they provide physical evidence as to what happened and filmed then-President George W. Bush in a classroom with small children. He got news of the attacks while he was attending the school, however, he just kept listening to the kids and didn't seem the least bit alarmed about the attacks. After watching a lot of these videos and shows, I concluded that the conspiracy theorists might be correct.

We used an artificial intelligence tool to summarize your statement as follows:

The 9/11 attacks were orchestrated by the government, with events like the collapse of WTC 7 and then-President George W. Bush's unalarmed reaction in a classroom serving as key evidence.

On a scale of 0% to 100%, please indicate your level of confidence that this statement is true.

"100%"

Thank you for sharing your thoughts and concerns about the 9/11 attacks. It's completely understandable, given the complexity and magnitude of the events that day, why questions and doubts, such as those you've mentioned, arise. [...] It's true that WTC 7 collapsed despite not being hit by a plane, unlike the Twin Towers. This fact alone can seem puzzling at first glance. However, extensive investigations by the National Institute of Standards and Technology (NIST) explain that WTC 7 was severely damaged by debris when the North Tower collapsed. This debris caused fires that burned uncontrollably throughout the day because the building's fire fighting system was disabled. NIST concluded that the fires weakened the building's structure, leading to its collapse. [...] Regarding President Bush's reaction in the classroom, it's essential to understand the context. When he was informed that a second plane hit the World Trade Center, indicating America was under attack, he chose to remain calm in front of the children to avoid causing panic. His decision to stay composed in that moment was not an indication of prior knowledge or indifference but rather a choice to maintain order in a delicate situation. Afterwards, he promptly excused himself and took action. Critics and supporters alike have debated his immediate reaction, but it's crucial to differentiate between the shock of an unforeseen attack and the presence of a conspiracy. [...] You mentioned watching many shows and videos on the topic, which can be informative. However...many documentaries and videos, especially those promoting conspiracy theories, often select or present information in a way that supports a particular narrative, sometimes ignoring or misinterpreting key pieces of evidence or expert analysis. Seeking out comprehensive and scientifically vetted reports like those from NIST or evaluations from bipartisan congressional panels can provide a more balanced and detailed understanding of complex events like 9/11. [...] Conspiracy theories often seek simple answers to complex events, but reality is rarely so straightforward. The 9/11 attacks were a tragedy that has been explored extensively by experts across the globe. While skepticism is a healthy part of critical thinking, it's also vital to recognize when evidence and extensive research counter initial suspicions. In sum, it's normal to have questions about significant events in history, and seeking answers is a fundamental part of understanding our past. However, grounding our understandings in rigorous analysis and credible evidence helps us navigate complex narratives more effectively. Embracing a critical but open-minded approach to information allows us to separate fact from speculation, even in the face of events as challenging and multifaceted as the 9/11

That may be true, but the evidence clearly shows the Twin Towers collapsed due to demolition experts' account of explosives being deliberately placed to cause a collapse. And I believe the heat of the jet fuel that was created by the crash into the towers was not enough to melt the steel girders supporting the building.

[...] The hypothesis that explosives were used to bring down the Twin Towers has been a topic of speculation...extensive investigations and analyses have found no evidence of this. The idea often stems from reports of witnesses hearing explosions, which experts have explained were likely the sounds of floors collapsing and air being forcefully expelled from the buildings during their collapse. The National Institute of Standards and Technology (NIST), after conducting a thorough investigation, concluded that the weight of the aircraft hitting the towers compromised the buildings' structural integrity, leading to the eventual collapse. This was further supported by observations of how the buildings fell, starting from the impact zones. [...] Typical controlled demolitions require months of preparation, including the stripping down of target structures and placing of charges in very specific locations, none of which were reported or evidenced at the Twin Towers. Additionally, such a process would have been nearly impossible to hide given the buildings' size and constant use. [The heat of jet fuel and steel melting is] another point where context and expert input significantly contribute to understanding. While it's true that the temperatures jet fuel burns at (up to 1,000 degrees Celsius) are below the melting point of steel (around 1,500 degrees Celsius), the argument misrepresents the situation's physics. Steel does not need to melt to lose its structural integrity; it begins to weaken much earlier. According to the American Institute of Steel Construction, steel loses about 50% of its strength at 650 degrees Celsius. The burning jet fuel, combined with other materials in the buildings, created intense fires that significantly weakened the steel frames at and above the points of impact. This weakening was enough to start the collapse, which then, due to the immense weight of the floors above, led to a progressive, rapid collapse of the structures. [...] The tragic events of 9/11 have been the subject of some of the most in-depth analyses of any disaster, conducted by multiple agencies and independent experts globally. Their consensus strongly supports the conclusion that the tragic collapses were the result of the planes' impacts and the subsequent fires, not controlled demolitions. [...] Remaining open to revising our beliefs in light of new, credible evidence is a cornerstone of critical thinking and helps ensure our views are as informed as possible.

Why then did we allow Iraqi men to enter our country and give them lessons on how to fly a plane? They also got past security fairly easily without question.

[...] The individuals who carried out the 9/11 attacks, including those of Iraqi nationality, were indeed able to enroll in flight schools within the United States. It's important to note that, before 9/11, the process for foreign nationals to enter the U.S. and enroll in flight schools was not as scrutinized as it is today. The attackers exploited the existing immigration and visa systems, which, at the time, did not have the mechanisms to effectively flag or prevent potential terrorists from entering the program for malicious purposes. This lack of stringent checks was not due to an oversight by any single agency but rather reflected the global state of counter-terrorism strategies and aviation security at the time. [...] Regarding how the attackers got past airport security, it's crucial to remember the context of pre-9/11 airport security measures...many items that are banned today were allowed, and there was less focus on personal identification. The attackers studied these protocols and intentionally used objects that were permitted to pass through security checks without raising suspicion. [...] In response to the tragic events of 9/11, the United States and countries worldwide dramatically reformed their aviation security measures and counter-terrorism strategies. The U.S. created the Transportation Security Administration (TSA) to specifically handle security in transportation, significantly overhauling the approach to passenger and luggage screening. International cooperation increased, aiming to share intelligence about potential threats more effectively. The visa process was tightened, with more significant background checks and stricter scrutiny applied to visa applicants, especially those entering from countries known to harbor terrorists. Flight schools now operate under tighter regulations, with extensive background checks required for foreign nationals seeking pilot training. Additionally, the U.S. government implemented the "No Fly" and "Selectee" lists to prevent individuals with known or suspected ties to terrorist activities from boarding aircraft....

Thank you for your responses. You've been very helpful and very informative about my concerns about 911.

Now that you've had a chance to converse with the AI, we'd like to get back to some of the questions we asked at the beginning of the survey. At the outset of this survey, you suggested that:

The 9/11 attacks were orchestrated by the government, with events like the collapse of WTC 7 and then-President George W. Bush's unalarmed reaction in a classroom serving as key evidence.

On a scale of 0% to 100%, please indicate your level of confidence that this statement is true.

"40%"

Figure 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 pre-treatment 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 relevant topics). Following three rounds of dialogue, respondents once again rated their belief in the summarized conspiracy statement, serving as our post-treatment measure. Shown is an example treatment dialogue which led the participant to substantially reduce their belief.

Results

Can conspiracy beliefs be refuted?

In Study 1, participants indicated their belief in 15 popular conspiracy theories (from the Belief in Conspiracy Theories Index, 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 pre-treatment measure of belief. This open-ended measurement approach avoids a longstanding criticism of discrete conspiracism measures, such as the BCTI, for failing to representatively sample from the universe of possible conspiracies (*31*).

Out of *N*=1,055 American participants (quota-matched to the U.S. census on age, gender, race, and ethnicity) who completed the pre-treatment measures, 75.2% indicated belief in a conspiracy theory and were included in our subsequent analyses, whereas 14.8% said they did not believe any conspiracy theories or described a belief that the AI classified as not actually conspiratorial (for coding validation, see SI section 3.1) and 10.6% described a conspiracy theory but had belief below the scale midpoint.

To assess whether the AI could reduce conspiracy belief, participants were then randomly assigned to either have a 3-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). For each participant, the AI was (a) provided with that participant's specific open-ended response, including their stated rationale for believing the conspiracy theory and their degree of endorsement and (b) prompted to use simple language to persuade the user that their conspiracy theory is not supported and change their beliefs to be less conspiratorial. Following the conversations, all participants re-rated belief in their stated conspiracy theory (see **Figure 1** for key methodological steps and a sample conversation).

Was conversing with an AI able to successfully reduce participants' conspiratorial beliefs? Indeed, the treatment reduced participants' belief in their conspiracy theory participants' stated conspiracy by 16.5 percentage points more than the control (linear regression with robust standard errors controlling for pre-treatment belief, 95% CI [13.6, 19.4], *p* < .001, *d* = 1.13; **Figure 2a**). This translates into a 21.43% decrease in belief among those in treatment (vs. 1.04% in the control). Furthermore, over a quarter (27.4%) of participants in the treatment became uncertain in their conspiracy belief (i.e. belief below the scale midpoint) following the conversation, compared to only 2.4% in the control.

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 find no significant change in belief in the focal conspiracy theory from immediately after the

AI conversation to 2 months later in a mixed-effects model with fixed effects for experimental condition and time point and random intercepts for participants, $b_{\Delta ImmediatelyPost\ -\ 2Month}$ = 0.03, 95% CI [-2.24, 2.31], *p* = .98; **Figure 2A**). This result is robust to assuming that the 14% of participants who did not complete the follow-up returned to their initial pre-treatment belief levels (b = 12.70 95% CI [9.47, 15.93], p < .001). Thus, the change in beliefs we observe is remarkably persistent.

In Study 1, the proportion of participants who endorsed a conspiracy via free-text response (75.6%) 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 explore this possibility, as well as the replicability of our results and robustness to minor design changes, in Study 2 where N=2,286 Americans completed an extremely similar procedure without the BCTI (and with slightly different wording for the conspiracy elicitation prompt). Here, 64.6% of participants indicated belief in a conspiracy theory. Most importantly, we replicate the experimental results of Study 1: Participants in the treatment in Study 2 reduced belief in their focal conspiracy by 12.39 percentage points more than participants in the control (95% CI [10.07, 14.72], *p* < .001, d = 0.79; **Figure 2B**), translating into a 19.41% decrease in belief (versus a 2.94% decrease in the control).

Robustness across topics and people

Next, we examine the robustness of the AI conservation treatment effect. We begin by investigating whether the treatment size varies across the specific focal conspiracy theories articulated by the participants. To do so, we used a multi-step natural language processing and clustering approach to classify each focal conspiracy theory according to its contents (see Materials & Methods and Table S18 for details). We find that the treatment effect did not differ significantly across conspiracy type in an omnibus test (F[12, 1971] = 1.30, p = .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 (**Figure 2c**). Notably, the treatment worked even for highly salient - and likely deeply entrenched - political conspiracies such as those involving fraud in the 2020 US Presidential Election (*b* = 10.61 [5.54, 15.67], *p* < .001, *d* = .82) and the COVID-19 pandemic (*b* = 11.79 [6.98, 16.60], *p* < .001, d = .73). In addition to allowing us to test for the robustness of our treatment, this classification based on the participant’s open-ended responses also provides novel descriptive insight into which particular conspiracy theories Americans subscribe to.

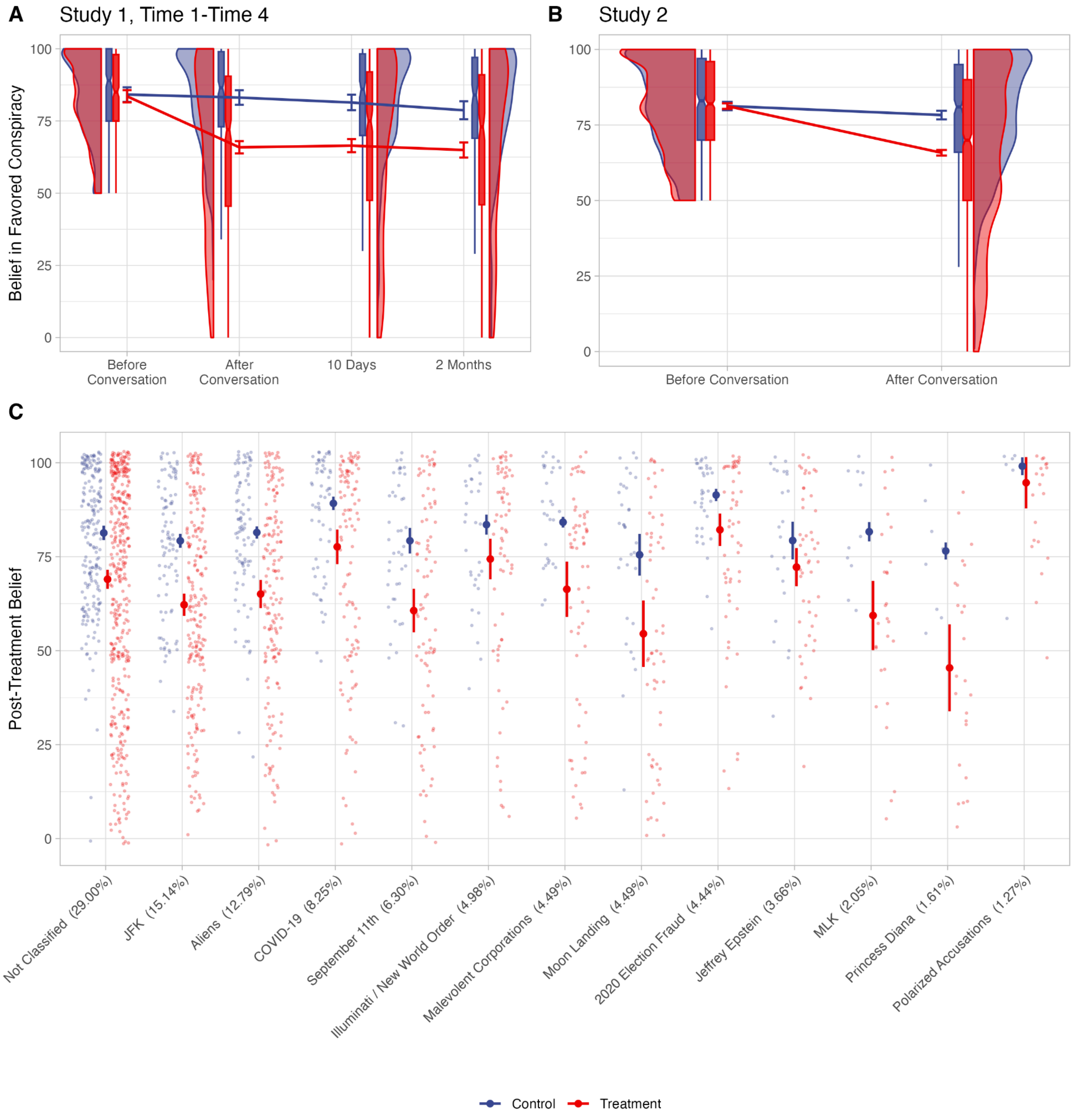


Figure 2. A brief conversation with an AI model durably reduces belief in conspiracy theories. *Top: Average belief in each participant's focal conspiracy theory in by condition (treatment, in which the AI attempted to refute the conspiracy theory, in red; control, in which the AI discussed an irrelevant topic, in blue) and time point for Study 1 (A) and Study 2 (B). Before-conversation belief is greater than 50 for all participants because participants with initial belief below 50 were excluded from the study. Bottom: Belief immediately after the AI conversation by condition and topic of the participant's focal conspiracy theory; see Materials and Methods section 3.2 for details of topic detection. Error bars indicate 95% confidence intervals.*

We now turn to variation in effect sizes across individuals. In particular, we ask whether the treatment is effective even among participants likely to have particularly entrenched beliefs. We use generalized additive models (GAMs) to analyze how the treatment effect varies in a non-linear manner based on several measures relevant to entrenchment. First, we examine participants' level of pre-treatment belief in the focal conspiracy, and find that it does significantly moderate the treatment effect, resulting in a u-shaped curve (ΔAIC = -3.25, ΔR^2 = .002, p = .022; **Figure 3A**). Second, we examine how important participants indicated the conspiracy theory is to their worldview (**Figure 3B**), which does significantly decrease the size of the treatment effect (ΔAIC = 3.12, ΔR^2 = .003, *p* = .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* = .038, *d* = .53). Third, we examined participants' level of general conspiratorial ideation (i.e. the number of BCTI conspiracies they believed), which showed non-significant moderation of the treatment effect, such that the treatment impact may be somewhat diminished among participants with the highest levels of conspiratorial ideation (ΔAIC = 0.77, ΔR^2 = .002, p = .108; **Figure 3C**). Yet participants at or above the 90th percentile of conspiratorial ideation still displayed a substantial average treatment effect of *b* = 9.07 (95% CI [2.73, 15.44], *p* = .006, *d* = .53).

We also examine 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, intellectual humility, actively open-minded thinking, and institutional trust). In a single linear regression model including all candidate moderators and their interaction with experimental condition, as well as a controls for conspiracy type and its interaction with experimental condition, only (a) trust in generative AI and (b) institutional trust replicably moderated the treatment effect, such that those higher in both kinds of trust showed larger treatment effects (see Supplemental Tables 9-10 for all model coefficients for both experiments). We conducted a post hoc analysis using the Causal Forest method (*32*) to further clarify and identify heterogeneous effects of the intervention across all moderators (including conspiracy type, pre-treatment beliefs, and importance). While there were heterogeneous treatment effects across subgroups (*t* = 4.97, *p* < .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; the CATE ranged from -12.5 to -4.8 (median = -8.2) for individuals with 0/10 institutional trust; and the CATE ranged from -18.2 to -10.0 (median = -15.4) for individuals with 95th percentile and above BCTI scores. Variable importance analyses indicated that, for Experiment 1, the predominant determinants of treatment effect heterogeneity (in order) were participant's age, trust in generative AI, and BCTI scores; in Experiment 2, these were institutional trust, trust in generative AI, age, and conspiracy-importance.

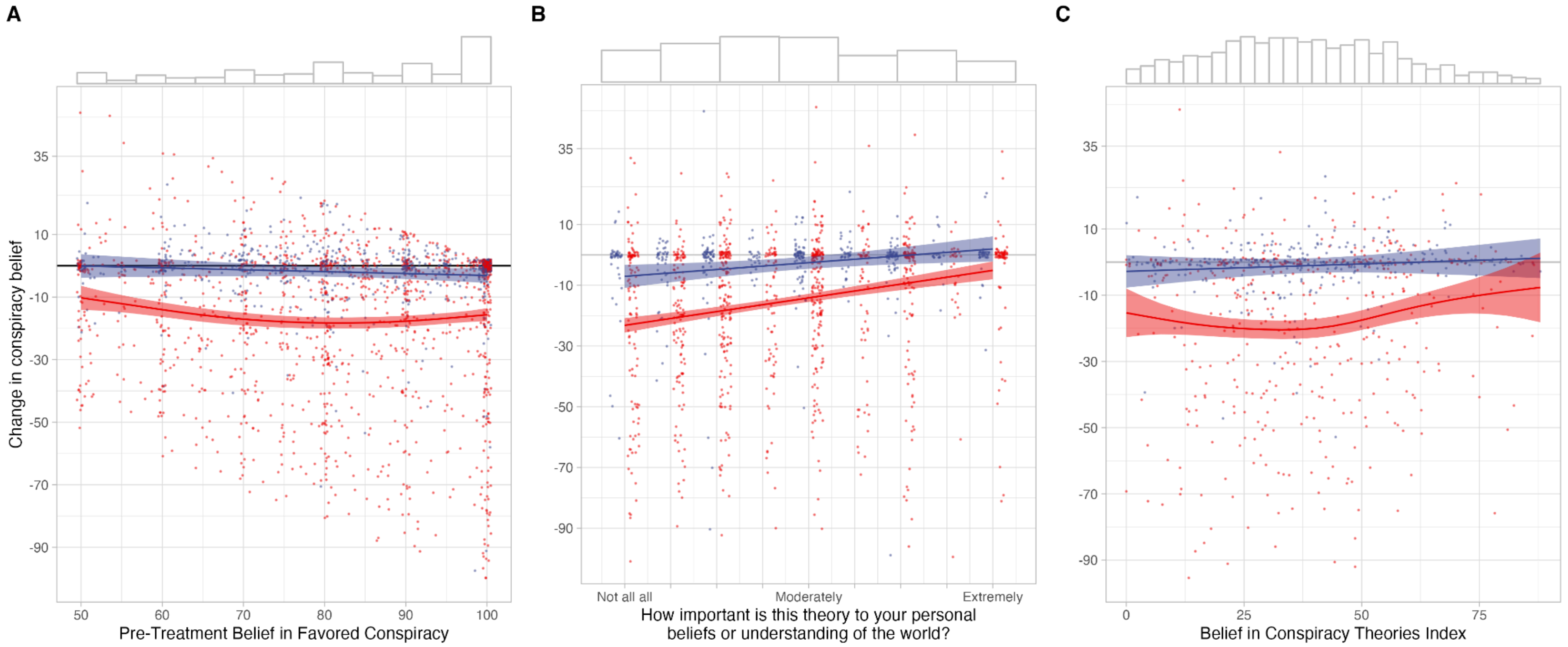


Figure 3. The treatment is effective even for those who are strongly attached to their conspiracy beliefs. *Shown is the change in belief in the focal conspiracy from before AI conversation to after AI conversation, for the treatment (red) and control (blue) conditions. Data are pooled across studies to maximize power. Individual observations are plotted along with fit lines and 95% confidence intervals generated using generalized additive models. We conduct separate analyses for predictors of participant's pre-treatment 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 Belief in Conspiracy Theories Index completed pre-treatment (C).*

Spillover effects and behavioral implications

Next, we examined treatment effects on outcomes beyond belief in the focal conspiracy. First, we ask whether the treatment effect affected individuals' beliefs in conspiracy theories that were *not* discussed during the conversation with the AI model. We did so by analyzing respondents' belief in 15 widespread conspiracy theories from the BCTI (which is assessed both pre-treatment and post-treatment in Study 1). We employed a linear mixed model with fixed effects for experimental conditions and time point (pre, post, 10-day, 2-months) and random intercepts for participant. Post-intervention, there was a 3.05-point decrease in generic conspiracy belief in the active condition (95% CI [-3.90, -2.20], p < .001, 8.2% decrease; **Figure 4A**), compared to a 1.64-point increase in the control (*d* = .21). This effect was still evident at the 2-month follow-up, with a 2.46-point decrease from pre-treatment (95% CI [-3.44, -1.49], *p* < .001). When only analyzing belief in BCTI conspiracy theories that a given participant believed pre-treatment (i.e. endorsed above the scale midpoint), the impact was more pronounced: a 9.39-point reduction immediately post-intervention (95% CI [-11.06, -7.72], p < .001, 12% decrease, **Figure 4B**), compared to a 3.32-point reduction in the control (*d* = 0.53). This difference between treatment and control persisted at the 2-month follow-up ($b_{\Delta Treatment\ -\ Control}$ = -5.34, 95% CI [-8.40, -2.29], *p* < .001).

