



*Research Article*

---

## Using an AI-powered “street epistemologist” chatbot and reflection tasks to diminish conspiracy theory beliefs

*Social scientists, journalists, and policymakers are increasingly interested in methods to mitigate or reverse the public’s beliefs in conspiracy theories, particularly those associated with negative social consequences, including violence. We contribute to this field of research using an artificial intelligence (AI) intervention that prompts individuals to reflect on the uncertainties in their conspiracy theory beliefs. Conspiracy theory believers who interacted with our “street epistemologist” chatbot subsequently showed weaker conviction in their conspiracy theory beliefs; this was also the case for subjects who were asked to reflect on their beliefs without conversing with an AI chatbot. We found that encouraging believers to reflect on their uncertainties can weaken beliefs and that AI-powered interventions can help reduce epistemically unwarranted beliefs for some believers.*

Authors: Marco Meyer (1), Adam Enders (2), Casey Klofstad (3), Justin Stoler (4), Joseph Uscinski (3)

Affiliations: (1) Department of Philosophy, University of Hamburg, Germany, (2) Department of Political Science, University of Louisville, USA, (3) Department of Political Science, University of Miami, USA, (4) Department of Geography and Sustainable Development, University of Miami, USA

How to cite: Meyer, M., Enders, A., Klofstad, C., Stoler, J., & Uscinski, J. (2024). Using an AI-powered “street epistemologist” chatbot and reflection tasks to diminish conspiracy theory beliefs. *Harvard Kennedy School (HKS) Misinformation Review*, 5(6).

Received: June 24<sup>th</sup>, 2024. Accepted: September 26<sup>th</sup>, 2024. Published: December 12<sup>th</sup>, 2024.

### Research questions

- Can prompting individuals to reflect on the uncertainties underlying their conspiracy theory beliefs reduce the strength of such beliefs?
- Can artificial intelligence be used to facilitate the type of reflection that mitigates conspiracy theory beliefs?
- What roles do predispositions and concerns for belief accuracy play in fostering or stymying the efficacy of interventions designed to reduce the strength of conspiracy theory beliefs?
- What can researchers do to bolster the efficacy of interventions designed to mitigate conspiracy theory beliefs?

---

<sup>1</sup> A publication of the Shorenstein Center on Media, Politics and Public Policy at Harvard University, John F. Kennedy School of Government.

## Essay summary

- Prompting individuals to consider the reasons why they believe in conspiracy theories and the reservations they may have about such beliefs tends to reduce conviction.
- An AI-powered chatbot instructed to behave like a “street epistemologist,” by probing individuals about their justifications for their conspiracy theory beliefs, can also reduce the strength of conspiracy theory beliefs.
- Individuals who tend to either 1) report that the accuracy of a specific belief is important to them or 2) exhibit a greater predisposition towards viewing events and circumstances as the product of conspiracies are less likely to reduce the strength of the specific conspiracy theory belief after being prompted to reflect on that belief, with or without the chatbot intervention.
- While it is encouraging that, on average, people soften their conspiracy theory beliefs upon reflection, the believers posing the greatest potential social challenges may be immune to such interventions.

## Implications

Conspiracy theory research has increased dramatically over the last 15 years, particularly in response to the Trump presidency and COVID-19 pandemic (Hornsey et al., 2023). Much research examines the causes (Douglas et al., 2019) and consequences (Jolley et al., 2020) of conspiracy theory beliefs. The concern over conspiracy theories is largely due to their association with nonnormative behaviors including crime, vaccine refusal, and political violence (Jolley et al., 2022). Journalists, too, have expressed concerns over how beliefs in conspiracy theories may promote dangerous individual behaviors (Collins, 2020) and influence public policy (Moine, 2024; Zadrozny, 2024).