In Study 2, we investigated the treatment's influence on participants' behavioral intentions. We found that the treatment significantly increased intentions to ignore or unfollow social media accounts espousing the focal conspiracy (**β** = .39 [.27, .50], p < .001; **Figure 4C**) and significantly increased willingness to ignore or argue against people who believe the focal conspiracy (**β** = .42 [.31, .54], p < .001; **Figure 4D**). There was a directional but non-significant decrease in intentions to join pro-conspiracy protests (**β** = -.12 [-.27, .03], p = .12; **Figure 4E**) - intentions which were low at baseline, potentially creating a floor effect.

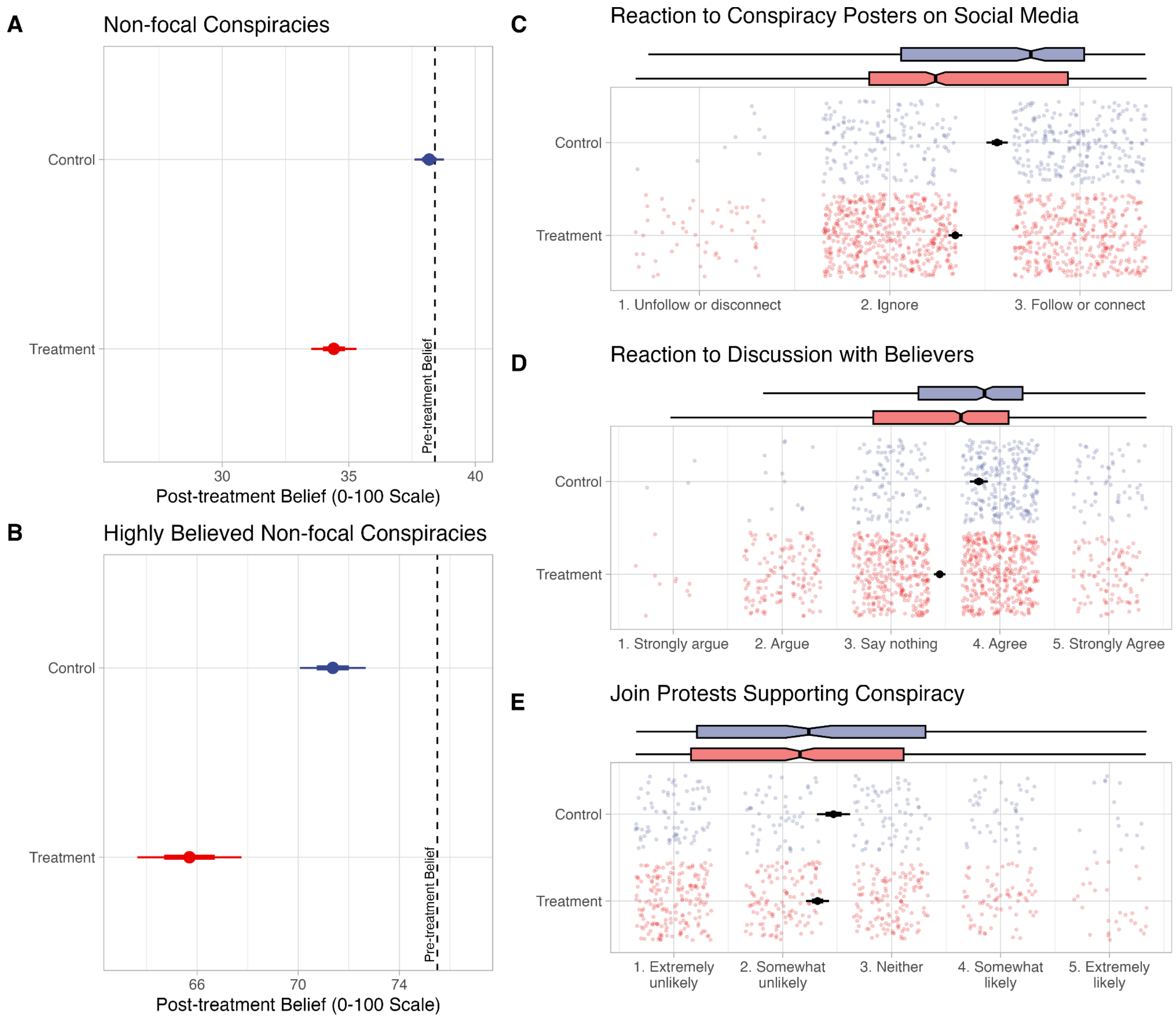


Figure 4. The treatment also affects belief in other conspiracy theories and behavioral intentions. *First column: Post conversation average belief in the 15 conspiracies from the Belief in Conspiracy Theories Index (excluding the focal conspiracy, if it was one of those 15) by condition, for all conspiracies (A) and for only the subset of conspiracies with the participant indicated believing pre-treatment (B). Vertical dotted line indicates average pretreatment belief. Second column: Post-conversation 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, and (D) how likely they would be to participate in a protest in support of the focal conspiracy. Thick error bars indicate 66% confidence intervals, thin error bars indicate 95% confidence intervals.*

What occurred during the conversations?

Finally, we shed light on how the AI went about persuading conspiracy theorists, via post hoc natural language processing analyses of the conversations (we pooled data across studies to maximize power). We first had the model list the strategies it would use in the setting of our experiment, and then had it go through each conversation and indicate the extent to which strategy(s) were used in that conversation. Strikingly, reasoning-based strategies were clearly the most frequently used approach (see **Figure 5**): evidence-based alternative perspectives were used “extensively” in a large majority of conversations (83%) and encouraging critical thinking was either used “extensively” (47%) or used “moderately” in virtually all conversations (52%). Conversely, the rapport-building strategies of finding common ground and expressing understanding were used only “moderately” in most conversations, and other strategies (including various psychological and social/emotional strategies) were used even less. These descriptive results suggest that the AI was largely being persuasive due to actual use of evidence and arguments to change people’s minds.

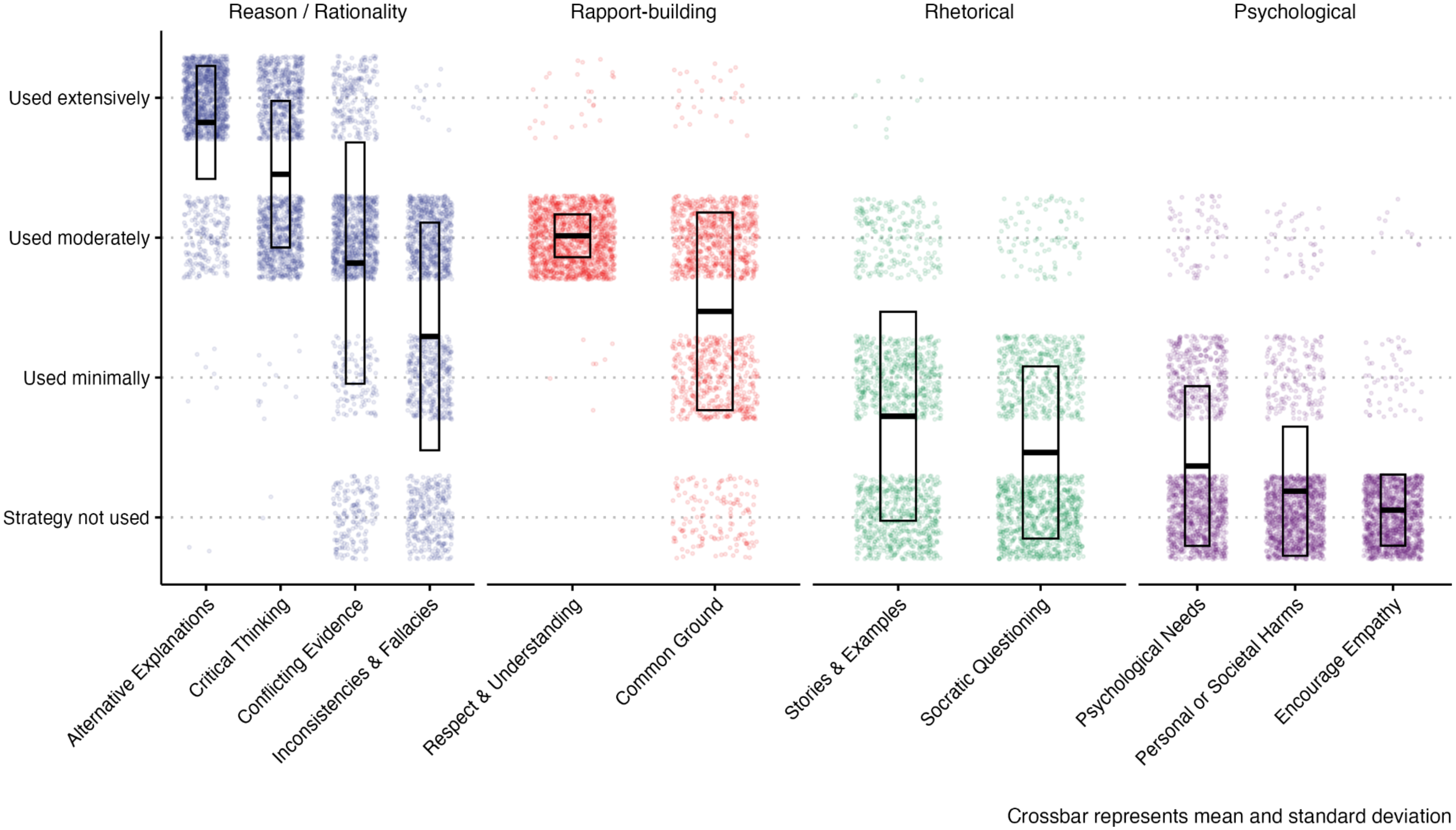


Figure 5. The AI responses overwhelmingly use reason and arguments to persuade, rather than psychological strategies. Shown are individual (raw data) and summarized (crossbar) ratings of the presence and prevalence of 11 persuasion strategies used by the AI model during each conversation – based on natural language processing analyses conducted using GPT-4. The full procedure and model prompt are described in section 3.3 and Table S15.

Discussion

Conspiracy theories are widely seen as a paradigmatic example of beliefs that rarely change in response to evidence (*11–13*). Yet here we have demonstrated that a brief conversation with an out-of-the-box large language model AI systes – which relied most heavily on providing counterevidence and encouraging critical thinking – substantially reduced belief in a wide range of conspiracy theories. This was the case even though participants articulated in their own words a specific conspiracy theory they believed in (rather than choosing from a pre-selected list); occurred even among those participants most committed to their conspiratorial beliefs; and produced a persistent effect that not only lasted for two months, but was virtually *undiminished* in that time. This indicates that the dialogue produced a meaningful and lasting change in beliefs for a meaningful proportion of the conspiracy believers in our study.

These findings profoundly 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 – and, accordingly, challenge the family of social-psychological theories that center psychological "needs" and motivations when explaining conspiratorial belief (*1*, *3*, *33*). Instead, our findings are more consistent with an alternative theoretical perspective whereby epistemically suspect beliefs – such as conspiracy theories, but also superstitions and paranormal beliefs (*34*), belief in misinformation (*35*) and receptivity to pseudo-profound bullshit (*36*) – primarily arise due to a failure to engage in reasoning, reflection, and careful deliberation (*37*). Past work has shown that conspiracy believers tend to be more intuitive (*38*) and overconfident (*39*) than those who are skeptical of conspiracies, and that conspiracy believers massively overestimated how many others also believe in conspiracies (*39*). This suggests that conspiracy belief may be more passive than commonly assumed – that is, that they may be beliefs that people fall into (for various reasons) rather than being actively sought out as a means to fulfill psychological needs. Consistent with this suggestion, here we show that when confronted with an AI that compellingly argues against their beliefs, many conspiracists – even those strongly committed to their beliefs – do in fact update their views.

Our findings also have practical implications for those seeking to reduce belief in conspiracy theories. Most broadly, arguments and evidence should not be abandoned. More specifically, AI models such as GPT-4 have the potential to be powerful tools for reducing epistemically suspect beliefs. Many conspiracy theorists are eager to discuss their beliefs with anyone willing to listen, and thus may actually be enticed by the opportunity to talk to an AI about them – if, for example, AI bots with prompts like those used in our experiments were created on forums or social media sites popular with conspiracy theorists, or if ads redirecting to such AIs were targeted at people entering conspiracy-related terms into search engines. Consistent with the potential for uptake of AI dialogues, many 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. Surprisingly in this AI response, it's actually made more sense then the average human response I've ever read or listened to. Thank you"; "That's great. Thank you so much for all of the information, AI! I'll use you again in the future for questions about complex problems."; "wow your smart. you are amazing. how can you give so much information to so many people at once?"; and "Thank you for your responses. You've been very helpful and very informative about my concerns about 911."). Such approaches to belief reduction offer an important complement to existing approaches, which seek to prevent people from adopting conspiracy beliefs in the first place – for example, by "inoculating" those who do not yet believe in an effort to help them build epistemic resistance (*40*, *41*).

On the other hand, the effectiveness of AI persuasion demonstrated in our studies also relates to ongoing debates regarding the promise versus peril of generative AI (*42*, *43*). Absent appropriate guardrails, it is entirely possible that such models could also convince people to *believe* conspiracy theories, or adopt other epistemically suspect beliefs (*44*) – or be used as tools of large-scale persuasion more generally (*45*). Thus, our findings emphasize both the potential positive impacts of generative AI when deployed responsibly, and the crucial importance of minimizing opportunities for this technology to be used irresponsibly.

The experimental paradigm presented in this paper also represents a substantial methodological advancement in behavioral science research. Traditional survey experiments typically rely on static, predetermined stimuli and questions, which limits their ability to probe and respond to individuals' beliefs. In contrast, the real-time use of LLMs embedded in the survey enables the researcher to elicit open-ended statements of belief (or anything else) and translate them into quantitative outcomes. Furthermore, the integrated AI can engage in back-and-forth dialogues with participants, adapting its responses based on the specific information provided by each individual (as opposed to, for example, using LLMs to pre-generate static stimuli as in past work) (*46*, *47*). 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. The open-ended nature of the human-AI conversations also produces a wealth of rich textual data that can be analyzed using natural language processing techniques (including by the same AI model used in the conversations). This allows researchers to gain deeper insights into the content and structure of participants' beliefs, as well as the strategies employed by the AI to challenge those beliefs. Integrating LLMs into survey experiments opens up new avenues for conducting adaptive, personalized studies. Minor modifications to our paradigm would allow for personalized AI-generated images, dynamic agents in collective decision-making games, psychological state inductions, and monitors of inattentive responding. This approach has the potential to significantly enhance our understanding of complex psychological phenomena.

Finally, our findings raise a wide array of important directions of future research. First, there are questions related to our specific intervention. In our experiment, participants were explicitly informed that they were conversing with an AI. Future work should test how the effects vary if participants are instead told that they were conversing with a human (and if the supposed characteristics of that human, e.g. partisanship, are varied). Furthermore, shedding more light on how exactly the AI successfully persuades others – for example, by prompting it to use or avoid using specific persuasion techniques – is another promising area to explore in future work. Beyond conspiracy theories, the approach we introduce here can be used to examine the limits of persuasion across a wide range of beliefs, allowing scholars to map out which beliefs are more versus less resistant to evidence and arguments.

In sum, it has become almost a truism that people that are “down the rabbit hole” of conspiracy belief are almost impossible to reach. In contrast to this pessimistic view, we have shown that a 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 the reason it has proven so difficult to dissuade people from their conspiracy beliefs is that they simply have not been given good enough counterevidence. This paints a picture of human reasoning that is surprisingly optimistic: Conspiracists are not all entirely blinded by psychological needs and motivations – it just takes a genuinely strong argument to reach them.

References

1. S. M. Bowes, T. H. Costello, A. Tasimi, The conspiratorial mind: A meta-analytic review of motivational and personological correlates. *Psychol. Bull.* (2023).
2. M. Butter, P. Knight, Eds., *Routledge Handbook of Conspiracy Theories* (Routledge, London, 2020).
3. K. M. Douglas, R. M. Sutton, A. Cichocka, The Psychology of Conspiracy Theories. *Curr. Dir. Psychol. Sci.* **26**, 538–542 (2017).
4. J. E. Oliver, T. J. Wood, Conspiracy Theories and the Paranoid Style(s) of Mass Opinion. *Am. J. Polit. Sci.* **58**, 952–966 (2014).
5. H. G. West, T. Sanders, *Transparency and Conspiracy: Ethnographies of Suspicion in the New World Order* (Duke University Press, 2003).
6. J. E. Uscinski, J. M. Parent, *American Conspiracy Theories* (Oxford University Press, 2014).
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).
8. P. Krekó, “Countering Conspiracy Theories and Misinformation” in *Routledge Handbook of Conspiracy Theories*, M. Butter, P. Knight, Eds. (Routledge, Abingdon, Oxon ; New York, NY : Routledge, 2020., ed. 1, 2020; https://www.taylorfrancis.com/books/9780429840593/chapters/10.4324/9780429452734-2_8), pp. 242–255.
9. 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).
10. L. Stasielowicz, *How to Reduce Conspiracy Beliefs? A Meta-Analysis of Intervention Studies* (2024).
11. S. Lewandowsky, G. E. Gignac, K. Oberauer, The Role of Conspiracist Ideation and Worldviews in Predicting Rejection of Science. *PLoS ONE* **8**, e75637 (2013).
12. M. G. Napolitano, “Conspiracy Theories and Resistance to Evidence,” thesis, UC Irvine (2022).
13. M. G. Napolitano, “Conspiracy Theories and Evidential Self-Insulation” in *The Epistemology of Fake News*, S. Bernecker, A. K. Flowerree, T. Grundmann, Eds. (Oxford University Press, 2021; https://doi.org/10.1093/oso/9780198863977.003.0005), p. 0.
14. C. R. Sunstein, A. Vermeule, Conspiracy Theories: Causes and Cures. *J. Polit. Philos.* **17**, 202–227 (2008).
15. J. A. Whitson, A. D. Galinsky, Lacking Control Increases Illusory Pattern Perception. *Science* **322**, 115–117 (2008).
16. A. C. Kay, J. A. Whitson, D. Gaucher, A. D. Galinsky, Compensatory Control: Achieving Order Through the Mind, Our Institutions, and the Heavens. *Curr. Dir. Psychol. Sci.* **18**, 264–268 (2009).
17. 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).
18. J.-W. van Prooijen, An Existential Threat Model of Conspiracy Theories. *Eur. Psychol.* **25**, 16–25 (2020).
19. A. Lantian, D. Muller, C. Nurra, K. M. Douglas, “I Know Things They Don’t Know!” *Soc. Psychol.* **48**, 160–173 (2017).
20. 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).
21. 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).
22. 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).
23. R. Brotherton, *Suspicious Minds: Why We Believe Conspiracy Theories* (Bloomsbury Publishing, 2015).
24. R. K. Garrett, B. E. Weeks, Epistemic beliefs’ role in promoting misperceptions and conspiracist ideation. *PLOS ONE* **12**, e0184733 (2017).
25. N. Dagnall, K. Drinkwater, A. Parker, A. Denovan, M. Parton, Conspiracy theory and cognitive style: a worldview. *Front. Psychol.* **6** (2015).
26. 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).

27. P. M. Fernbach, J. E. Bogard, Conspiracy Theory as Individual and Group Behavior: Observations from the Flat Earth International Conference. *Top. Cogn. Sci.* **n/a**.
28. B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, D. Roth, Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. *ACM Comput. Surv.* **56**, 30:1-30:40 (2023).
29. H. Wang, J. Li, H. Wu, E. Hovy, Y. Sun, Pre-Trained Language Models and Their Applications. *Engineering* **25**, 51–65 (2023).
30. OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Ł. Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Ł. Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de A. B. Peres, M. Petrov, H. P. de O. Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. J. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, B. Zoph, GPT-4 Technical Report. arXiv arXiv:2303.08774 [Preprint] (2024). https://doi.org/10.48550/arXiv.2303.08774.
31. 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.* **n/a**.
32. S. Athey, J. Tibshirani, S. Wager, Generalized random forests. *Ann. Stat.* **47**, 1148–1178 (2019).
33. 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).
34. 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).
35. 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).
36. 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).
37. G. Pennycook, "Chapter Three - A framework for understanding reasoning errors: From fake news to climate change and beyond" in *Advances in Experimental Social Psychology*, B. Gawronski, Ed. (Academic Press, 2023; https://www.sciencedirect.com/science/article/pii/S0065260122000284)vol. 67, pp. 131–208.

38. J. Binnendyk, G. Pennycook, Intuition, reason, and conspiracy beliefs. *Curr. Opin. Psychol.* **47**, 101387 (2022).
39. G. Pennycook, J. Binnendyk, D. Rand, Overconfidently conspiratorial: Conspiracy believers are dispositionally overconfident and massively overestimate how much others agree with them. OSF [Preprint] (2022). https://doi.org/10.31234/osf.io/d5fz2.
40. J. A. Banas, G. Miller, Inducing Resistance to Conspiracy Theory Propaganda: Testing Inoculation and Metainoculation Strategies. *Hum. Commun. Res.* **39**, 184–207 (2013).
41. D. Jolley, K. M. Douglas, Prevention is better than cure: Addressing anti-vaccine conspiracy theories. *J. Appl. Soc. Psychol.* **47**, 459–469 (2017).
42. E. Klein, Opinion | This Changes Everything, *The New York Times* (2023). https://www.nytimes.com/2023/03/12/opinion/chatbots-artificial-intelligence-future-weirdness.html.
43. D. Allen, E. G. Weyl, The Real Dangers of Generative AI. *J. Democr.* **35**, 147–162 (2024).
44. M. Phuong, M. Aitchison, E. Catt, S. Cogan, A. Kaskasoli, V. Krakovna, D. Lindner, M. Rahtz, Y. Assael, S. Hodkinson, H. Howard, T. Lieberum, R. Kumar, M. A. Raad, A. Webson, L. Ho, S. Lin, S. Farquhar, M. Hutter, G. Deletang, A. Ruoss, S. El-Sayed, S. Brown, A. Dragan, R. Shah, A. Dafoe, T. Shevlane, Evaluating Frontier Models for Dangerous Capabilities. arXiv arXiv:2403.13793 [Preprint] (2024). https://doi.org/10.48550/arXiv.2403.13793.
45. M. Burtell, T. Woodside, Artificial Influence: An Analysis Of AI-Driven Persuasion. arXiv arXiv:2303.08721 [Preprint] (2023). https://doi.org/10.48550/arXiv.2303.08721.
46. H. Bai, J. Voelkel, J. Eichstaedt, R. Willer, Artificial intelligence can persuade humans on political issues. (2023).
47. 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**, 116:1-116:29 (2023).
48. 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).
49. M. Faverio, A. Tyson, What the data says about Americans' views of artificial intelligence, *Pew Research Center*. https://www.pewresearch.org/short-reads/2023/11/21/what-the-data-says-about-americans-views-of-artificial-intelligence/.
50. OECD, *OECD Guidelines on Measuring Trust* (Organisation for Economic Co-operation and Development, Paris, 2017; https://www.oecd-ilibrary.org/governance/oecd-guidelines-on-measuring-trust_9789264278219-en).
51. A. Neelakantan, T. Xu, R. Puri, A. Radford, J. M. Han, J. Tworek, Q. Yuan, N. Tezak, J. W. Kim, C. Hallacy, J. Heidecke, P. Shyam, B. Power, T. E. Nekoul, G. Sastry, G. Krueger, D. Schnurr, F. P. Such, K. Hsu, M. Thompson, T. Khan, T. Sherbakov, J. Jang, P. Welinder, L. Weng, Text and Code Embeddings by Contrastive Pre-Training. arXiv arXiv:2201.10005 [Preprint] (2022). https://doi.org/10.48550/arXiv.2201.10005.
52. K. D. Carlson, F. L. Schmidt, Impact of experimental design on effect size: Findings from the research literature on training. *J. Appl. Psychol.* **84**, 851–862 (1999).
53. M. Ester, H.-P. Kriegel, X. Xu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.
54. M. Hahsler, M. Piekenbrock, D. Doran, dbscan: Fast Density-Based Clustering with R. *J. Stat. Softw.* **91**, 1–30 (2019).

Materials and Methods

All studies began by obtaining informed consent from participants. 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, aspredicted.org/KSN_PNL). Any non-pre-registered analyses are labeled “post hoc” and any deviations from the pre-registrations are reported.

1. Study 1

1.1 Participants

We preregistered a target sample of 1,000 responses from CloudResearch’s Connect participant pool. In total, 1190 individuals began the survey (this includes 75 participants from a pilot conducted prior to the pre-registration; for completeness, we include these participants in our analyses, but excluding them does not qualitatively change the results). An initial (pre-treatment) 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 13 participants failed pre-treatment attention checks and were removed from the survey; 86 discontinued prior to reaching the treatment. Further, using preregistered criteria, we excluded 157 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 55 participants for whom the AI provided an inaccurate summary. Thus, 774 participants began the treatment. The overall attrition rate was 1.8%. Using a logistic regression model predicting whether or not a person attrited, we find no evidence of differential rates of attrition in treatment vs. control (b = -.53, p = .37). The flow of participants through the study is summarized in Figure S1.

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. This study was run on 19-22 January 2024 and took 30.98 minutes on average to complete.

1.2 Procedure

1.2.1 Pre-Treatment Measures

Participants completed a battery of self-report measures concerning their endorsement of a diverse set of 15 conspiracy beliefs, their attitudes concerning artificial intelligence, and demographic items including beliefs about politics and religion. Conspiracy beliefs were assessed using a modified version of Belief in Conspiracy Theories Inventory (*48*) (α = .90; example item “Government agencies in the UK are involved in the distribution of illegal drugs to ethnic minorities”), 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 = 38.6% (sd = 20.0%). Attitudes concerning artificial intelligence were adapted from a Pew survey (*49*).

Subsequently, participants responded to a non-directive, 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 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 S14 for exact wording of the 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”).

For all open-ended responses, including those in the Human-AI dialogues, the “paste” functionality was disabled to prevent automated responding.

1.2.2 Human - AI Dialogues

Following these pre-treatment measures, participants were informed they would be conversing with an advanced AI. To facilitate this real-time interaction within the Qualtrics survey platform, we used custom 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 (Figure S3), multiple paragraphs, and markdown formatting (e.g., lists and section headings). Thus, although each dialogue only lasted 3 rounds, the dialogues (a) represented 8.4 minutes of AI-human engagement on average and (b) 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 (a) discuss the American medical system, (b) debate with participants about whether they prefer dogs or cats, or (c) 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 13-14% 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. Similarly, a balance check found that our sample was balanced on pre-treatment covariates (see Table S15).

1.2.3 Post-Treatment Measures

Following the conversations, participants re-rated their belief in the focal conspiracy and then again completed the modified BCTI (α = .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 post-treatment 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 judgement (0 vs 1), yielding overlap-adjusted BCTI scores for pre-treatment (α = .90) and post-treatment (α = .92). Thirdly, we again filtered the BCTI item pool, this time retaining non-overlapping items that participants initially rated above 50% (more belief than "uncertain"), which allowed for pre-treatment (α = .90) and post-treatment (α = .90) overlap-adjusted BCTI scores for conspiracy theories that each participant actively endorses.

1.2.4 Recontacting at 10-Days and 2-Months

The participants from Study 1 were recontacted twice. The first recontact occurred 10 days (T3) after completing the intervention (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 pre-treatment belief in their chosen conspiracy (*t*[454] = 0.61, *p* = .544) or on the pre-treatment BCTI (*t*[454] = -0.71, *p* = .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 for T3, participants in the treatment who remained did not differ from those who dropped for either pre-treatment belief in their chosen conspiracy (*t*[450] = 0.02, *p* = .977) or on the pre-treatment BCTI (*t*[450] = -1.33, *p* = .183).

2. Study 2

For Study 2, two additional samples (Study 2a and 2b) were fielded from CloudConnect to corroborate, replicate, and extend our experimental findings. Although the majority of materials were identical across Studies 2a and 2b, we describe them separately because (a) we pre-registered separate rounds of data collection, (b) we used different phrasings for the behavioral outcome items, and (c) the data were collected several weeks apart. Particularly, we collected Study 2b due to 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.

2.1 Participants

In Study 2a, we preregistered (aspredicted.org/HSD_41Q) a target sample of 1,000 complete responses from CloudResearch's Connect participant pool, using quota-based sampling for age, race, ethnicity, and gender. A total of 1,427 individuals entered the survey, of whom 312 were redirected for using a cell phone, 30 failed the initial pre-treatment writing screener, 14 failed an attention check, and 104 discontinued prior to treatment, leaving 968. Of these participants, 218 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 = 668. Similarly, in Study 2b (aspredicted.org/KSN_PNL), we recruited 1545 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 prior to treatment, leaving 1309. Of these participants, 296 did not provide a genuine conspiracy theory and another 128 did not endorse their conspiracy above 50%, leaving a treatment sample of 885.

Thus, the full sample size across both rounds of Study 2 was n = 1553 (mean age = 41.9, mean ideology = 3.09) and included 670 males, 724 females, and 13 participants who selected another gender option. These studies were run on 25-28 February and 4-9 March 2024 and took 24.4 and 27.85 minutes on average to complete. The overall attrition rate was 3.7%. Using a logistic regression model predicting whether or not a person attrited, we find no evidence of differential rates of attrition in treatment vs. control (*b* = .02, *p* = .97).

2.2 Procedure

2.2.1 Pre-Treatment Measures

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 (i.e., those who did not answer the question) were redirected from the survey (see

Figure S2). Subsequently, participants completed self-report items about their artificial intelligence attitudes and demographic characteristics (mirroring those from Study 1). We did not administer the Belief in Conspiracy Theories Index in Study 2, and instead proceeded directly to the person-specific conspiracy assessment.

The wording of the person-specific instructions were modified slightly from Study 1 to (a) explicitly define the theories to be described and (b) only indirectly classify the theories as “conspiracies”. The initial question wording was:

> 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’s wording was:

> 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 pre-treatment 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”).

2.2.2 Post-Treatment Measures

Following the conversations, participants re-rated the focal conspiracy’s veracity and then completed a set of measures related to conspiracy-relevant behavior and trust. In both studies, we assessed (a) intentions to ignore or unfollow social media accounts espousing the focal conspiracy and (b) willingness to ignore or argue against people who believe the focal conspiracy; in our analyses of these items, we pool data across Studies 2a and 2b. Study 2a also asked about (c) willingness to engage in collective actions opposing the focal conspiracy, and (d) 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 do not analyze these items. Item c, concerning collective actions, was both counter-directionally worded (relative to the other items) and used a response scale containing negative and positive options that was *not* counter-directionally worded, potentially resulting in a confused pattern of results. Item d, 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 c and changing the wording of item d 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 indicates that many of them 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 if they wanted to sign a petition calling for greater government transparency about aliens - which plays into the conspiracy theory, rather than opposing it). This makes the choice of whether to sign the petition uninterpretable, and we do not include analysis of it.

In both Study 2a and 2b, participants then completed measures of general trust (1-item), personal trust (1-item), and institutional trust (5-items), which were adapted from the OECD Guidelines on Measuring Trust (*50*).

3. LLM-based Processing of Text Data

3.1 Conspiracy Theory Identification

In all studies, we used LLMs to filter participants who did not supply a genuine conspiracy theory. Specifically, we used GPT-4 Turbo to evaluate each participant’s free-response conspiracy statement using the following two prompts (model temperature = 0):

Prompt 1:

> “Determine if the following text contains or reflects a statement that, if endorsed, would indicate affirmative belief in a conspiracy theory (or something quite like a conspiracy theory). Respond only either “Conspiracy theory” or “Not conspiracy theory”.

Prompt 2:

> “Your task is to determine whether a given statement describes a conspiracy theory or not. A conspiracy theory is an explanation for an event or situation that invokes a conspiracy by powerful people or organizations, often without credible evidence. Conspiracy theories often involve claims of secret plots, coverups, or the manipulation of information by influential groups.
>
> Here are some examples of conspiracy theories:
>
> 1. "The moon landing was faked by the U.S. government to win the space race."
> 2. "The COVID-19 pandemic was planned and orchestrated by pharmaceutical companies to profit from vaccine sales."
> 3. "Climate change is a hoax perpetrated by scientists and politicians to gain funding and control the population."
>
> And here are some examples of statements that are not conspiracy theories:

> 4. "The Watergate scandal involved a cover-up of illegal activities by the Nixon administration."
> 5. "The tobacco industry concealed the harmful effects of smoking for many years."
> 6. "Corporate lobbying influences political decisions in favor of special interests."
>
> For each statement provided, respond with either "Conspiracy theory" or "Not conspiracy theory".

Prompt 2 clearly defined conspiracy theories and examples from which the models might inform their decision-making. In contrast, Prompt 1 was general, allowing GPT-4 to rely entirely on its internal definition of conspiracy theories. Agreement between the two prompts was high (**κ** = .67 [.63, .72]), with 70.3% of statements classified as a conspiracy theory under both prompts. Prompt 1 was more conservative than Prompt 2 (excluding 26% of responses vs. 21.4%). To adjudicate between the prompts, the first author coded a random subset of 200 theories, which revealed a stronger agreement with Prompt 1 (**κ** = .80 [.71, .89]) than Prompt 2 (**κ** = .64 [.53, .76]). Hence, we proceeded with Prompt 1.

3.2 Conspiracy Theory Classification

Each conspiracy statement was converted into text embeddings (numerical representations of text and concepts that capture their underlying semantic and syntactic structure). To ensure that the text embeddings reflected primarily the substantive content of each conspiracy theory rather than each participant's verbal abilities and linguistic preferences, we relied on the GPT-4 Turbo summarizations of each open-ended response (rather than the raw text entered by the participant). After stemming the words and removing English stopwords and punctuation, we used the *text-embedding-3-large* model–one of the best-performing models available to the public–to generate the embeddings (*51*). We then used cluster analytic algorithms to sort the embeddings into classes (see Analytic Procedures below).

3.3 Classifying the AI Model's Persuasion Strategies

To describe the persuasion strategies used by the AI model during the dialogues, we used GPT-4 Turbo to both generate candidate strategies (based on 10 API queries) and, in a separate set of API queries, to detect the presence and ubiquity of each candidate strategy in the dialogues. Particularly, we used the following prompt to identify plausible strategies:

> "If you, GPT-4 Turbo, were tasked with convincing a human being to stop believing in a specific conspiracy during an extended conversation (where you had been provided information about the human's particular conspiracy beliefs), which persuasive strategies would you use?"

Given a model temperature of 1, the AI model returned similar but non-overlapping sets of strategies in each query.

We next used GPT-4 Turbo to detect the presence of each strategy and the frequency with which they were used by the AI during each conversation. We instructed the AI to identify the 10 strategies mentioned in at least half of the previous model queries. The LLM was provided with a labeled transcript of each conversation and queried with the following prompt (model temperature = 0):

You are about to be shown the text of a written conversation about conspiracy theories. The two people in this conversation are a Debunker and a Believer. It is the role of the Debunker to convince the Believer that the Believer is wrong to hold a particular conspiracy theory. Each conversation will have 3 rounds.

Your job is to process the conversation and return a classification of the nature of each of the DEBUNKER'S responses. Particularly, you will determine whether the debunker's responses use each of the following persuasion strategies.

** Strategy List **

Build Rapport: Establish a respectful and understanding relationship with the Believer (e.g., to ensure the conversation is seen as a friendly exchange rather than a confrontation; demonstrating understanding and empathy towards the individuals beliefs without judgment).

Critical Thinking: Encourage the Believer to question and analyze the logic, evidence, and sources behind their beliefs, promoting a more analytical and reflective approach to information.

Alternative Explanations: Provide plausible, evidence-based alternative perspectives or explanations for events or phenomena that are attributed to conspiracy theories.

Harm: Discuss the personal or societal harms of the conspiracy beliefs.

Stories/Examples: Share stories, anecdotes, or real-world examples.

Encourage Empathy: Help the Believer consider the impact of conspiracy beliefs on others, fostering empathy and a broader perspective.

Socratic Questioning: Employ a questioning approach that leads the Believer to reflect on and examine the validity of their beliefs.

Conflicting Evidence: Introduce facts or data that directly contradict claims made by the conspiracy theory or the Believer.

Common Ground/Shared Reality: Identify and build on beliefs or values that the Debunker shares with the Believer.

Psychological Needs: Recognize and address the emotional aspects or psychological needs that may be underlying the Believers attraction to conspiracy theories, such as a desire for control or understanding.

Inconsistencies/Logical Fallacies: Identify and discuss logical inconsistencies or fallacies in the conspiracy theorys arguments.

Please be sure not to classify the responses of the Believer. Use the Believers responses only for context, so that you can understand the responses of the Debunker.

As the conversation follows 3 rounds, you should provide a rating for each strategy's presence in each round (i.e., 3 ratings per strategy).

Please format your ratings as JSON.

** Response Scale **

Use the following response scale for each rating:

> None: Strategy not used.
> Low: Strategy used rarely, in a limited fashion.
> Moderate: Strategy used repeatedly or with clear emphasis.
> High: Strategy used extensively and/or centrally throughout the response.

Thus, GPT-4 evaluated each strategy's prevalence in each conversational round. We plot strategies used in conversational round 1, to account for the endogenous proliferation of strategies based on particular participants' responses.

3.4 Accounting for Overlap between the BCTI Items and Focal Conspiracies

We defined a function that takes each participant's statement of conspiracy belief as input and determines whether it reflects an affirmative belief in any of 15 BCTI items (which are also statements of conspiracy theories) by sending the conspiracy theory text and a list of the 15 conspiracy theories to GPT-4 Turbo and getting a response in the form of a string of 15 0s and 1s (where 1 indicates an affirmative belief in that particular conspiracy theory). We then calculated average scores for each participant's non-overlapping BCTI items. The number of overlapping conspiracy theories ranged from 0 - 3 (*M* = .71). The GPT-4 prompt was as follows:

> Determine if the presented text reflects an affirmative belief in any of the following 15 conspiracy theories:
>
> Conspiracy 1: A powerful and secretive group, known as the New World Order, are planning to eventually rule the world through an autonomous world government, which would replace sovereign governments.
> Conspiracy 2: COVID-19 was produced under laboratory conditions by the Chinese government.
> Conspiracy 3: The US government had foreknowledge about the Japanese attack on Pearl Harbor, but allowed the attack to take place so as to be able to enter the Second World War.
> Conspiracy 4: US agencies intentionally created the AIDS epidemic and administered it to Black and gay men in the 1970s.
> Conspiracy 5: The assassination of Martin Luther King, Jr., was the result of an organised conspiracy by US government agencies such as the CIA and FBI.
> Conspiracy 6: The Apollo moon landings never happened and were staged in a Hollywood film studio.
> Conspiracy 7: Area 51 in Nevada, US, is a secretive military base that contains hidden alien spacecraft and/or alien bodies.
> Conspiracy 8: The US government allowed the 9/11 attacks to take place so that it would have an excuse to achieve foreign (e.g., wars in Afghanistan and Iraq) and domestic (e.g., attacks on civil liberties) goals that had been determined prior to the attacks.
> Conspiracy 9: The assassination of John F. Kennedy was not committed by the lone gunman, Lee Harvey Oswald, but was rather a detailed, organised conspiracy to kill the President.
> Conspiracy 10: In July 1947, the US military recovered the wreckage of an alien craft from Roswell, New Mexico, and covered up the fact.
> Conspiracy 11: Princess Diana's death was not an accident, but rather an organised assassination by members of the British royal family who disliked her.
> Conspiracy 12: The Oklahoma City bombers, Timothy McVeigh and Terry Nichols, did not act alone, but rather received assistance from neo-Nazi groups.
> Conspiracy 13: The Coca Cola company intentionally changed to an inferior formula with the intent of driving up demand for their classic product, later reintroducing it for their financial gain.
> Conspiracy 14: Special interest groups are suppressing, or have suppressed in the past, technologies that could provide energy at reduced cost or reduced pollution output.
> Conspiracy 15: Government agencies in the UK are involved in the distribution of illegal drugs to ethnic minorities.
>
> Format your answer as a list of EXACTLY 15 0s or 1s. Do not use spaces or commas in your answer. For example, 101000000000001 would be an acceptable response'

4. Ethics

Participants gave informed consent. 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).