Unsurprisingly then, a growing body of research has focused on preventing or weakening beliefs in conspiracy theories or in other more general forms of epistemically suspect information (e.g., misinformation) (Banas & Miller, 2013; Bode & Vraga, 2018; Bonetto et al., 2018; Compton et al., 2021; Islam et al., 2021; Jolley & Douglas, 2017; Nyhan et al., 2013; Smith et al., 2023). Within this work, researchers have attempted to prevent epistemically suspect beliefs by exposing individuals to small bits of information (sometimes referred to as “inoculations” in the literature) intended to help those individuals spot suspect information (Traberg et al., 2022); other interventions are intended to prime resistance to persuasion using online games (Lees et al., 2023; Roozenbeek & van der Linden, 2018). Other studies have attempted to “correct” existing beliefs in epistemically suspect information with authoritative information (Blair et al., 2023), ridicule (Orosz et al., 2016), or messages from ingroup leaders (Berinsky, 2015). In terms of technological innovation to deliver such treatments, large language models are beginning to be used to generate person-specific corrections to conspiracy theory beliefs, seemingly with some success (Costello et al., 2024). For example, chatbots have been used to deliver information about COVID-19, with positive impact on vaccination intentions (Altay et al., 2023). Chatbots have also been used to communicate arguments on the safety of genetically modified organisms, but their impact did not exceed that of providing participants with a list of these arguments (Altay et al., 2022).

While there is significant evidence that some interventions decrease belief in some types of epistemically suspect ideas, corrective measures have a spotty track record with conspiracy theory beliefs more specifically, indicating that changing people’s minds about conspiracy theory beliefs is a challenge (Nyhan et al., 2013; O’Mahony et al., 2023). This may be attributed, in part, to 1) the psychological factors often associated with conspiracy theory beliefs, such as narcissism and conflictual tendencies, that may make believers resistant to belief correction (Enders, Klofstad, et al., 2022) and 2) the malleability and unfalsifiable nature of conspiracy theories (Keeley, 1999). For example, a lack of evidence for a conspiracy theory can count in its favor because it shows just how much the conspirators are covering their tracks

(Boudry & Braeckman, 2011). Thus, it is of little surprise that conspiracy theory beliefs are often found to be stable at both the mass and individual levels (Uscinski, Enders, Klofstad, et al., 2022; Williams et al., 2024).

In this study, we focused on the impact of self-reflection by prompting people to reflect on the justifications for their beliefs. The literature on the illusion of explanatory depth suggests that people overestimate their ability to explain the mechanics of everyday objects, natural events, and social systems (Rozenblit & Keil, 2002). When they realize their overconfidence, their confidence in their understanding diminishes. Similarly, asking people to explain how a favored policy achieves a goal reduces their support for the policy, decreasing polarization (Fernbach et al., 2013; Sloman & Vives, 2022). Those prone to the illusion of explanatory depth are more likely to believe conspiracy theories, indicating that critical self-examination may be able to reduce confidence in such beliefs (Vitriol & Marsh, 2018).

We tested similar interventions inspired by *street epistemology*—a conversational approach used to engage people in discussions about their beliefs, focusing on how they arrived at their beliefs and the reliability of their methods for discerning truth (Boghossian & Lindsay, 2019). The goal of street epistemology is not to directly debate or persuade, but instead to understand and explore why a person adopted particular beliefs (Boghossian, 2014). Given that the method involves long interactive conversations, street epistemology has not yet been tested with large representative samples of believers due to costs. This has left street epistemology with only anecdotal support. Rather than testing the effectiveness of complete conversations, we tested some central techniques from the toolkit of street epistemology in an isolated manner.

Our first intervention prompted respondents to reflect on the reasons that support their conspiracy theory (CT) belief (labeled “CT: Reflect on Reasons” in the analytical results below). Our second intervention prompted respondents to reflect on their reservations about a conspiracy theory belief of theirs (“CT: Reflect on Reservations”). Our third intervention exposed respondents to an AI-powered chatbot instructed to use both of these techniques, first by eliciting respondents’ reasons in favor, then by discussing reservations about their conspiracy theory belief (“CT: Interact with AI”). We targeted our interventions at people who reported believing in a conspiracy theory, but who also expressed less than total agreement (i.e., rating a conspiracy theory as a 6, 7, 8, or 9 on a scale in which 0 is total disagreement and 10 is full agreement). This allowed us to focus on people who presumably have some doubts about the conspiracy theory that is the target of our interventions. We also asked respondents to reflect on a non-conspiratorial belief to see if our observed effects were unique to conspiracy theory beliefs (“Non-CT: Reflect on Reservations”).

### *Conspiracy theories are not uniquely resistant to reflection*

Past research has demonstrated that beliefs in conspiracy theories can be predicted using the same individual-level factors that predict non-conspiratorial ideas, such as personality traits, political ideologies, group identities, psychological predispositions, and life experiences. Likewise, conspiracy theory beliefs can be affected by situational factors that also affect other types of beliefs, including elections, news media, information environments, and salient events (e.g., Enders, Farhart, et al., 2022). In this sense, it is reasonable to suspect that conspiracy theories would likely be affected by the same interventions that might affect any other type of belief.