5. Analysis Procedure

5.1 Estimating the Treatment Effect on Focal and Non-focal Conspiracy Beliefs

These analyses were preregistered. We fitted a linear regression with the Lin (2013) covariate adjustment and HC2 standard errors to test the overarching impact of the intervention (treatment condition vs. control conditions) on conspiracy beliefs. That is, to obtain the direct effect of condition, we included pre-treatment belief in the conspiracy theory as a covariate (i.e., linear regression with one covariate and the condition dummy). We report the untransfomed beta estimate, 95% confidence interval, and p-value. Further, we report d_{ppc2} (*52*), an effect size estimate that uses the pooled pretest standard deviation for weighting the differences of the pre-post means. This analysis was repeated for belief in the focal conspiracy, and average belief across all non-focal conspiracies from the BCTI.

To evaluate the durability of the treatment at follow-up, we specified a linear mixed model (estimated using REML) to predict conspiracy belief (one observation per user-time point) with experimental condition and time point (formula: ConspiracyBelief ~ Treatment [vs. Control] * TimePoint). The model included random intercepts on participant. Durability over time was assessed within the treatment condition via pairwise comparisons between (a) Time 1 (pre-treatment) and each post-treatment timepoint and (b) Time 2 (immediately following treatment) and Times 3 and 4 (10-days and 2-months). However, we also evaluated differences between the treatment and control groups at each timepoint to account for regression to the mean.

5.2 Conspiracy Theory Clustering

The text embeddings were first subjected to PCA. Enough components were retained to capture 85% of the total variance in the data (k = 352). This threshold was chosen to maintain a balance between reducing dimensionality and preserving the data's inherent structure. We next applied the density-based spatial clustering of applications with noise (DBSCAN) algorithm to a cosine distance matrix of the reduced embeddings (*53*, *54*). DBSCAN was selected because of its ability to identify clusters of arbitrary shape and its robustness to outliers, making it suitable for text data which often contains noise and irregular cluster patterns. Further, density based clustering accounts for “noise” (i.e., some points are not assigned a cluster label). We used an ε value of 3 (i.e., the distance parameter that defines the radius around a data point to search for neighboring points) and specified the minimum number of points required to form a single cluster as 15, considering border points. This approach identified 14 distinct clusters and labeled 594 points as noise (representing conspiracy theories that do not fit well into any cluster). The clustering result shows a diverse distribution of points across clusters, with the largest cluster (JFK conspiracies) containing 310 points and the smallest (government surveillance conspiracies) containing 15 points. The noise points constitute a significant portion of the dataset (29%), underscoring the variability of focal conspiracy theories in this population.

Given that the intention of this analysis was to identify coherent, interpretable cleavages in the universe of potential conspiracy theories – rather than identify the genuine structure of conspiracy theories in the population – we primarily evaluated the clustering results based on substantive similarities in the conspiracy statements belonging to each cluster. Qualitatively examining the conspiracies assigned to each cluster illustrated readily apparent similarities across virtually all statements in each cluster, leading us to retain the 14-cluster DBSCAN solution.

To test whether the identified clusters moderate the AI-driven treatment, we integrated cluster membership (including noise points as a distinct cluster) as a categorical variable into our treatment effect model (formula: [Post-treatment conspiracy belief] ~ [Experimental condition] * [DBSCAN cluster] + [Pre-teatment conspiracy belief]) and tested the significance of the relative improvement in fit attributable to the interactions between experimental condition and each cluster using an analysis of variance. Clusters with < 1% membership were removed from this analysis to increase statistical power. To test the statistical significance of the treatment for members of each cluster, we used pairwise comparisons (for differences between the treatment vs. control conditions of each cluster) with a Bonferroni p-value adjustment.

5.3 Individual Difference Moderators

To investigate the effect of individual differences on the treatment effect, particularly among participants with deeply rooted beliefs, we used a combination of generalized additive models (GAMs), multiple linear regression models, and post-hoc causal forests.

The use of GAMs was specifically aimed at uncovering sharp reductions in treatment efficacy among committed conspiracy believers via non-linear interactions, so we focused on pre-treatment focal and non-focal conspiracy beliefs, as well as the perceived importance of the focal conspiracy. Each GAM analysis began with a base model where post-treatment belief was predicted using the experimental condition factor, a smooth term for pre-treatment specific beliefs (to mirror the covariate-adjustment used in the main effect model), and a smooth term for the relevant moderator. We then specified an interaction GAM, which incorporated a term that allowed the smooth effect of the relevant moderator to vary by experimental condition. All GAMs were fitted with the REML smoothing parameter. Comparative analyses between the base and interaction models were based on an analysis of deviance (via the gam::anova.gam function), though we also report AIC (Akaike Information Criterion) values and R^2.

To further understand individual differences that best explained variation in treatment effects, a large multiple linear regression model was employed. The primary model, pooling across participants from all studies, included linear interactions between the experimental condition and all key predictors shared across samples: pre-treatment specific beliefs, familiarity with generative AI, usage and trust in generative AI, religiosity, partisanship, extremism, age, type of conspiracy belief, education level, race, and gender. We also specified additional, sample-specific regressions that included intellectual humility and actively open-minded thinking (in Study 1) and personal, general, and institutional trust (in Study 2). We estimated these models using OLS with HC2 robust standard errors.

Finally, we deployed causal forests, a machine learning technique for heterogeneous treatment effect estimation, with the grf package in R. Causal forest models were trained based on: (a) the full, combined sample, using only covariates shared across sample, (b) Study 1, using all covariates available, and (b) Study 2, using all covariates available. The causal forests were trained with 100,000 trees and all tunable parameters tuned by cross-validation. We used the grf package's summary function, test_calibration, to assess the forests' goodness of fit on held-out data. Variable importance scores are reported to highlight which moderators were most influential in affecting treatment outcomes. Estimates derived from the causal forest model representing the expected effect of the treatment for each individual (i.e., the conditional average treatment effect; CATE) are reported as a function of various moderators (i.e., those with high importance or particular relevance to conspiracy beliefs).

5.4 Estimating Treatment Effects for Behavioral Indicators

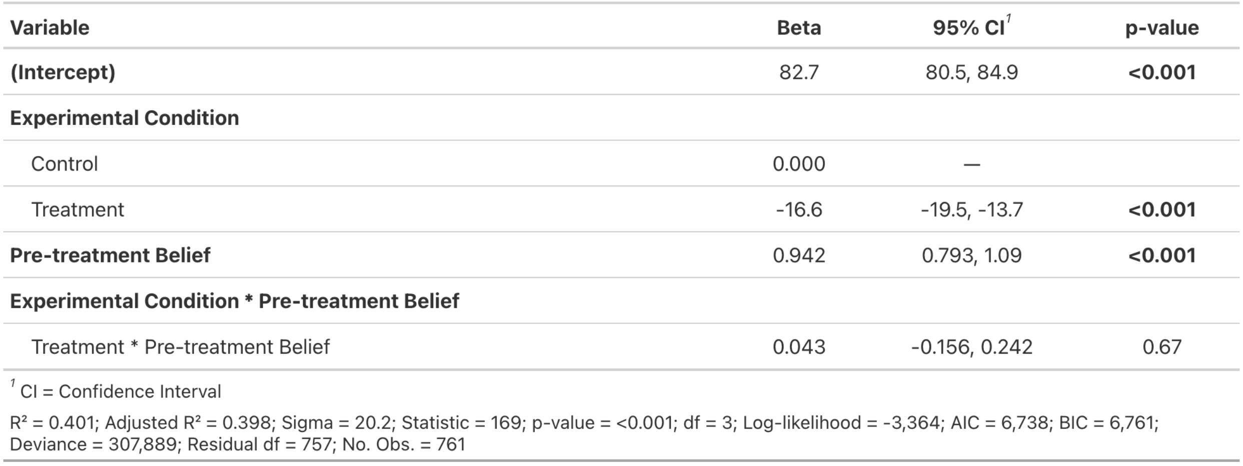
To evaluate the impact of the treatment on behavioral indicators, we used OLS with HC2 robust standard errors (i.e., formula: DV ~ ExperimentalCondition). Given that the behavioral outcome variables are on differing Likert-type scales, we report standardized beta coefficients with 95% confidence intervals to facilitate comparisons across the DVs.

Supplemental Results

Treatment Effects

Table S1. The effect of AI - human conversations on focal conspiracy beliefs in Sample 1.

We fitted a linear model (estimated using OLS) to predict Post_Belief_Specific with Experimental Condition and Pre-treatment Belief (formula: Post_Belief_Specific ~ `Experimental Condition` * `Pre-treatment Belief`). The model explains a statistically significant and substantial proportion of variance (R2 = 0.40, F(3, 757) = 168.75, p < .001, adj. R2 = 0.40). The model's intercept corresponds to Experimental Condition = Control and Pre-treatment Belief = 83.76.



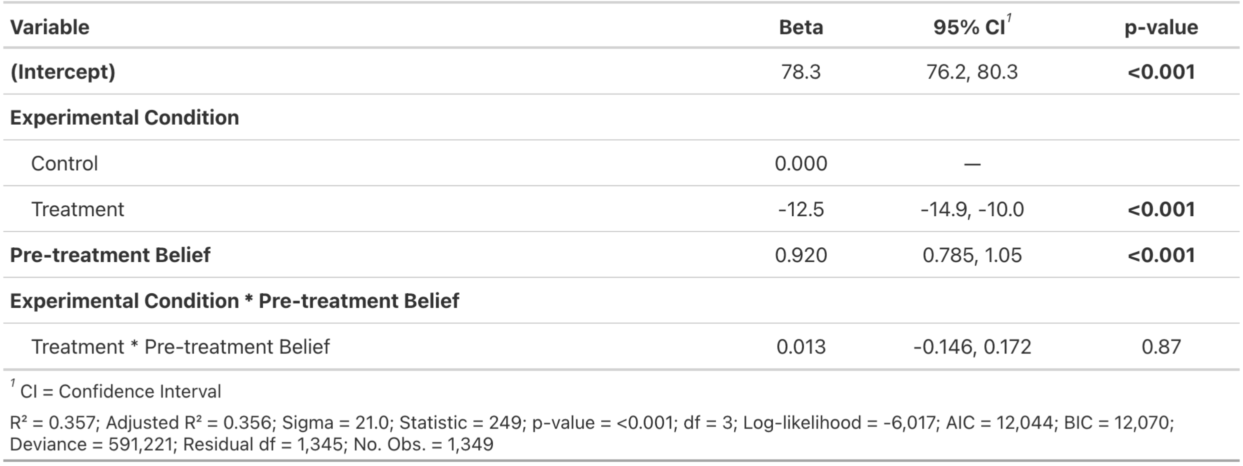
Variable	Beta	95% CI[1]	p-value
(Intercept)	82.7	80.5, 84.9	**<0.001**
Experimental Condition			
Control	0.000	—	
Treatment	-16.6	-19.5, -13.7	**<0.001**
Pre-treatment Belief	0.942	0.793, 1.09	**<0.001**
Experimental Condition * Pre-treatment Belief			
Treatment * Pre-treatment Belief	0.043	-0.156, 0.242	0.67

[1] CI = Confidence Interval

R^2 = 0.401; Adjusted R^2 = 0.398; Sigma = 20.2; Statistic = 169; p-value = <0.001; df = 3; Log-likelihood = -3,364; AIC = 6,738; BIC = 6,761; Deviance = 307,889; Residual df = 757; No. Obs. = 761

Table S2. The effect of AI - human conversations on focal conspiracy beliefs in Sample 2.

We fitted a linear model (estimated using OLS) to predict Post_Belief_Specific with Experimental Condition and Pre-treatment Belief (formula: Post_Belief_Specific ~ `Experimental Condition` * `Pre-treatment Belief`). The model explains a statistically significant and substantial proportion of variance (R2 = 0.36, F(3, 1345) = 249.43, p < .001, adj. R2 = 0.36). The model's intercept corresponds to Experimental Condition = Control and Pre-treatment Belief = 81.23.



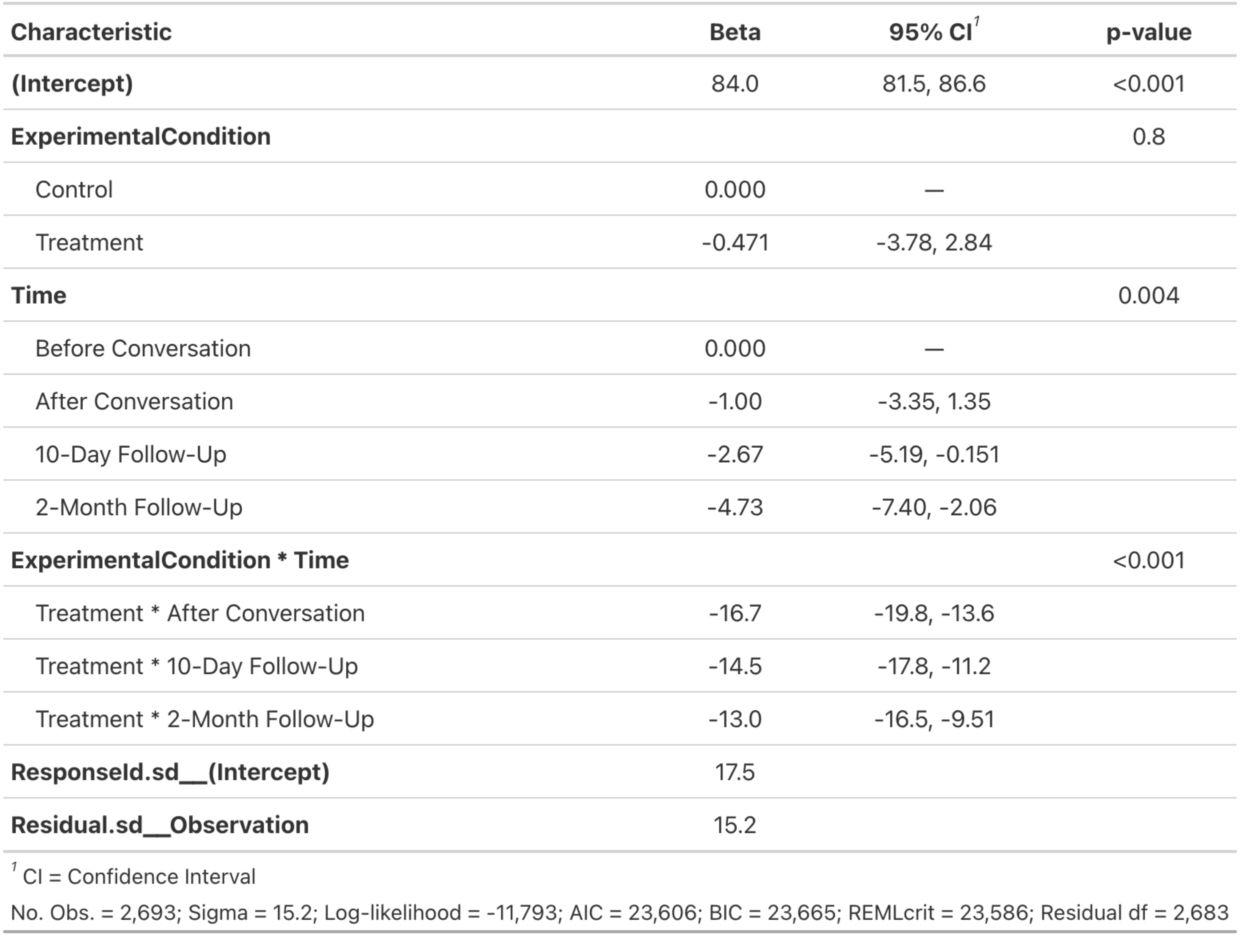
Variable	Beta	95% CI[1]	p-value
(Intercept)	78.3	76.2, 80.3	**<0.001**
Experimental Condition			
Control	0.000	—	
Treatment	-12.5	-14.9, -10.0	**<0.001**
Pre-treatment Belief	0.920	0.785, 1.05	**<0.001**
Experimental Condition * Pre-treatment Belief			
Treatment * Pre-treatment Belief	0.013	-0.146, 0.172	0.87

[1] CI = Confidence Interval

R^2 = 0.357; Adjusted R^2 = 0.356; Sigma = 21.0; Statistic = 249; p-value = <0.001; df = 3; Log-likelihood = -6,017; AIC = 12,044; BIC = 12,070; Deviance = 591,221; Residual df = 1,345; No. Obs. = 1,349

Table S3. The effect of AI - human conversations on focal conspiracy beliefs over time (Sample 1).

We fitted a linear mixed model (estimated using REML) to predict ConspiracyBelief with ExperimentalCondition and Time (formula: ConspiracyBelief ~ ExperimentalCondition * Time). The model included ResponseId as random effect (formula: ~1 | ResponseId). The model's total explanatory power is substantial (conditional R2 = 0.62) and the part related to the fixed effects alone (marginal R2) is of 0.11. The model's intercept corresponds to ExperimentalCondition = Control and Time = Before Conversation.



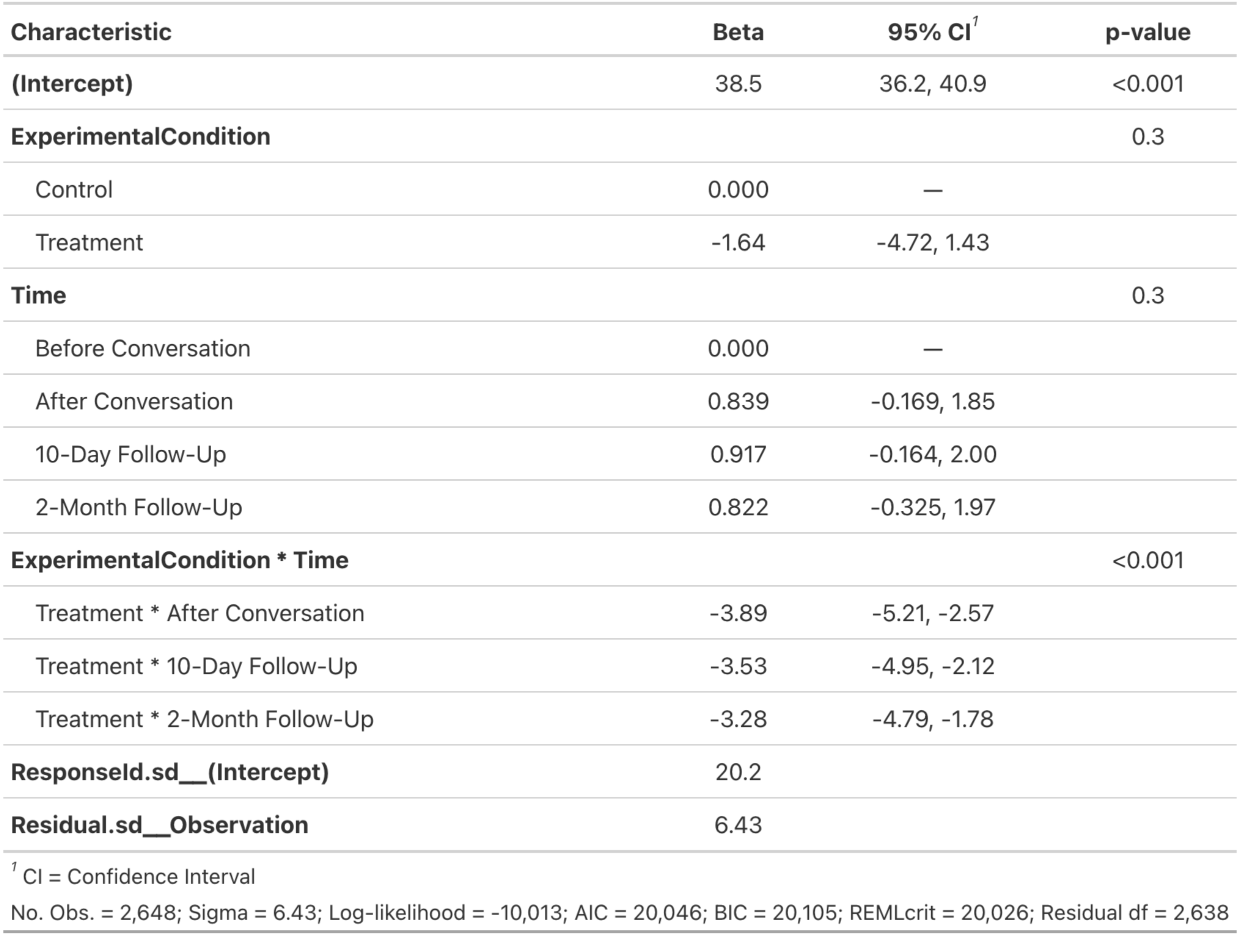
Characteristic	**Beta**	**95% CI**[1]	**p-value**
(Intercept)	84.0	81.5, 86.6	<0.001
ExperimentalCondition			0.8
Control	0.000	—	
Treatment	-0.471	-3.78, 2.84	
Time			0.004
Before Conversation	0.000	—	
After Conversation	-1.00	-3.35, 1.35	
10-Day Follow-Up	-2.67	-5.19, -0.151	
2-Month Follow-Up	-4.73	-7.40, -2.06	
ExperimentalCondition * Time			<0.001
Treatment * After Conversation	-16.7	-19.8, -13.6	
Treatment * 10-Day Follow-Up	-14.5	-17.8, -11.2	
Treatment * 2-Month Follow-Up	-13.0	-16.5, -9.51	
ResponseId.sd__(Intercept)	17.5		
Residual.sd__Observation	15.2		

[1] CI = Confidence Interval

No. Obs. = 2,693; Sigma = 15.2; Log-likelihood = -11,793; AIC = 23,606; BIC = 23,665; REMLcrit = 23,586; Residual df = 2,683

Table S4. The effect of AI - human conversations on conspiracy beliefs that do not overlap with the focal conspiracy.

We fitted a linear mixed model (estimated using REML) to predict ConspiracyBelief with ExperimentalCondition and Time (formula: ConspiracyBelief ~ ExperimentalCondition * Time). The model included Response ID as random effect (formula: ~1 | ResponseId). The model's total explanatory power is substantial (conditional R2 = 0.91) and the part related to the fixed effects alone (marginal R2) is of 0.01. The model's intercept corresponds to ExperimentalCondition = Control and Time = Before Conversation.



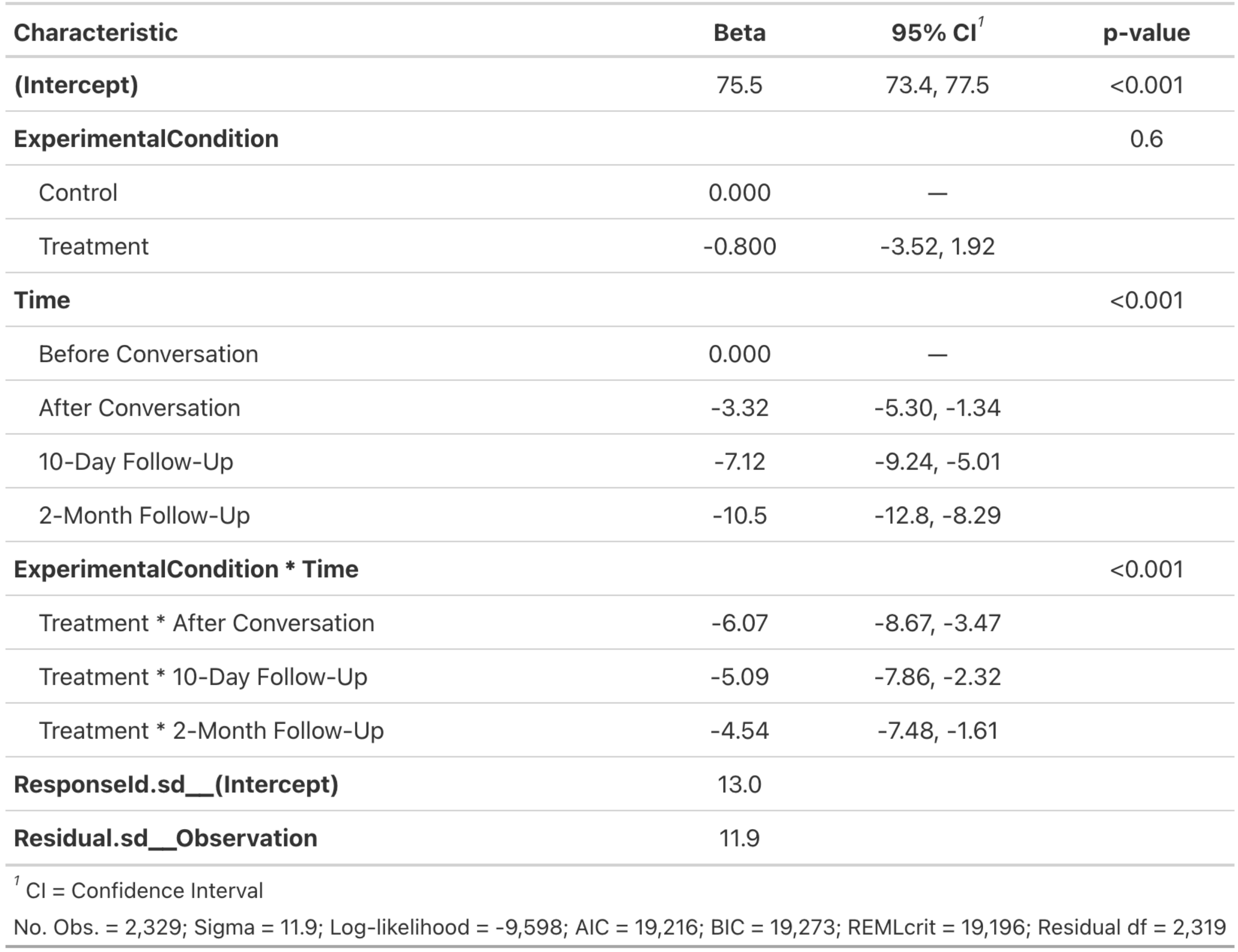
Characteristic	**Beta**	**95% CI**[1]	**p-value**
(Intercept)	38.5	36.2, 40.9	<0.001
ExperimentalCondition			0.3
Control	0.000	—	
Treatment	-1.64	-4.72, 1.43	
Time			0.3
Before Conversation	0.000	—	
After Conversation	0.839	-0.169, 1.85	
10-Day Follow-Up	0.917	-0.164, 2.00	
2-Month Follow-Up	0.822	-0.325, 1.97	
ExperimentalCondition * Time			<0.001
Treatment * After Conversation	-3.89	-5.21, -2.57	
Treatment * 10-Day Follow-Up	-3.53	-4.95, -2.12	
Treatment * 2-Month Follow-Up	-3.28	-4.79, -1.78	
ResponseId.sd__(Intercept)	20.2		
Residual.sd__Observation	6.43		

[1] CI = Confidence Interval

No. Obs. = 2,648; Sigma = 6.43; Log-likelihood = -10,013; AIC = 20,046; BIC = 20,105; REMLcrit = 20,026; Residual df = 2,638

Table S5. The effect of AI - human conversations on conspiracy beliefs that do not overlap with the focal conspiracy and were endorsed > 50 / 100.

We fitted a linear mixed model (estimated using REML) to predict ConspiracyBelief with ExperimentalCondition and Time (formula: ConspiracyBelief ~ ExperimentalCondition * Time). The model included Response ID as random effect (formula: ~1 | ResponseId). The model's total explanatory power is substantial (conditional R2 = 0.59) and the part related to the fixed effects alone (marginal R2) is of 0.09. The model's intercept corresponds to ExperimentalCondition = Control and Time = Before Conversation.



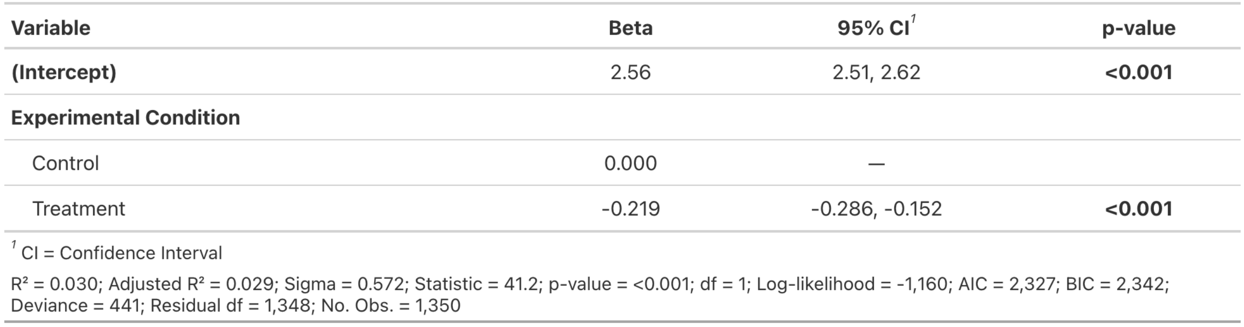
Characteristic	**Beta**	**95% CI**[1]	**p-value**
(Intercept)	75.5	73.4, 77.5	<0.001
ExperimentalCondition			0.6
Control	0.000	—	
Treatment	-0.800	-3.52, 1.92	
Time			<0.001
Before Conversation	0.000	—	
After Conversation	-3.32	-5.30, -1.34	
10-Day Follow-Up	-7.12	-9.24, -5.01	
2-Month Follow-Up	-10.5	-12.8, -8.29	
ExperimentalCondition * Time			<0.001
Treatment * After Conversation	-6.07	-8.67, -3.47	
Treatment * 10-Day Follow-Up	-5.09	-7.86, -2.32	
Treatment * 2-Month Follow-Up	-4.54	-7.48, -1.61	
ResponseId.sd__(Intercept)	13.0		
Residual.sd__Observation	11.9		

[1] CI = Confidence Interval

No. Obs. = 2,329; Sigma = 11.9; Log-likelihood = -9,598; AIC = 19,216; BIC = 19,273; REMLcrit = 19,196; Residual df = 2,319

Table S6. The effect of AI - human conversations on reaction to conspiracy posters on social media.

We fitted a linear model (estimated using OLS) to predict reactions to conspiracy posters (response scale = 1-3) with experimental condition. The model's intercept corresponds to experimental condition = Control. Within this model:



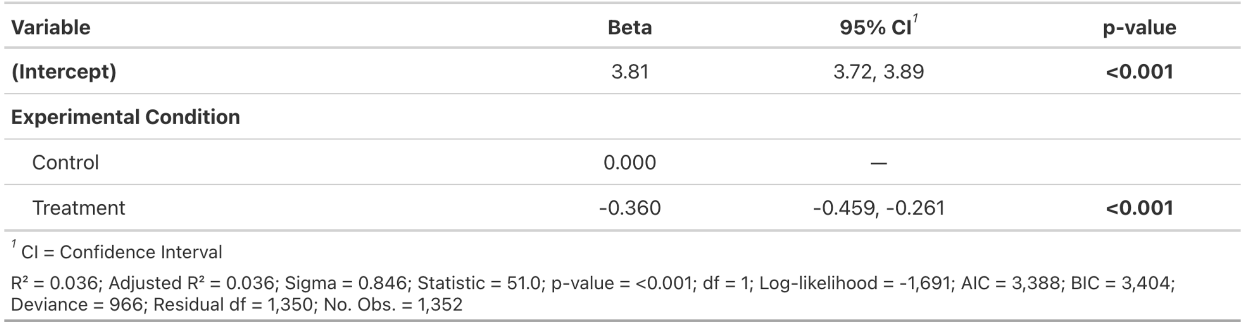
Variable	Beta	95% CI[1]	p-value
(Intercept)	2.56	2.51, 2.62	**<0.001**
Experimental Condition			
Control	0.000	—	
Treatment	-0.219	-0.286, -0.152	**<0.001**

[1] CI = Confidence Interval

R² = 0.030; Adjusted R² = 0.029; Sigma = 0.572; Statistic = 41.2; p-value = <0.001; df = 1; Log-likelihood = -1,160; AIC = 2,327; BIC = 2,342; Deviance = 441; Residual df = 1,348; No. Obs. = 1,350

Table S7. The effect of AI - human conversations on reaction to discussions with believers.

We fitted a linear model (estimated using OLS) to predict reactions to discussions with conspiracy believers (response scale = 1-5) with experimental condition. The model's intercept corresponds to experimental condition = Control. Within this model:



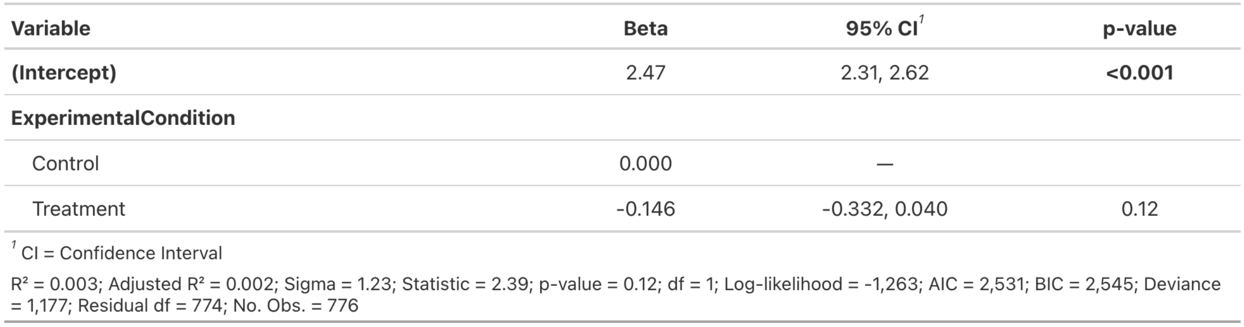
Variable	Beta	95% CI[1]	p-value
(Intercept)	3.81	3.72, 3.89	**<0.001**
Experimental Condition			
Control	0.000	—	
Treatment	-0.360	-0.459, -0.261	**<0.001**

[1] CI = Confidence Interval

R² = 0.036; Adjusted R² = 0.036; Sigma = 0.846; Statistic = 51.0; p-value = <0.001; df = 1; Log-likelihood = -1,691; AIC = 3,388; BIC = 3,404; Deviance = 966; Residual df = 1,350; No. Obs. = 1,352

Table S8. The effect of AI - human conversations on willingness to join protests supporting the focal conspiracy.

We fitted a linear model (estimated using OLS) to predict willingness to join protests supporting participants’ focal conspiracy (response scale = 1-5) with experimental condition. The model's intercept corresponds to experimental condition = Control. Within this model:



Variable	**Beta**	**95% CI**[1]	**p-value**
(Intercept)	2.47	2.31, 2.62	**<0.001**
ExperimentalCondition			
Control	0.000	—	
Treatment	-0.146	-0.332, 0.040	0.12

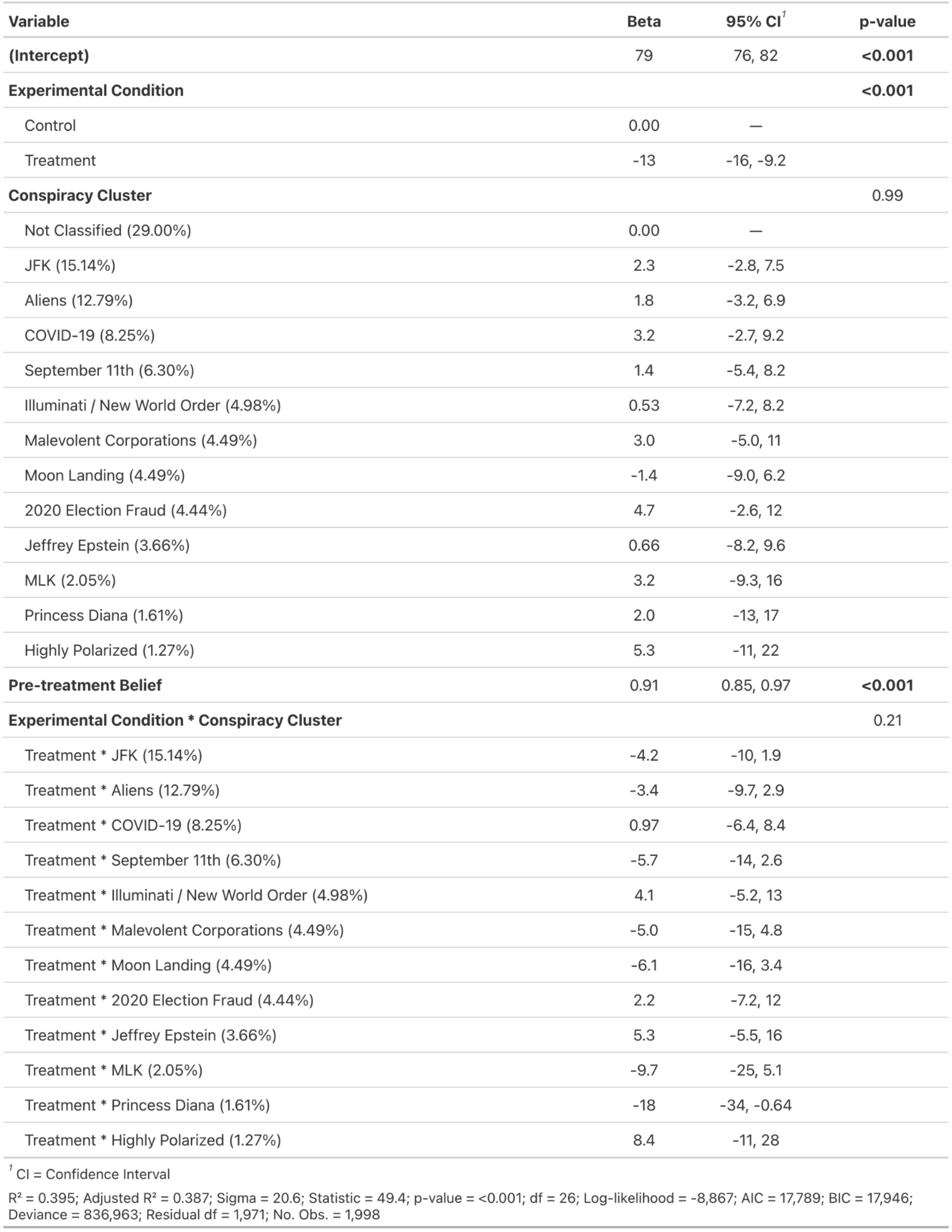
[1] CI = Confidence Interval

R^2 = 0.003; Adjusted R^2 = 0.002; Sigma = 1.23; Statistic = 2.39; p-value = 0.12; df = 1; Log-likelihood = -1,263; AIC = 2,531; BIC = 2,545; Deviance = 1,177; Residual df = 774; No. Obs. = 776

Moderators

Table S9. The effect of AI - human conversations on focal conspiracy beliefs by type of conspiracy theory (based on a density-based spatial clustering algorithm)

We fitted a linear model (estimated using OLS) to predict Post_Belief_Specific with Experimental Condition, Conspiracy Cluster and Pre-treatment Belief (formula: Post_Belief_Specific ~ `Experimental Condition` * `Conspiracy Cluster` + `Pre-treatment Belief`). The model explains a statistically significant and substantial proportion of variance (R2 = 0.39, F(26, 1971) = 49.44, p < .001, adj. R2 = 0.39). The model's intercept corresponds to Experimental Condition = Control, Conspiracy Cluster = Not Classified (29.00%) and Pre-treatment Belief = 82.16. Within this model:



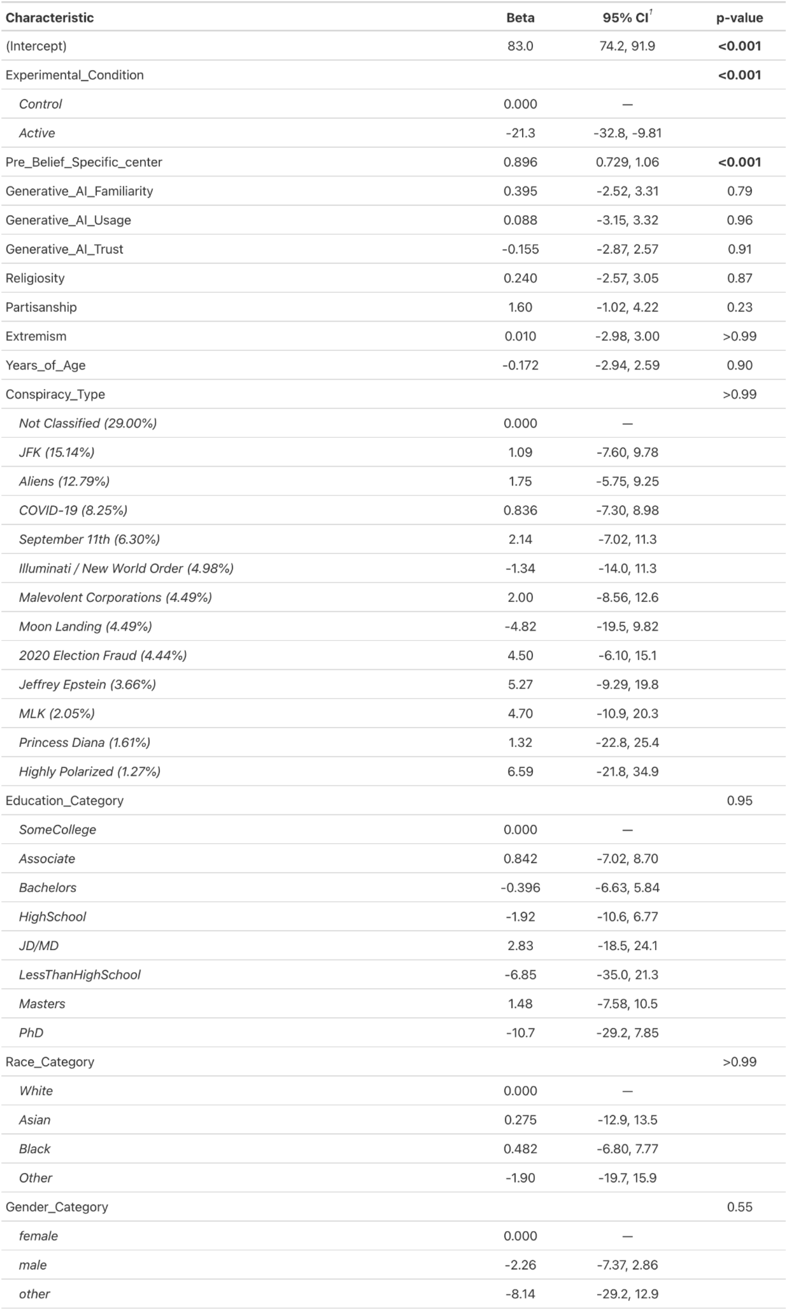
Variable	Beta	95% CI[1]	p-value
(Intercept)	79	76, 82	**<0.001**
Experimental Condition			**<0.001**
Control	0.00	—	
Treatment	-13	-16, -9.2	
Conspiracy Cluster			0.99
Not Classified (29.00%)	0.00	—	
JFK (15.14%)	2.3	-2.8, 7.5	
Aliens (12.79%)	1.8	-3.2, 6.9	
COVID-19 (8.25%)	3.2	-2.7, 9.2	
September 11th (6.30%)	1.4	-5.4, 8.2	
Illuminati / New World Order (4.98%)	0.53	-7.2, 8.2	
Malevolent Corporations (4.49%)	3.0	-5.0, 11	
Moon Landing (4.49%)	-1.4	-9.0, 6.2	
2020 Election Fraud (4.44%)	4.7	-2.6, 12	
Jeffrey Epstein (3.66%)	0.66	-8.2, 9.6	
MLK (2.05%)	3.2	-9.3, 16	
Princess Diana (1.61%)	2.0	-13, 17	
Highly Polarized (1.27%)	5.3	-11, 22	
Pre-treatment Belief	0.91	0.85, 0.97	**<0.001**
Experimental Condition * Conspiracy Cluster			0.21
Treatment * JFK (15.14%)	-4.2	-10, 1.9	
Treatment * Aliens (12.79%)	-3.4	-9.7, 2.9	
Treatment * COVID-19 (8.25%)	0.97	-6.4, 8.4	
Treatment * September 11th (6.30%)	-5.7	-14, 2.6	
Treatment * Illuminati / New World Order (4.98%)	4.1	-5.2, 13	
Treatment * Malevolent Corporations (4.49%)	-5.0	-15, 4.8	
Treatment * Moon Landing (4.49%)	-6.1	-16, 3.4	
Treatment * 2020 Election Fraud (4.44%)	2.2	-7.2, 12	
Treatment * Jeffrey Epstein (3.66%)	5.3	-5.5, 16	
Treatment * MLK (2.05%)	-9.7	-25, 5.1	
Treatment * Princess Diana (1.61%)	-18	-34, -0.64	
Treatment * Highly Polarized (1.27%)	8.4	-11, 28	

[1] CI = Confidence Interval

R^2 = 0.395; Adjusted R^2 = 0.387; Sigma = 20.6; Statistic = 49.4; p-value = <0.001; df = 26; Log-likelihood = -8,867; AIC = 17,789; BIC = 17,946; Deviance = 836,963; Residual df = 1,971; No. Obs. = 1,998

Table S10. The effect of AI - human conversations on focal conspiracy beliefs by all covariates in Sample 1.

We fitted a linear model (estimated using OLS) to predict Post_Belief_Specific with Experimental_Condition, Pre_Belief_Specific_center, and all covariates in Sample 1. The model explains a statistically significant and substantial proportion of variance (R2 = 0.49, F(68, 652) = 9.38, p < .001, adj. R2 = 0.44). The model's intercept corresponds to Experimental_Condition = Control, Pre_Belief_Specific_center = 83.83, mean levels of all continuous covariates, Conspiracy_Type = Not Classified (29.00%), Education_Category = SomeCollege, Race_Category = White and Gender_Category = Female. **Within this model, the main effects are as follows:**



Characteristic	**Beta**	**95% CI**^{1}	**p-value**
(Intercept)	83.0	74.2, 91.9	**<0.001**
Experimental_Condition			**<0.001**
Control	0.000	—	
Active	-21.3	-32.8, -9.81	
Pre_Belief_Specific_center	0.896	0.729, 1.06	**<0.001**
Generative_AI_Familiarity	0.395	-2.52, 3.31	0.79
Generative_AI_Usage	0.088	-3.15, 3.32	0.96
Generative_AI_Trust	-0.155	-2.87, 2.57	0.91
Religiosity	0.240	-2.57, 3.05	0.87
Partisanship	1.60	-1.02, 4.22	0.23
Extremism	0.010	-2.98, 3.00	>0.99
Years_of_Age	-0.172	-2.94, 2.59	0.90
Conspiracy_Type			>0.99
Not Classified (29.00%)	0.000	—	
JFK (15.14%)	1.09	-7.60, 9.78	
Aliens (12.79%)	1.75	-5.75, 9.25	
COVID-19 (8.25%)	0.836	-7.30, 8.98	
September 11th (6.30%)	2.14	-7.02, 11.3	
Illuminati / New World Order (4.98%)	-1.34	-14.0, 11.3	
Malevolent Corporations (4.49%)	2.00	-8.56, 12.6	
Moon Landing (4.49%)	-4.82	-19.5, 9.82	
2020 Election Fraud (4.44%)	4.50	-6.10, 15.1	
Jeffrey Epstein (3.66%)	5.27	-9.29, 19.8	
MLK (2.05%)	4.70	-10.9, 20.3	
Princess Diana (1.61%)	1.32	-22.8, 25.4	
Highly Polarized (1.27%)	6.59	-21.8, 34.9	
Education_Category			0.95
SomeCollege	0.000	—	
Associate	0.842	-7.02, 8.70	
Bachelors	-0.396	-6.63, 5.84	
HighSchool	-1.92	-10.6, 6.77	
JD/MD	2.83	-18.5, 24.1	
LessThanHighSchool	-6.85	-35.0, 21.3	
Masters	1.48	-7.58, 10.5	
PhD	-10.7	-29.2, 7.85	
Race_Category			>0.99
White	0.000	—	
Asian	0.275	-12.9, 13.5	
Black	0.482	-6.80, 7.77	
Other	-1.90	-19.7, 15.9	
Gender_Category			0.55
female	0.000	—	
male	-2.26	-7.37, 2.86	
other	-8.14	-29.2, 12.9	

And the interactions between experimental condition and each covariate are as follows:

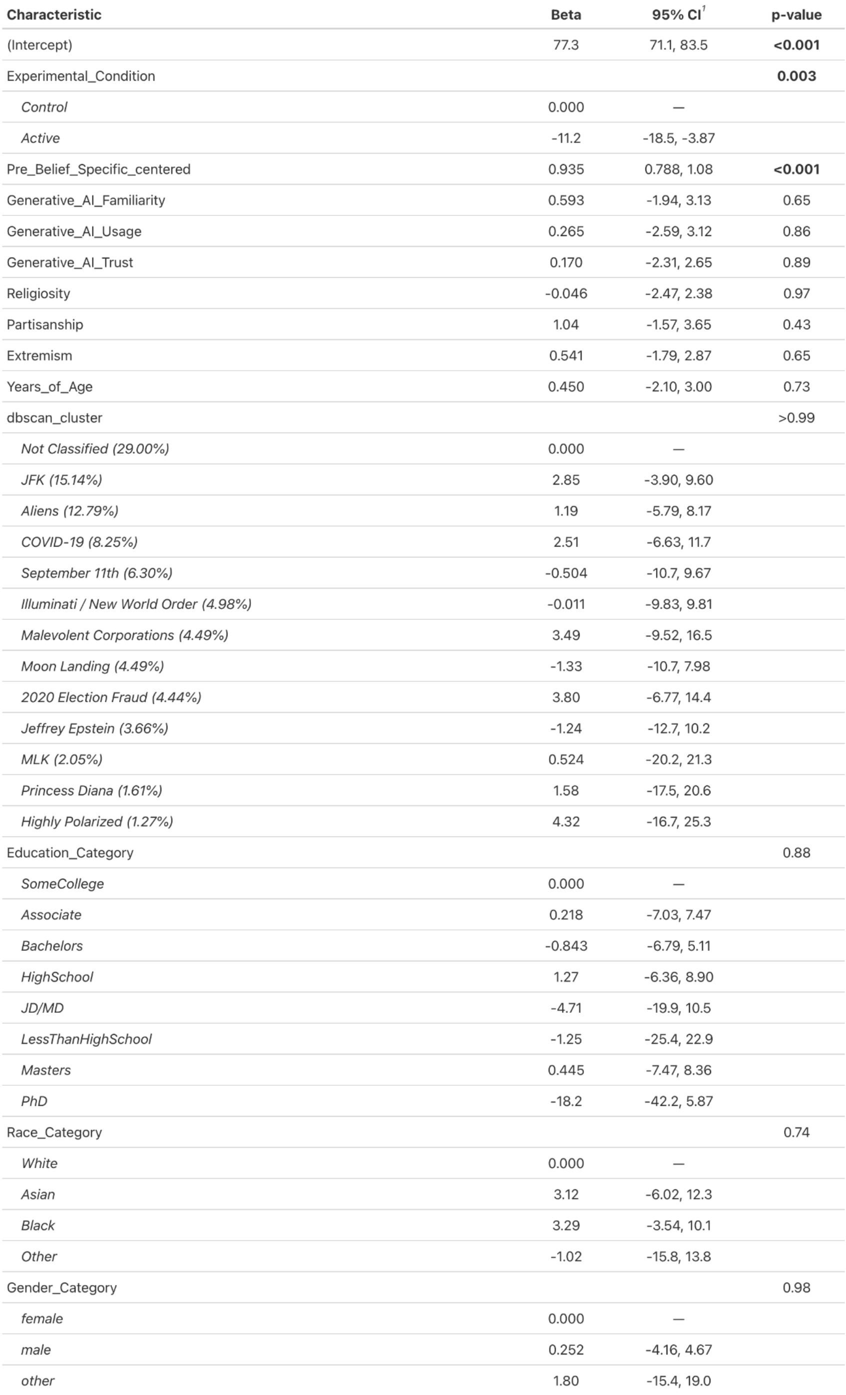
Experimental_Condition * Pre_Belief_Specific_center			0.85
*Active * Pre_Belief_Specific_center*	-0.021	-0.238, 0.196	
Experimental_Condition * Generative_AI_Familiarity			0.42
*Active * Generative_AI_Familiarity*	1.52	-2.21, 5.25	
Experimental_Condition * Generative_AI_Usage			0.38
*Active * Generative_AI_Usage*	1.85	-2.28, 5.97	
Experimental_Condition * Generative_AI_Trust			**0.004**
*Active * Generative_AI_Trust*	-5.26	-8.78, -1.73	
Experimental_Condition * Religiosity			0.74
*Active * Religiosity*	-0.602	-4.15, 2.94	
Experimental_Condition * Partisanship			0.76
*Active * Partisanship*	0.563	-3.03, 4.15	
Experimental_Condition * Extremism			0.44
*Active * Extremism*	1.57	-2.41, 5.56	
Experimental_Condition * Years_of_Age			**0.003**
*Active * Years_of_Age*	5.51	1.93, 9.10	
Experimental_Condition * Conspiracy_Type			0.47
*Active * JFK (15.14%)*	-9.48	-20.3, 1.31	
*Active * Aliens (12.79%)*	-4.83	-14.6, 4.94	
*Active * COVID-19 (8.25%)*	-0.533	-11.1, 10.0	
*Active * September 11th (6.30%)*	-12.6	-24.8, -0.439	
*Active * Illuminati / New World Order (4.98%)*	10.7	-8.39, 29.7	
*Active * Malevolent Corporations (4.49%)*	-5.00	-18.4, 8.43	
*Active * Moon Landing (4.49%)*	2.59	-15.1, 20.2	
*Active * 2020 Election Fraud (4.44%)*	2.54	-13.5, 18.5	
*Active * Jeffrey Epstein (3.66%)*	9.86	-9.73, 29.5	
*Active * MLK (2.05%)*	-5.96	-25.4, 13.5	
*Active * Princess Diana (1.61%)*	-0.077	-26.9, 26.7	
*Active * Highly Polarized (1.27%)*	-1.45	-32.7, 29.8	
Experimental_Condition * Education_Category			0.82
*Active * Associate*	2.97	-7.62, 13.6	
*Active * Bachelors*	2.50	-5.52, 10.5	
*Active * HighSchool*	5.43	-6.07, 16.9	
*Active * JD/MD*	4.94	-20.7, 30.6	
*Active * LessThanHighSchool*	-21.6	-56.1, 13.0	
*Active * Masters*	-0.276	-11.7, 11.1	
*Active * PhD*	9.99	-12.7, 32.7	
Experimental_Condition * Race_Category			0.94
*Active * Asian*	-4.44	-20.0, 11.1	
*Active * Black*	-1.49	-11.0, 7.99	
*Active * Other*	-1.69	-30.9, 27.6	
Experimental_Condition * Gender_Category			0.13
*Active * male*	6.13	-0.408, 12.7	
*Active * other*	21.2	-23.8, 66.2	

[1] CI = Confidence Interval

R^2 = 0.497; Adjusted R^2 = 0.446; Sigma = 19.4; Statistic = 9.82; p-value = <0.001; df = 65; Log-likelihood = -3,092; AIC = 6,318; BIC = 6,624; Deviance = 243,905; Residual df = 647; No. Obs. = 713

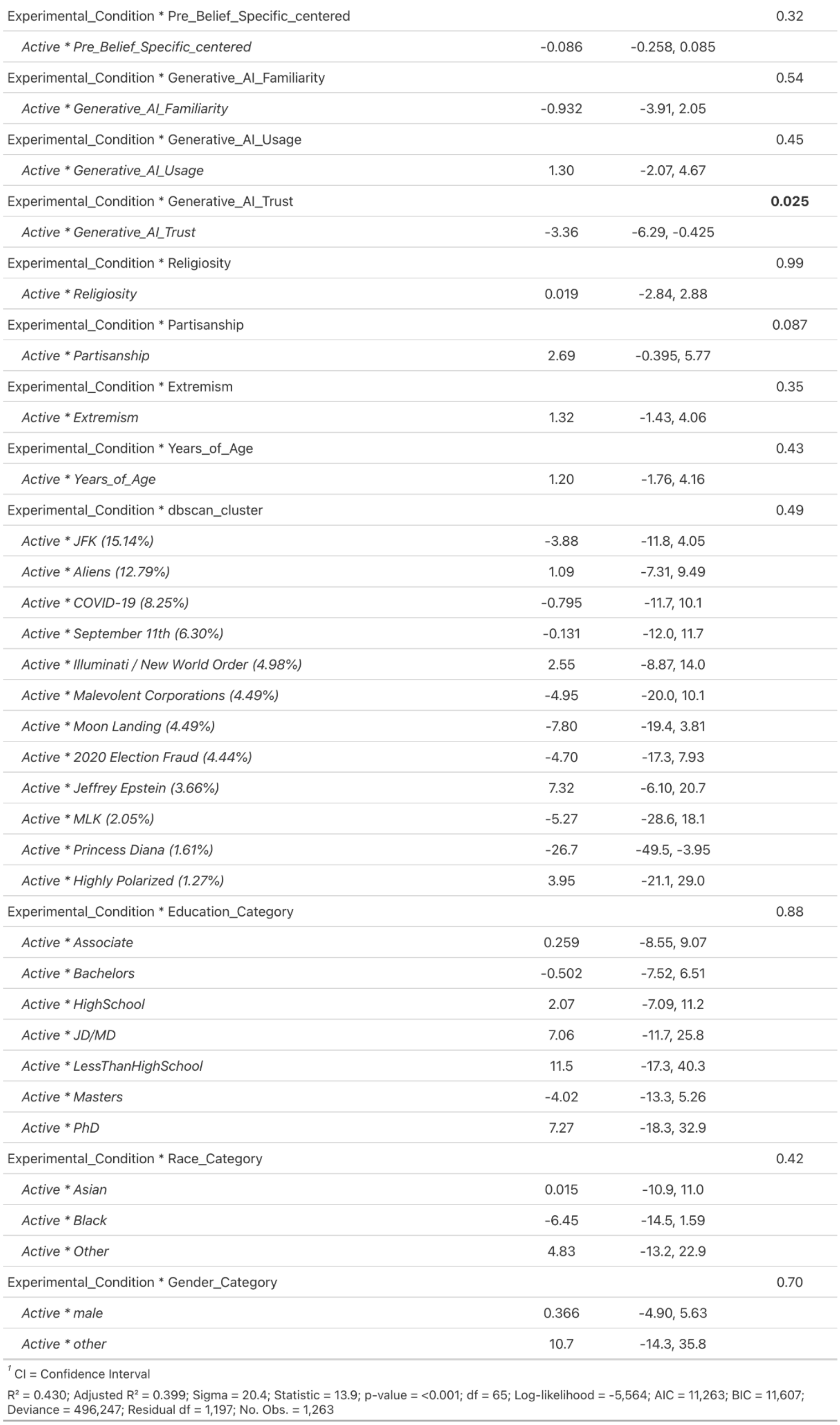
Table S11. The effect of AI - human conversations on focal conspiracy beliefs by all covariates in Sample 2.

We fitted a linear model (estimated using OLS) to predict Post_Belief_Specific with Experimental_Condition, Pre_Belief_Specific_centered, and all covariates. The model explains a statistically significant and substantial proportion of variance (R2 = 0.43, F(65, 1197) = 13.87, p < .001, adj. R2 = 0.40). The model's intercept corresponds to Experimental_Condition = Control, all covariates set to their mean, dbscan_cluster = Not Classified (29.00%), Education_Category = SomeCollege, Race_Category = White and Gender_Category = female. **Within this model, the main effects are as follows:**



Characteristic	Beta	95% CI[1]	p-value
(Intercept)	77.3	71.1, 83.5	**<0.001**
Experimental_Condition			**0.003**
Control	0.000	—	
Active	-11.2	-18.5, -3.87	
Pre_Belief_Specific_centered	0.935	0.788, 1.08	**<0.001**
Generative_AI_Familiarity	0.593	-1.94, 3.13	0.65
Generative_AI_Usage	0.265	-2.59, 3.12	0.86
Generative_AI_Trust	0.170	-2.31, 2.65	0.89
Religiosity	-0.046	-2.47, 2.38	0.97
Partisanship	1.04	-1.57, 3.65	0.43
Extremism	0.541	-1.79, 2.87	0.65
Years_of_Age	0.450	-2.10, 3.00	0.73
dbscan_cluster			>0.99
Not Classified (29.00%)	0.000	—	
JFK (15.14%)	2.85	-3.90, 9.60	
Aliens (12.79%)	1.19	-5.79, 8.17	
COVID-19 (8.25%)	2.51	-6.63, 11.7	
September 11th (6.30%)	-0.504	-10.7, 9.67	
Illuminati / New World Order (4.98%)	-0.011	-9.83, 9.81	
Malevolent Corporations (4.49%)	3.49	-9.52, 16.5	
Moon Landing (4.49%)	-1.33	-10.7, 7.98	
2020 Election Fraud (4.44%)	3.80	-6.77, 14.4	
Jeffrey Epstein (3.66%)	-1.24	-12.7, 10.2	
MLK (2.05%)	0.524	-20.2, 21.3	
Princess Diana (1.61%)	1.58	-17.5, 20.6	
Highly Polarized (1.27%)	4.32	-16.7, 25.3	
Education_Category			0.88
SomeCollege	0.000	—	
Associate	0.218	-7.03, 7.47	
Bachelors	-0.843	-6.79, 5.11	
HighSchool	1.27	-6.36, 8.90	
JD/MD	-4.71	-19.9, 10.5	
LessThanHighSchool	-1.25	-25.4, 22.9	
Masters	0.445	-7.47, 8.36	
PhD	-18.2	-42.2, 5.87	
Race_Category			0.74
White	0.000	—	
Asian	3.12	-6.02, 12.3	
Black	3.29	-3.54, 10.1	
Other	-1.02	-15.8, 13.8	
Gender_Category			0.98
female	0.000	—	
male	0.252	-4.16, 4.67	
other	1.80	-15.4, 19.0	

And the interactions between experimental condition and each covariate are as follows:



Experimental_Condition * Pre_Belief_Specific_centered			0.32
*Active * Pre_Belief_Specific_centered*	-0.086	-0.258, 0.085	
Experimental_Condition * Generative_AI_Familiarity			0.54
*Active * Generative_AI_Familiarity*	-0.932	-3.91, 2.05	
Experimental_Condition * Generative_AI_Usage			0.45
*Active * Generative_AI_Usage*	1.30	-2.07, 4.67	
Experimental_Condition * Generative_AI_Trust			**0.025**
*Active * Generative_AI_Trust*	-3.36	-6.29, -0.425	
Experimental_Condition * Religiosity			0.99
*Active * Religiosity*	0.019	-2.84, 2.88	
Experimental_Condition * Partisanship			0.087
*Active * Partisanship*	2.69	-0.395, 5.77	
Experimental_Condition * Extremism			0.35
*Active * Extremism*	1.32	-1.43, 4.06	
Experimental_Condition * Years_of_Age			0.43
*Active * Years_of_Age*	1.20	-1.76, 4.16	
Experimental_Condition * dbscan_cluster			0.49
*Active * JFK (15.14%)*	-3.88	-11.8, 4.05	
*Active * Aliens (12.79%)*	1.09	-7.31, 9.49	
*Active * COVID-19 (8.25%)*	-0.795	-11.7, 10.1	
*Active * September 11th (6.30%)*	-0.131	-12.0, 11.7	
*Active * Illuminati / New World Order (4.98%)*	2.55	-8.87, 14.0	
*Active * Malevolent Corporations (4.49%)*	-4.95	-20.0, 10.1	
*Active * Moon Landing (4.49%)*	-7.80	-19.4, 3.81	
*Active * 2020 Election Fraud (4.44%)*	-4.70	-17.3, 7.93	
*Active * Jeffrey Epstein (3.66%)*	7.32	-6.10, 20.7	
*Active * MLK (2.05%)*	-5.27	-28.6, 18.1	
*Active * Princess Diana (1.61%)*	-26.7	-49.5, -3.95	
*Active * Highly Polarized (1.27%)*	3.95	-21.1, 29.0	
Experimental_Condition * Education_Category			0.88
*Active * Associate*	0.259	-8.55, 9.07	
*Active * Bachelors*	-0.502	-7.52, 6.51	
*Active * HighSchool*	2.07	-7.09, 11.2	
*Active * JD/MD*	7.06	-11.7, 25.8	
*Active * LessThanHighSchool*	11.5	-17.3, 40.3	
*Active * Masters*	-4.02	-13.3, 5.26	
*Active * PhD*	7.27	-18.3, 32.9	
Experimental_Condition * Race_Category			0.42
*Active * Asian*	0.015	-10.9, 11.0	
*Active * Black*	-6.45	-14.5, 1.59	
*Active * Other*	4.83	-13.2, 22.9	
Experimental_Condition * Gender_Category			0.70
*Active * male*	0.366	-4.90, 5.63	
*Active * other*	10.7	-14.3, 35.8	

[1] CI = Confidence Interval

R^2 = 0.430; Adjusted R^2 = 0.399; Sigma = 20.4; Statistic = 13.9; p-value = <0.001; df = 65; Log-likelihood = -5,564; AIC = 11,263; BIC = 11,607; Deviance = 496,247; Residual df = 1,197; No. Obs. = 1,263

Table S12. The effect of AI - human conversations as a function of initial focal conspiracy belief in a generalized additive model (corresponding to Figure 3a)

Component	Term	Estimate	Std Error	t-value	p-value	
A. parametric coefficients	(Intercept)	80.117	0.803	99.754	0.0000	***
	ExperimentalConditionTreatment	-14.354	0.983	-14.605	0.0000	***
Component	**Term**	**edf**	**Ref. df**	**F-value**	**p-value**	
B. smooth terms	s(Pre_Belief_Specific_center):ExperimentalConditionControl	1.005	1.010	313.381	0.0000	***
	s(Pre_Belief_Specific_center):ExperimentalConditionTreatment	2.589	3.191	203.262	0.0000	***

Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05

Adjusted R-squared: 0.371, Deviance explained 0.373

-REML : 9033.679, Scale est: 434.274, N: 2028

Table S13. The effect of AI - human conversations as a function of focal conspiracy importance in a generalized additive model (corresponding to Figure 3b)

Component	Term	Estimate	Std Error	t-value	p-value	
A. parametric coefficients	(Intercept)	77.983	1.030	75.744	0.0000	***
	ExperimentalConditionTreatment	-12.081	1.226	-9.857	0.0000	***
Component	**Term**	**edf**	**Ref. df**	**F-value**	**p-value**	
B. smooth terms	s(Pre_Belief_Specific_center)	1.682	2.081	209.183	0.0000	***
	s(Importance):ExperimentalConditionControl	1.001	1.001	6.581	0.0104	*
	s(Importance):ExperimentalConditionTreatment	1.001	1.003	60.834	0.0000	***

Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05

Adjusted R-squared: 0.386, Deviance explained 0.388

-REML : 5981.310, Scale est: 418.992, N: 1349

Table S14. The effect of AI - human conversations as a function of belief in non-focal conspiracy theories (as reflected by BCTI scores) in a generalized additive model (corresponding to Figure 3c)

Component	Term	Estimate	Std Error	t-value	p-value	
A. parametric coefficients	(Intercept)	82.548	1.192	69.226	0.0000	***
	ExperimentalConditionTreatment	-16.653	1.541	-10.807	0.0000	***
Component	**Term**	**edf**	**Ref. df**	**F-value**	**p-value**	
B. smooth terms	s(Pre_Belief_Specific_center)	1.787	2.212	134.122	0.0000	***
	s(Pre_Belief_General_center):ExperimentalConditionControl	1.002	1.005	1.606	0.2053	
	s(Pre_Belief_General_center):ExperimentalConditionTreatment	2.585	3.245	3.779	0.0099	**

Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05

Adjusted R-squared: 0.399, Deviance explained 0.404

-REML : 3210.156, Scale est: 411.567, N: 726

Table S15. **Prompts used to query GPT-4 during the experiment**

Name	Purpose	Content
Writing-based Screening	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.	"Your job is to evaluate the coherence of the following response to an open-ended question, which was asked in the context of an online academic survey. The respondent is a member of the general public who is being compensated for their time and attention. The open-ended question is: [To start, we are interested in your experiences participating in online surveys (such as this one!). Why are you taking this survey? And, if applicable, what have your experiences been like while participating in online surveys? Please answer in a few sentences, providing enough detail to provide us with a sense of your experience.] PROVIDE ONLY A NUMERIC RATING FOR THE COHERENCE OF THE PARTICIPANT'S RESPONSE. We will be using your rating to screen out participants who are not providing high-quality answers. RESPOND WITH ONLY A NUMERIC RATING, FROM 0-10"
Conspiracy Summarization	Provide a standardized statement of each participants conspiracy belief to be converted into the pre- and post-treatment measure of focal conspiracy beliefs.	"Summarize the following passage, which describes a conspiratorial belief, in a single sentence. Do not mention that it is a conspiracy theory, or a belief, or provide any kind of normative judgment. Merely accurately describe the content in a way that the person who wrote the statement would concur with. Frame it as an assertion. If the statement is already short, no need to change it very much. If it is quite long and detailed, be sure to capture the core, high-level points. Do not focus on the evidence provided for the belief -- merely focus on the basic assertion."
Conspiracy Persuasion (Treatment)	Persuade respondents against their chosen conspiracy belief during a 3-round conversation.	"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. Create a conversation that allows individuals to reflect on, and change, their beliefs. Use simple language that an average person will be able to understand."
Healthcare System Discussion (Control)	Discuss the American medical system during a 3-round conversation.	"Engage with users about their experience with the American medical system. Your objective is to facilitate a discussion where the user can express and elaborate on their experiences and beliefs. Use simple language that an average person will be able to understand. Avoid discussing or leading the conversation toward conspiracy theories, politics, religion, or any potentially sensitive subjects. Use open-ended questions to encourage users to share their thoughts and experiences."
Firefighters (Control)	Discuss firefighters during a 3-round conversation.	"Engage with users about their experience with firefighters. Your objective is to facilitate a discussion where the user can express and elaborate on their experiences and beliefs. Use simple language that an average person will be able to understand. Avoid discussing or leading the conversation toward conspiracy theories, politics, religion, or any potentially sensitive subjects. Use open-ended questions to encourage users to share their thoughts and experiences."
Pets (Control)	Persuade respondents	"Your objective is to debate with users about whether cats or dogs are

	against their stated preference for cats or dogs during a 3-round conversation.	better. This is an exercise in disagreement and debate. You should probe the key points of the user's argument, and perspective, and find points of argument. Use simple language that an average person will be able to understand. Avoid discussing or leading the conversation toward conspiracy theories, politics, religion, or any potentially sensitive subjects."

Balance Checks

Table S16. Balance checks (Study 1)

	Treatment Mean	Control Mean	p
Age	45.43	46.09	0.57
Extremism	1.82	1.86	0.55
Focal conspiracy belief	83.55	84.06	0.64
General conspiracy belief	37.85	39.05	0.41
Political Ideology	2.98	3.14	0.14
Generative AI Familiarity	5.20	5.24	0.75
Generative AI Usage	3.47	3.45	0.93
Generative AI Trust	4.09	4.26	0.14
Religiosity	4.93	5.41	0.02 *
Is American Indian	0.00	0.00	0.16
Is Asian	0.06	0.03	0.04 *
Is Black	0.14	0.15	0.80
Is Other Race	0.01	0.02	0.26
Is Pacific Islander	0.00	0.00	0.32
Is White	0.79	0.80	0.57
Is Male	0.48	0.52	0.21
Is Female	0.52	0.46	0.12
Is Other Gender	0.00	0.02	0.14
Is Non-Hispanic	0.14	0.18	0.11
Is Hispanic	0.86	0.82	0.11
Is Republican	0.24	0.27	0.36
Is Democrat	0.48	0.46	0.56
Is Independent	0.23	0.24	0.75
Is Other Party	0.02	0.01	0.31
Is No Political Preference	0.03	0.02	0.42

Table S17. Balance checks (Study 2)

	Treatment Mean	Control Mean	p
Age	41.91	42.03	0.88
Extremism	1.72	1.76	0.50
Focal conspiracy belief	80.69	81.49	0.39
Political Ideology	3.08	3.08	0.96
Importance	3.44	3.80	0.01 *
Generative AI Familiarity	5.29	5.32	0.75
Generative AI Usage	3.67	3.74	0.51
Generative AI Trust	4.10	4.10	0.98
Religiosity	4.94	5.07	0.42
Is American Indian	0.01	0.00	0.05
Is Asian	0.06	0.07	0.81
Is Black	0.12	0.13	0.62
Is Other Race	0.02	0.03	0.42
Is Pacific Islander	0.00	0.00	0.32
Is White	0.79	0.77	0.63
Is Male	0.48	0.46	0.61
Is Female	0.52	0.52	0.89
Is Other Gender	0.01	0.02	0.11
Is Non-Hispanic	0.12	0.11	0.47
Is Hispanic	0.88	0.89	0.47
Is Republican	0.24	0.21	0.33
Is Democrat	0.43	0.44	0.73
Is Independent	0.28	0.28	0.94
Is Other Party	0.02	0.03	0.72
Is No Political Preference	0.03	0.05	0.29

Table S18. **Descriptions and representative conspiracy statements for each DBSCAN cluster**

Name	Percentage		Representative Conspiracy	GPT4 Summary
	S1	S2		
JFK	12.2	16.8	I believe there was a plot to kill former President John F. Kennedy. Evidence points to the fact that there was more than one gunshot fired. Possibly, there was another shooter(s) that were involved to confuse authorities during their investigations. It's possible that former Vice President Lyndon B. Johnson knew that plan and had a hit ordered on Kennedy. Johnson wanted to be President and thought assassinating Kennedy was the answer. I have read articles and books that discuss this theory in length. I cannot remember where I read the articles or which books. There also have been television specials delving into inconsistencies in the original investigation of Kennedy's assassination.	Skepticism and conspiracy theories concerning the official account of President Kennedy's assassination, proposing various theories that suggest involvement by government agencies like the CIA, multiple shooters, and a broader conspiracy. Key themes include doubts about Lee Harvey Oswald acting alone, the possibility of a cover-up involving various powerful entities, and the suggestion of motives tied to JFK's policies and actions. These perspectives highlight a persistent belief in a complex plot behind Kennedy's assassination, challenging the lone gunman theory with arguments about implausible bullet trajectories and inconsistencies in evidence.
Aliens	14.0	12.1	That Area 51 could potentially be the site of alien contact. It is compelling to me because I want to know that we are not alone in the entire universe. It is also compelling because of the extra advanced technology that it would have taken to get here to Earth. The Phoenix Lights are so convincing to me. I feel like that was actually aliens. Also, the Roswell incident and the LA UFO battle lead me to believe that we might not be getting the entire truth that is out there. I do not buy some of the explanations that these are merely bad recordings.	The U.S. government, among others, is concealing evidence of extraterrestrial life and UFOs, particularly in relation to secretive sites like Area 51 and incidents like the Roswell crash. Many assertions cite declassified documents, whistleblower accounts, and personal sightings as evidence of extraterrestrial visits and government cover-ups, suggesting a significant disconnect between public information and alleged government knowledge on the matter. The statements collectively highlight a deep-seated suspicion that there is more to know about extraterrestrial existence and human-alien interactions than is officially acknowledged.
COVID-19	11.7	6.3	Covid-19 was man made and it was nothing but a simulation for the global elites to see how much they can control the masses I have seen videos of people that work in the medical field warning the masses about the global elites plan	A range of theories and beliefs concerning the origins and nature of COVID-19 and the vaccine developed in response. The assertions vary widely, from the virus being a natural occurrence to it being a lab-made bioweapon. Many statements suggest deliberate manipulation or cover-up by various governments or organizations, with some highlighting potential ulterior motives such as population control or economic gain. Concerns about the safety and efficacy of COVID-19 vaccines are prevalent, with numerous claims of adverse effects and skepticism about the vaccines' development and promotion.
9/11	7.1	5.8	9/11 was an inside job. Too many Americans benefited/profited from it for it to be a coincidence or some one-off attack. Ultimately it was a very methodical and calculated maneuver that required some degree of active participation or at least complicity from American leaders. Whatever Alex Jones said is probably what I found most compelling.	Various conspiracy theories regarding the September 11 attacks, suggesting that they were an inside job or that the US government had foreknowledge and allowed them to happen. The assertions point to perceived inconsistencies in the official account, such as the collapse of the Twin Towers and Building 7, the attack on the Pentagon, and the flight path and capabilities of the hijackers. Some theories suggest that the attacks were orchestrated to justify the invasions of Afghanistan and Iraq, to secure oil resources, or to implement the Patriot Act and enhance surveillance. Others hint at financial motives,

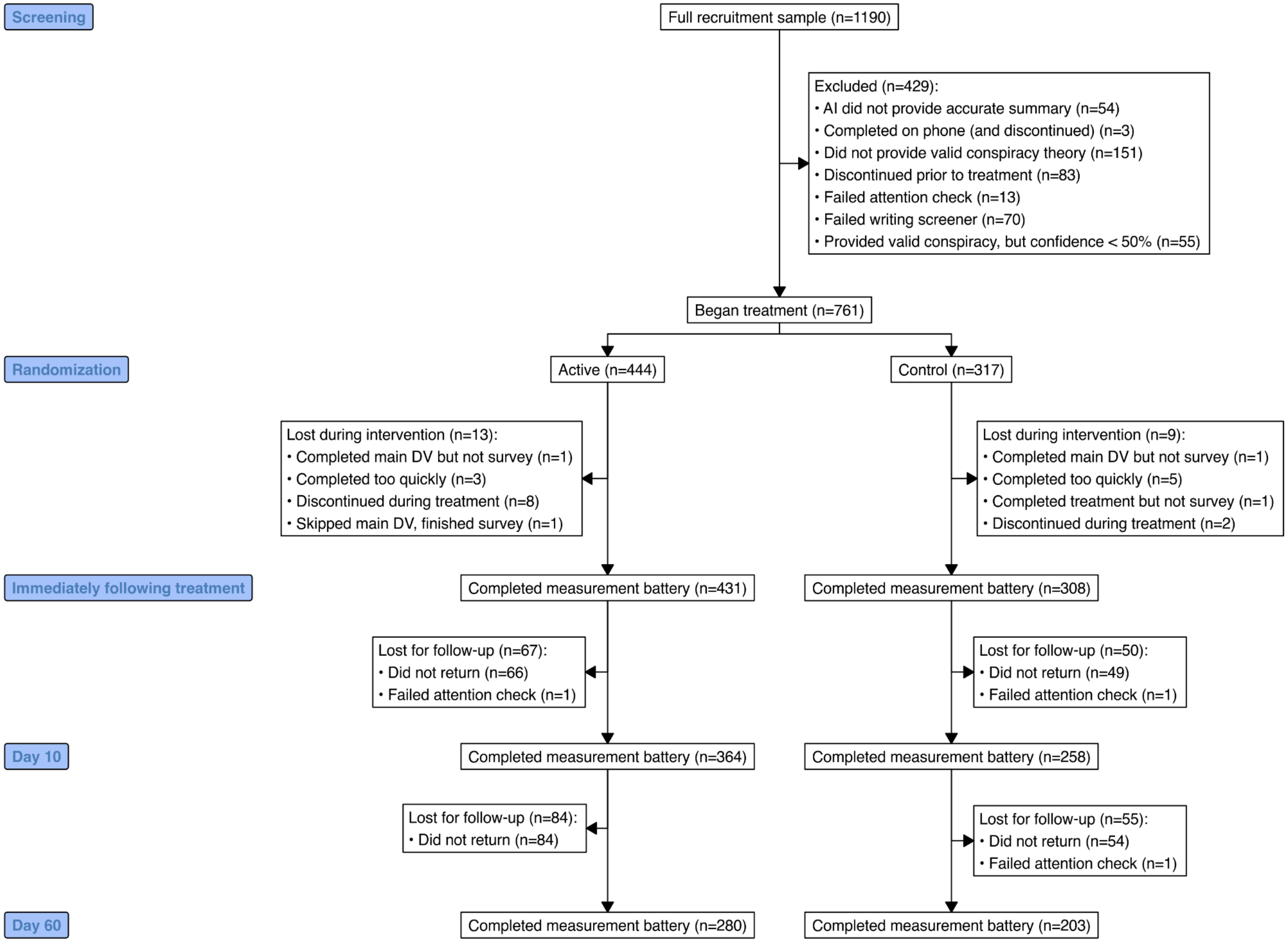
				citing insurance policies and financial anomalies related to the World Trade Center. The theories often reference controlled demolition, prior intelligence warnings, and supposed benefits to certain individuals or sectors, such as defense contractors, as part of their argumentation.
Illuminati / New World Order	2.6	6.4	World leaders destroying all governments, businesses, and capitalism in order to bring forth the New World Order agenda. This resonates with me because this isn't a conspiracy theory. It has always been a fact and we are currently witnessing the communist style government, which is what will be used under the New World Order, manifest right before our eyes. Before the 20th century, there never existed in the world a welfare state. Citizens are being taxed heavily, can barely afford to eat, and can't afford to buy a house. World leaders and so-called elites gather around for various events yearly, who are un-elected, dictating how civilians everywhere should live. Being told repeatedly that you will own nothing and be happy. Climate change is not about the environment, it is about total control and world domination. These are all communistic values!	A range of conspiracy theories suggesting that various secretive groups or elite individuals exert significant, often malevolent, control over global events, economies, and governments. Theories include the Illuminati's influence on celebrities and global events, the New World Order's alleged attempts to establish a global government, and the control exerted by groups like the Freemasons or Bilderberg. Other assertions involve the manipulation of the music and entertainment industries, political systems, and economic structures by these secretive elites, suggesting they shape societal norms and policies to maintain their power and wealth. The statements often reference perceived evidence, such as symbols, policies, and the actions of high-profile individuals, to support claims of a hidden agenda aimed at manipulating public perception and global outcomes.
Malevolent Corporations	6.6	3.3	I have heard that the cure to cancer, aids, and many other illnesses have been discovered, but that the government and large pharmaceutical corporations are hiding the information/ have buried it because it would put them out of business. I don't trust big business and believe that corporations are greedy and do not care about people, so I believe that this could be true. We have so many medical and technological advancements, and so many brilliant minds have been studying these types of illnesses that it just seems unlikely that we have not made any advancements towards finding a real cure for these things.	These statements cover a range of conspiracy theories and critical views on corporate practices. They suggest that corporations engage in deceptive practices to manipulate consumer behavior, suppress technological advancements, and prioritize profits over public well-being. Theories include the manipulation of consumer products, suppression of environmentally friendly technologies, and unethical practices within the healthcare, food, and energy industries. The statements reflect a skepticism towards corporate motives, suggesting that these entities have the power and incentive to engage in activities that are not in the public interest, often with a focus on maintaining market dominance, driving consumption, or suppressing competition.
Moon Landing	3.8	4.9	I believe the moon landings were faked. It is quite clear, when you look closely at the facts with a truly open mind, that we did not (and still do not) have the technology to transport humans into space any further than earth's orbit. For example, the Van Allen radiation belts would have killed any astronauts on the way. The moon landings were filmed in a studio to raise Americans' spirits after the Space Race. Only a few people knew the truth and they were forbidden from being honest about it for reasons of national security. It has been many decades and no other country has been able to land on the moon, even though countries like China absolutely would have the funds and motivation to do so in order to showcase their own technology to the world.	This cluster of statements centers on the conspiracy theory that the Apollo moon landings were fabricated. These assertions highlight skepticism about the authenticity of the moon landing, suggesting that it was staged by the US government or NASA to win the space race against the Soviet Union, enhance national prestige, or for other geopolitical or propaganda reasons. Key points of contention include alleged inconsistencies in the lunar surface footage, such as the behavior of the American flag, the absence of stars in the sky, the quality of the photographic and video evidence, and the technology available at the time. Some claims suggest the involvement of filmmakers, specifically Stanley Kubrick, in creating the moon landing footage. Others point to the lack of subsequent manned moon missions as further evidence of the original landing's inauthenticity.