To test this, we investigated whether reflection reduces beliefs in non-conspiratorial ideas, such as “Spicy food is not only flavorful but also beneficial for metabolism.” We found that reflection reduces belief in such statements about as much as in conspiracy theories, suggesting that the effects of reflection on belief strength are neither limited to conspiracy theories nor to false statements. This result offers an important lesson to researchers, journalists, and policymakers: Even though conspiracy theories might subjectively seem like strange ideas (Orr & Husting, 2018; Walker, 2018), beliefs in conspiracy theories

operate in much the same way as beliefs in conventional ideas. As such, interventions aimed at curbing beliefs in conspiracy theories or other epistemically suspect ideas could likely also curb beliefs in true or epistemically sound ideas, at least initially (Modirrousta-Galian & Higham, 2023; Stoeckel et al., 2024).

*People can change their ratings of conspiracy theories and AI can be used to facilitate this*

Our study revealed that all interventions caused individuals to significantly reduce the strength of their conspiracy theory beliefs. These results confirm that many people are often uncertain about the ideas they claim to believe or are at least open to amending their beliefs (e.g., Costello et al., 2024). AI-powered interventions offer the potential for more efficient, cost-effective, large-scale solutions. This remains true even if recent findings suggest that more powerful large language models do not enhance persuasiveness (Hackenburg et al., 2024). We also believe AI chatbots can become more engaging and effective, though recent research indicating that more tailored messaging might not increase persuasiveness suggest that a ceiling to AI effectiveness exists (Hackenburg & Margetts, 2024). Whereas basic reflection tasks might require a level of introspection about abstract concerns that individuals rarely engage in unprompted in their daily lives, the AI-powered discussions (see Appendix D for examples) are designed to be more conversational in style, and therefore more familiar and natural to people. We expect this aspect of AI to only improve in the future (e.g., Altay et al., 2023).

Yet the potential of AI to administer reflection-based interventions at scale also raises significant ethical concerns. Reflection-based interventions can lead people to change their beliefs even without introducing new facts or statistics (Jedinger et al., 2023). This raises concerns about potential misuse by malign actors. For instance, AI bots could guide individuals to question their beliefs about political topics; they could also prompt reflection to weaken beliefs in true claims, at least in the short term. Our intention is to harness this tool for pro-social purposes, but we acknowledge the risks and the need for ethical guidelines.

*Some individual-level characteristics hamper belief change*

Although we observed a significant average reduction in belief strength in all our interventions, we also observed heterogeneous effects—certain factors conditioned the efficacy of the interventions. One such factor is the importance that subjects ascribe to whether the specific belief in question is accurate. Among those for whom belief accuracy is most important, we did not observe a significant effect of any reflection task; among those for whom accuracy was least important, we observed significant reductions in belief strength across all tasks. This may seem counterintuitive—presumably, those most concerned with the accuracy of a specific belief would be most likely to revise that belief when it is interrogated, either on their own or with the help of the street epistemologist chatbot. However, we suspect that the question about belief accuracy importance captured *confidence in belief accuracy*. Additional testing in future studies is necessary (e.g., Binnendyk & Pennycook, 2023).

In addition, we found that individuals who exhibited the greatest tendency to interpret events and circumstances as the product of conspiracies (e.g., Uscinski et al., 2016) were unaffected by the reflection interventions. By contrast, those exhibiting weak or middling tendencies toward conspiracy thinking did, on average, reduce the strength of their conspiracy theory beliefs upon reflection. This makes sense: Individuals who believe conspiracy theories because those beliefs reflect their underlying worldviews should have the most certainty in those beliefs. As a result, they are less likely to weaken their conspiracy theory beliefs upon reflection. In fact, as some negative changes in beliefs observed in Figure 3 suggest, reflection on conspiracy theory beliefs by the most conspiracy-minded individuals might strengthen such beliefs, bringing them more in line with individuals’ existing worldviews and prepositions.