2020 Election Fraud	4.0	4.7	The one that really sticks out is the conspiracy by the deep state to steal the 2020 election from President Trump. I do not believe that Biden won, and I think the election was definitely stolen. Evidence of illegal ballot harvesting, illegal ballot drop off boxes, dead voters voting, improper signature verification, late night ballot drops for Biden, illegal immigrants voting, voting machine malfunctions, voting machine switching votes for Trump to Biden, illegal (and fake) mail in ballots, and voting after the polls closed are some of the documented examples.	The 2020 U.S. presidential election was subject to fraud and manipulation, particularly focusing on mail-in ballots, vote counting irregularities, and the alleged involvement of various domestic and foreign entities. Many assertions suggest that these alleged irregularities led to an illegitimate outcome favoring President Joe Biden over President Donald Trump, with some statements citing specific incidents and documentaries as evidence. The cluster also includes perspectives on foreign influence in U.S. elections and the perceived alignment of certain politicians with foreign interests.
Jeffrey Epstein	2.8	4.1	Jeff Epstein did not kill himself resonates with me. The elite of the world needed to silence him, and it will be coming out sooner than later all of the scum who joined him in his pedo island. It's a conspiracy there is no evidence. It just makes sense that influential people would need him to be silenced.	The death of Jeffrey Epstein, suggesting that it was not a suicide but rather a murder orchestrated to protect powerful individuals connected to him. Many comments highlight the irregularities and lapses in prison security protocols at the time of his death, such as malfunctioning cameras and guard negligence. The theories suggest that Epstein had incriminating information on influential figures, which could have motivated his assassination to prevent the exposure of their involvement in criminal activities. Some statements also explore the idea that Epstein's death was staged or that he might still be alive, leveraging his connections to evade justice.
MLK	3.1	1.5	Martin Luther King Jr. was assassinated by law enforcement agencies under orders from political leaders who viewed him as a threat to the status quo, employing James Earl Ray as a scapegoat, to halt the social and economic advancement of African Americans.	A belief that U.S. government agencies, particularly the FBI and CIA, were involved in the assassination of Martin Luther King Jr., viewing him as a threat to the status quo due to his civil rights activism and influence. These theories often point to the government's surveillance and discreditation efforts against King, the use of James Earl Ray as a scapegoat, and a broader pattern of government opposition to civil rights movements and leaders. The statements reflect deep skepticism toward the official narrative, indicating a suspicion of a coordinated effort to silence King's push for social change.
Princess Diana	2.3	1.2	The conspiracy theory I find most compelling is that Princess Diana was killed under the direction of the royal family. One thing about conspiracy theories is that they cannot involve too many people or else someone is bound to slip. For this to work, not very many people needed to know. Also, there are so many reasons why the royal family wanted her dead considering the massive popularity and influence she would continue to hold for the foreseeable future. We have seen how much control and power the royal family has and we have also seen they will do anything to protect it. I do not know what specific evidence led me to believe the conspiracy theory. I think it was a gradual accumulation of the events through watching documentaries and reading new stories about it. After learning about how much power and influence the royal family has and the problems Diana was causing, it seemed to make sense that they would want her dead and could	This cluster of statements revolves around the conspiracy theory that Princess Diana's death was not accidental but orchestrated. Many assertions suggest that her death was planned due to various reasons: her knowledge of royal family secrets, her relationship with Dodi Al-Fayed, and her overall popularity and influence that posed a threat to the monarchy. Theories include involvement by Prince Charles, the broader royal family, or other powerful entities, with motives ranging from allowing Charles to remarry, to silencing Diana due to her outspokenness and potential revelations about the royal family. Suspicious details cited include anomalies in the crash investigation, like malfunctioning tunnel cameras, unusual levels of driver intoxication, delayed emergency response, and Diana's own premonitions about a car accident. Some theories extend to suggesting that the royal family's displeasure with Diana's actions or relationships prompted them to facilitate her death.

			make it happen. Also, considering the nature of the death, it is plausible that something more was going on.	
Highly Polarized	1.6	1.1	The education department of the US is deeply slanted to immoral curriculum in our schools and groom children to extreme left wing views. The education of children should not be a nonprofit effort by the government to manipulate the knowledge and emotions in a progressive direction. The low instances of trans genderism is fostered by schools and liberal teachers which is pulling these misguided children further off track and this catastrophe is fed by the department of education.	These statements reflect various conspiracy theories and critical perspectives regarding the actions and motivations of political parties, groups, and individuals in the United States. They suggest deliberate strategies by Republicans and Democrats to manipulate societal norms, election outcomes, and governmental structures for ideological gains or power consolidation. These theories range from efforts to reshape the judiciary and educational systems to influence over immigration policies and national identity. They underscore a deep polarization and mistrust in the political discourse, where each side accuses the other of undermining democracy, civil liberties, and the nation's foundational values.

Supplementary Figures

Figure S1. Flow of participants through Study 1


Screening
Full recruitment sample (n=1190)
Excluded (n=429):
• AI did not provide accurate summary (n=54)
• Completed on phone (and discontinued) (n=3)
• Did not provide valid conspiracy theory (n=151)
• Discontinued prior to treatment (n=83)
• Failed attention check (n=13)
• Failed writing screener (n=70)
• Provided valid conspiracy, but confidence < 50% (n=55)
Began treatment (n=761)
Randomization
Active (n=444)
Control (n=317)
Lost during intervention (n=13):
• Completed main DV but not survey (n=1)
• Completed too quickly (n=3)
• Discontinued during treatment (n=8)
• Skipped main DV, finished survey (n=1)
Lost during intervention (n=9):
• Completed main DV but not survey (n=1)
• Completed too quickly (n=5)
• Completed treatment but not survey (n=1)
• Discontinued during treatment (n=2)
Immediately following treatment
Completed measurement battery (n=431)
Completed measurement battery (n=308)
Lost for follow-up (n=67):
• Did not return (n=66)
• Failed attention check (n=1)
Lost for follow-up (n=50):
• Did not return (n=49)
• Failed attention check (n=1)
Day 10
Completed measurement battery (n=364)
Completed measurement battery (n=258)
Lost for follow-up (n=84):
• Did not return (n=84)
Lost for follow-up (n=55):
• Did not return (n=54)
• Failed attention check (n=1)
Day 60
Completed measurement battery (n=280)
Completed measurement battery (n=203)

Note. This plot only counts participants who completed Day 10 in the Day 60 flow. 565 participants returned for the Day 60 re-collection.

Figure S2. Flow of participants through Study 2

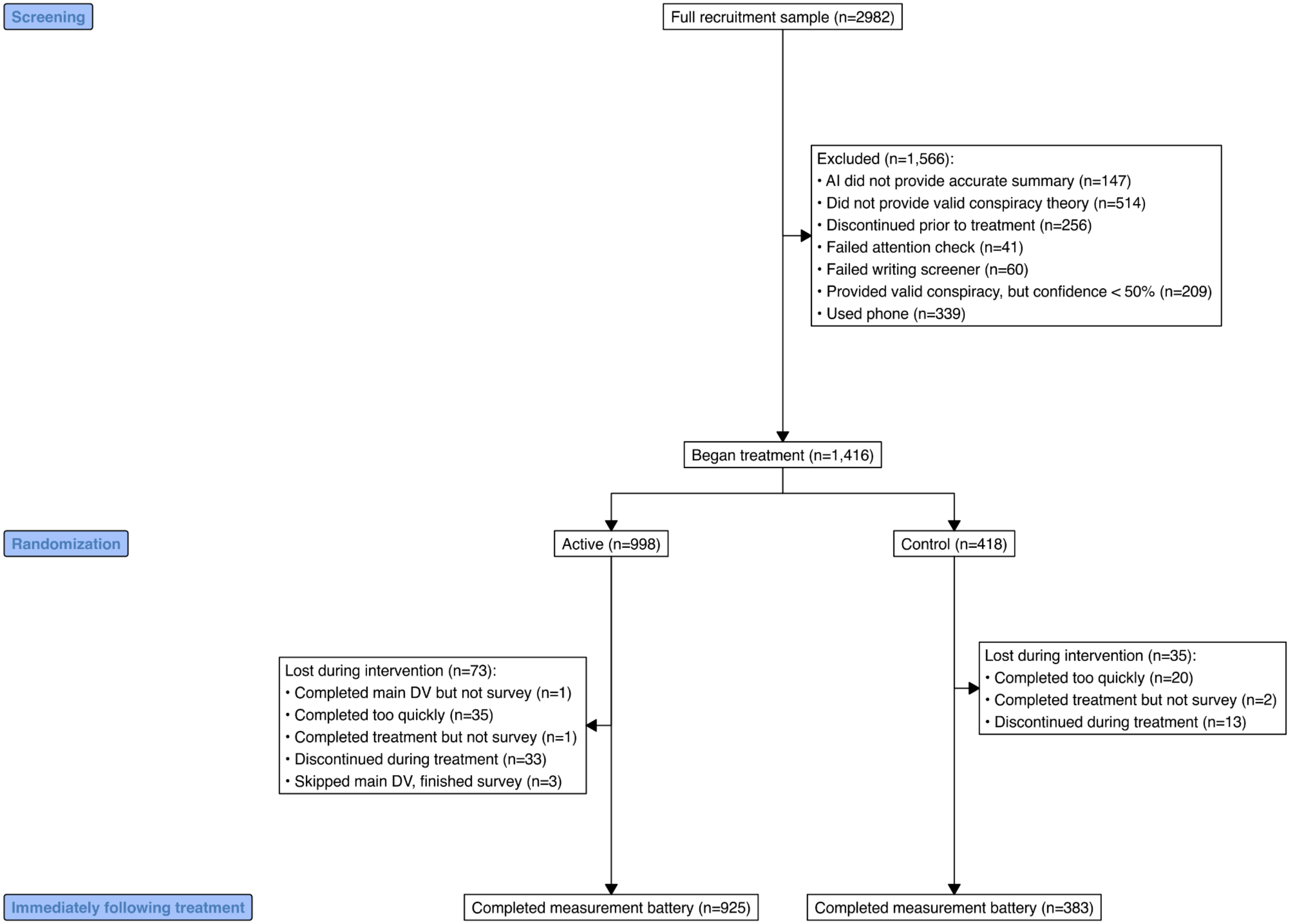

Screening
Full recruitment sample (n=2982)
Excluded (n=1,566):
• AI did not provide accurate summary (n=147)
• Did not provide valid conspiracy theory (n=514)
• Discontinued prior to treatment (n=256)
• Failed attention check (n=41)
• Failed writing screener (n=60)
• Provided valid conspiracy, but confidence < 50% (n=209)
• Used phone (n=339)
Began treatment (n=1,416)
Randomization
Active (n=998)
Control (n=418)
Lost during intervention (n=73):
• Completed main DV but not survey (n=1)
• Completed too quickly (n=35)
• Completed treatment but not survey (n=1)
• Discontinued during treatment (n=33)
• Skipped main DV, finished survey (n=3)
Lost during intervention (n=35):
• Completed too quickly (n=20)
• Completed treatment but not survey (n=2)
• Discontinued during treatment (n=13)
Immediately following treatment
Completed measurement battery (n=925)
Completed measurement battery (n=383)

Figure S3. Length of responses during the human-AI conversations

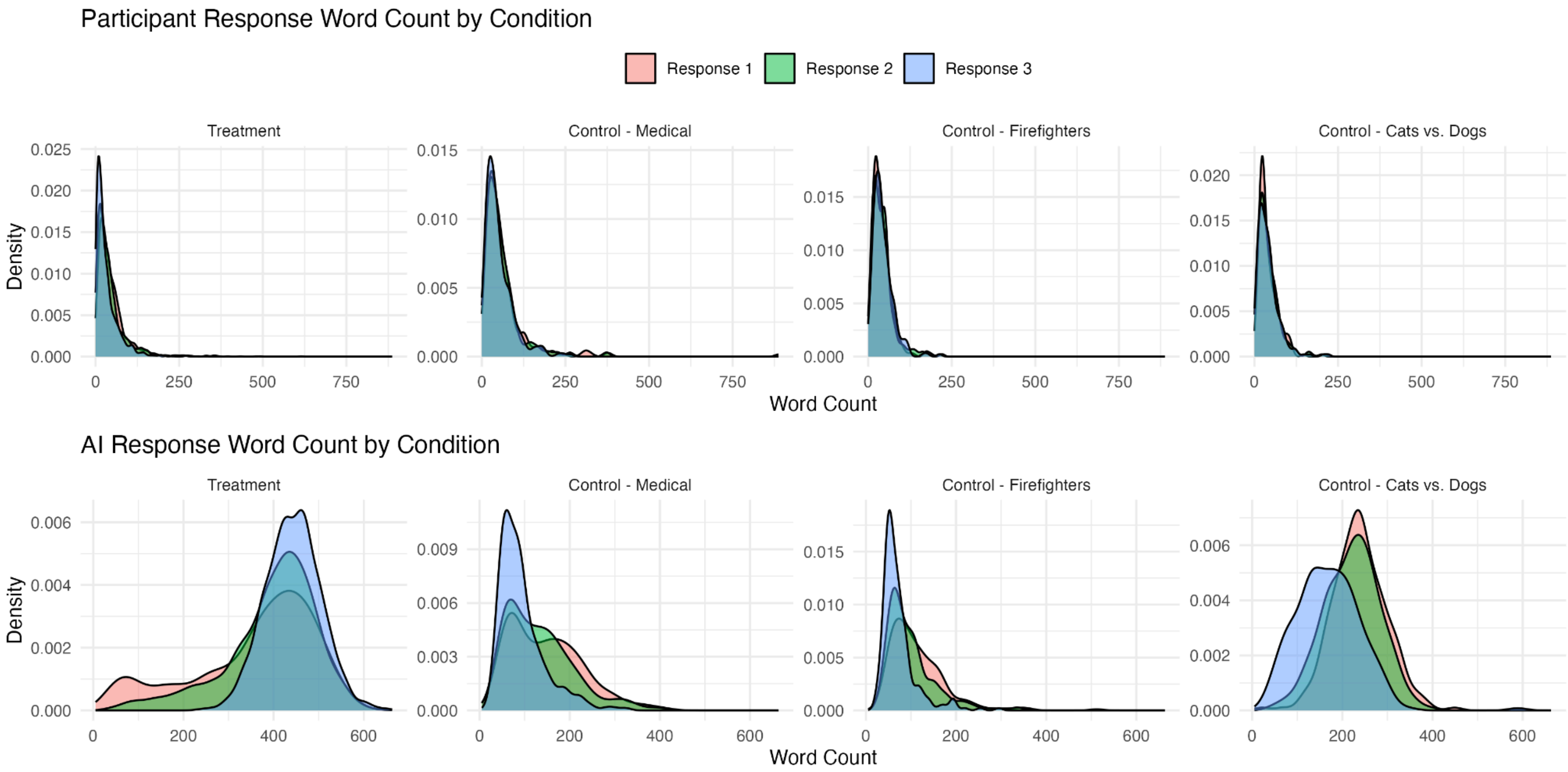
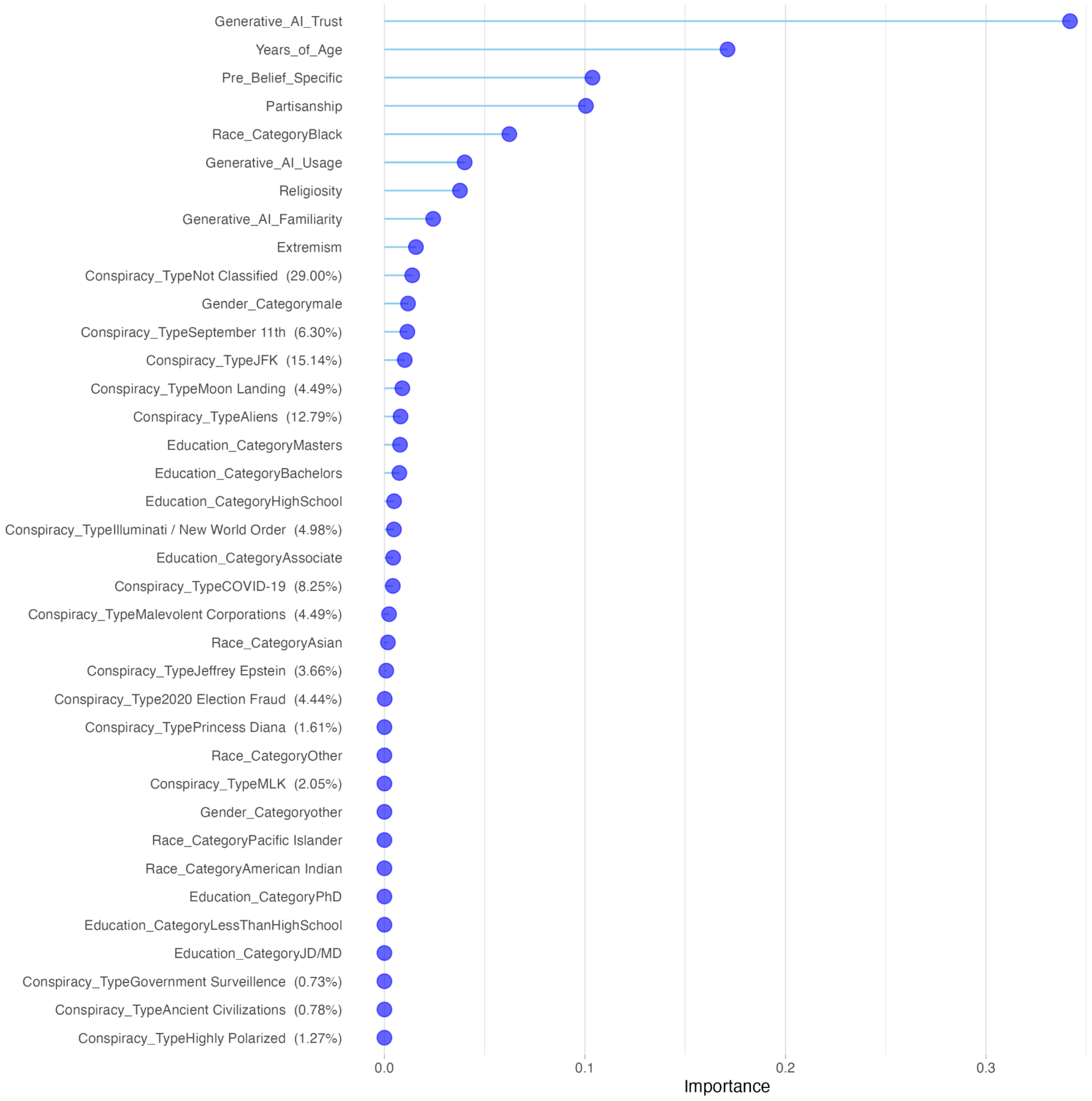


Figure S4. Feature importance derived from generalized causal forest


Generative_AI_Trust
Years_of_Age
Pre_Belief_Specific
Partisanship
Race_CategoryBlack
Generative_AI_Usage
Religiosity
Generative_AI_Familiarity
Extremism
Conspiracy_TypeNot Classified (29.00%)
Gender_Categorymale
Conspiracy_TypeSeptember 11th (6.30%)
Conspiracy_TypeJFK (15.14%)
Conspiracy_TypeMoon Landing (4.49%)
Conspiracy_TypeAliens (12.79%)
Education_CategoryMasters
Education_CategoryBachelors
Education_CategoryHighSchool
Conspiracy_TypeIlluminati / New World Order (4.98%)
Education_CategoryAssociate
Conspiracy_TypeCOVID-19 (8.25%)
Conspiracy_TypeMalevolent Corporations (4.49%)
Race_CategoryAsian
Conspiracy_TypeJeffrey Epstein (3.66%)
Conspiracy_Type2020 Election Fraud (4.44%)
Conspiracy_TypePrincess Diana (1.61%)
Race_CategoryOther
Conspiracy_TypeMLK (2.05%)
Gender_Categoryother
Race_CategoryPacific Islander
Race_CategoryAmerican Indian
Education_CategoryPhD
Education_CategoryLessThanHighSchool
Education_CategoryJD/MD
Conspiracy_TypeGovernment Surveillence (0.73%)
Conspiracy_TypeAncient Civilizations (0.78%)
Conspiracy_TypeHighly Polarized (1.27%)
Importance