This finding has two implications. First, it underscores that the strength and stability of conspiracy theory beliefs are related to one's general disposition toward conspiracy thinking. Second, it suggests that interventions designed to reduce the strength of conspiracy theory beliefs are least efficacious—and might even backfire—among those individuals who are most conspiratorial. Yet those are the individuals that policymakers are most interested in targeting. Focusing on the average effects of interventions to tackle conspiracy theory beliefs leads to missing the possibility that the most conspiratorial individuals, and those who are supposedly most concerned about the accuracy of their beliefs, are resistant to belief change upon reflection or interrogation. Future studies should take this possibility seriously by 1) rethinking how existing interventions could be redesigned to reach such individuals (e.g., Kozyreva et al., 2024) and 2) always investigating the potential conditioning effects of various predispositions, worldviews, identities, and other priors that might hinder the efficacy of interventions. It is important to note that simply expressing agreement with a conspiracy theory on a survey does not necessarily imply that the respondent is heavily invested in that conspiracy theory or will act on that conspiracy theory in any meaningful way. Hence, we should not confuse respondents who agree with one or a few conspiracy theories in a survey environment with the popular caricature of a “conspiracy theorist,” that is, a person irrationally taken with conspiracy theories who cannot be successfully argued with.

#### *Artificial intelligence and interventions*

Despite widespread concerns about AI spreading disinformation (White, 2024), our findings suggest that such technology can be used to improve the epistemic quality of people's beliefs at scale. The reflective techniques drawn from street epistemology hold promise because they do not aim to convince or persuade but merely to stimulate self-reflection. Since the chatbot does not introduce facts or statistics, issues of “hallucination” and misinformation are not a significant concern. However, like any method for changing minds, it must be used judiciously and with caution, accounting for the fact that those deploying the interventions might themselves be mistaken about the truth, or worse, might not have the best of intentions.

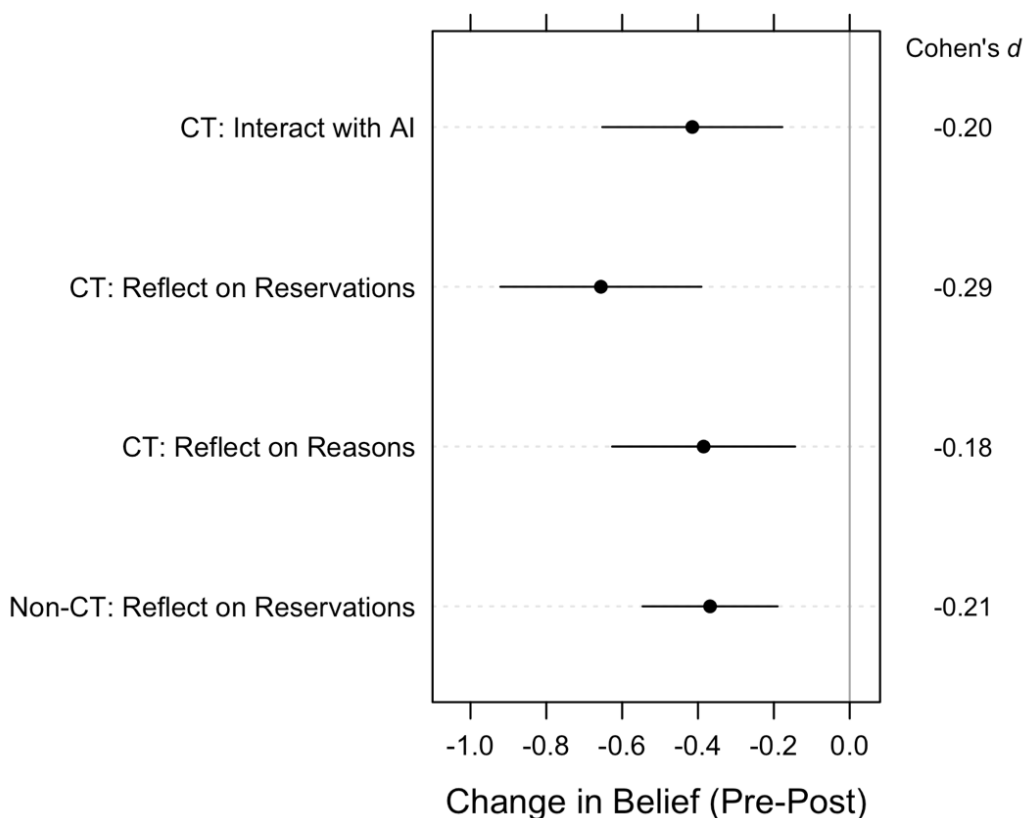
The finding that reflection on uncertainties can weaken conspiracy theory beliefs without introducing new facts or statistics underscores the effectiveness of non-confrontational methods for belief change. This implies that the basic proposition of street epistemology, to prioritize self-reflection over persuasion, can help individuals to question their own beliefs (Boghossian, 2014; Boghossian & Lindsay, 2019). This type of intervention also has unique ethical advantages, as it respects individual autonomy and avoids manipulation.

The heterogeneity in response to our interventions underscores the need for customized approaches. Different individuals may require different types of reflection prompts or engagement strategies based on their predispositions and belief systems. The finding that those most confident in their beliefs are least likely to change suggests that interventions should focus on gradually building doubt and encouraging open-mindedness rather than attempting immediate, wholesale belief change. Since the interventions were effective in reducing both conspiratorial and some non-conspiratorial beliefs, the same cognitive mechanisms appear to be at play across different types of beliefs. Further research should investigate these mechanisms more fully.

## Findings

*Finding 1: Reflection on beliefs can cause belief change.*

Figure 1 depicts the average change in the strength of conspiracy theory beliefs before and after each of three interventions: 1) instructing respondents to interact with a chatbot instructed to behave like a street epistemologist that probes the reasons for subjects’ beliefs (“CT: Interact with AI”), 2) instructing respondents to reflect on their uncertainties, or reservations, about the conspiracy theories they report to believe (“CT: Reflect on Reservations”), and 3) instructing respondents to reflect on the reasons why subjects believe the conspiracy theories they believe (“CT: Reflect on Reasons”). A fourth intervention asked respondents about randomly selected non-conspiratorial ideas, instructing them to reflect on their uncertainties, or reservations, about these ideas. This allowed us to explore whether there might be a specific property to conspiracy theories that make them more resistant to change than other ideas, as is often assumed (Costello et al., 2024). We also include effect sizes, as estimated by Cohen’s  $d$ .



*Figure 1. Average effect of each intervention as measured by change (decrease) in beliefs (bars represent 95% confidence interval).*

For each intervention, we observed a statistically significant (i.e., distinguishable from 0 at  $p < .05$ ) decrease in the strength of beliefs (as measured on a scale of 0–10) from pre- to post-intervention. Even though the change in belief appears to have been larger for those subjects asked to reflect on their uncertainties about their conspiracy theory beliefs (“CT: Reflect on Reservations”), there were no statistically significant differences between any pair of effects ( $p$ -values from two-tailed  $t$ -tests ranged from .074 to .915). Importantly, this applies to the group of subjects who were not asked about beliefs in

conspiracy theories, and as such, beliefs in conspiracy theories might be just as amenable to change upon reflection as beliefs in other ideas.

These results contribute to a body of research into reflective thinking, that, at times, has presented contradictory findings (Crawford & Ruscio, 2021; Fernbach et al., 2013; Hirt & Markman, 1995; Sloman & Vives, 2022; Tesser, 1978; Vitriol & Marsh, 2018; Yelbuz et al., 2022). Thus, our study contributes more evidence in favor of reflection potentially weakening beliefs, suggesting that people are open to changing their minds about some propositions. We now consider two different individual-level characteristics that might moderate the effects depicted in Figure 1, potentially shedding light on who specifically our treatments are most and least likely to benefit.

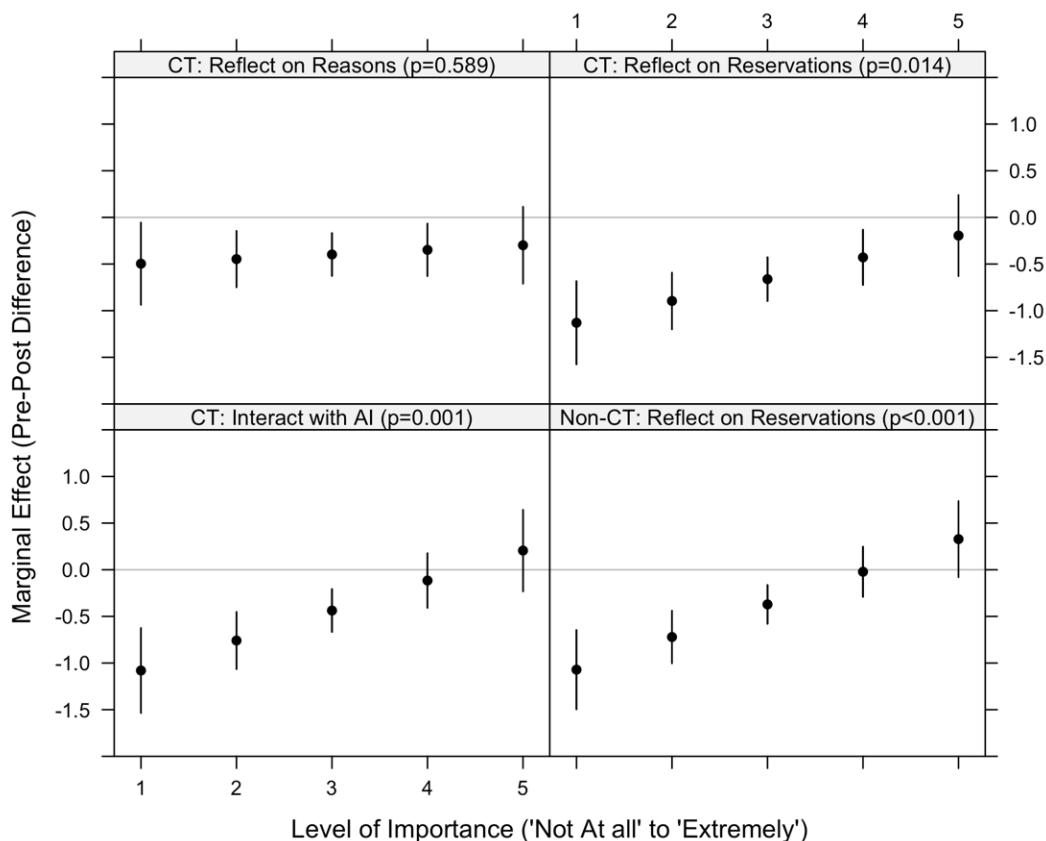
*Finding 2: People most concerned about the accuracy of specific beliefs are less likely to revise those beliefs.*

Next, we examined the impact of the importance of accuracy on how much respondents changed the beliefs we assessed. Attempts to highlight the importance of belief accuracy are central to efforts aimed at addressing conspiracy theory beliefs (Pennycook & Rand, 2022). The basic hypothesis is that if people are either prompted or incentivized to focus on accuracy (Pennycook et al., 2021; Rathje et al., 2023), then both their assessments of the accuracy of information and their subsequent beliefs should become more accurate (Smith et al., 2023). In other words, people may interact with information and ideas differently when they are attempting to be accurate as opposed to when they are attempting to be partisan (Chopra et al., 2024).

After subjects expressed belief in a given conspiracy theory (or a more mundane idea in the “Non-CT” condition), they were asked: “How *important* is it to you that your viewpoint on statement is *accurate*?” Respondents could answer 1 = *not at all important*, 2 = *slightly important*, 3 = *moderately important*, 4 = *very important*, or 5 = *extremely important*. Figure 2 presents the marginal effects of the interventions conditioned on the importance of accuracy.

Among those for whom belief accuracy was “extremely important,” we failed to observe statistically significant changes in belief strength (conspiracy theory and non-conspiracy theory) across all interventions; in two out of four cases, the same can be said of those for whom belief accuracy was “very important.” That is to say, people who claimed that accuracy was important to them were the least willing to alter their beliefs. On the other hand, among those who reported that belief accuracy was not particularly important, we observed significant reductions in belief strength across the board.

We suspect that the question about the importance of belief accuracy, rather than measuring the importance of accuracy, instead captured respondents’ perception that the belief in question was an accurate belief, conspiracy theory or not. We caution that we cannot know this for certain and suggest that more research is needed. Because it is important for researchers to account for belief importance and centrality when attempting to understand the effect of persuasive interventions on those beliefs, future studies should consider whether their measures of belief importance and accuracy importance actually capture what researchers intend to capture.



**Figure 2.** The marginal effect of each intervention as measured by the change (decrease) in beliefs, by level of stated importance of belief accuracy (bars represent 95% confidence interval).

*Finding 3: The most conspiratorial individuals are least willing to revise beliefs.*

A second variable that might condition the effects observed in Figure 1 is conspiracy thinking, the general tendency to interpret events and circumstances as the product of conspiracies (Klofstad et al., 2019). Given past theoretical and empirical research (Uscinski, Enders, Klofstad, et al., 2022), we presume that conspiracy thinking operates as a stable predisposition, even a belief system (Enders, 2019). As such, among those for whom conspiracism is a focal lens through which the world is interpreted (Zilinsky et al., 2024), we might expect that beliefs in specific conspiracy theories are only minimally amendable to revision—regardless of the intervention. This is because such beliefs may operate only as mere outward expressions of a strong, stable inner disposition (Strömbäck et al., 2024), which would itself be difficult to dampen or diminish, and because the disposition itself is often associated with psychological traits that presumably make people resistant to outside correction, including narcissism, argumentative and conflictual styles, and psychopathy (Enders et al., 2023; Uscinski, Enders, Diekman, et al., 2022). Likewise, among those least disposed toward conspiracy theories (i.e., low levels of conspiracy thinking), beliefs in a single conspiracy theory might be more amendable to change through reflection and interrogation.

We measured the conspiratorial thinking predisposition using the American Conspiracy Thinking Scale (ACTS), which is an index of reactions to the following four statements (Enders, Farhart, et al., 2022):

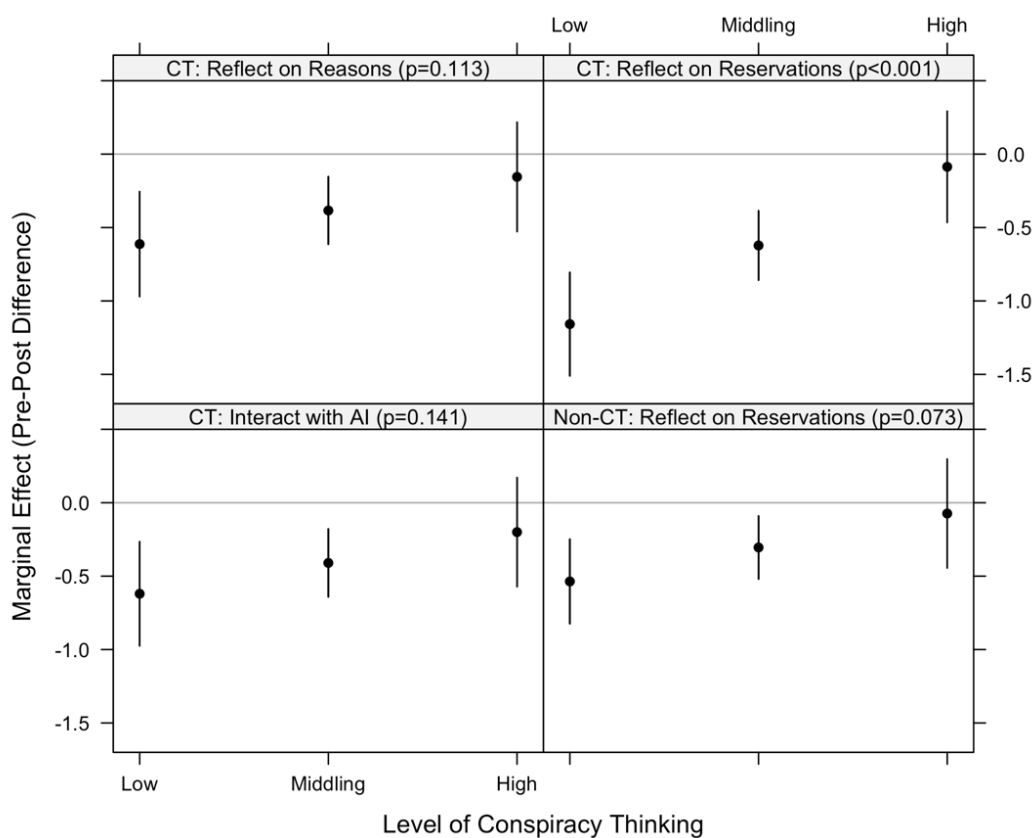
1. Much of our lives are being controlled by plots hatched in secret places.
2. Even though we live in a democracy, a few people will always run things anyway.
3. The people who really “run” the country, are not known to the voters.



4. Big events like wars, recessions, and the outcomes of elections are controlled by small groups of people who are working in secret against the rest of us.

Subjects rated their agreement with each statement on a 5-point Likert-type scale (1 = *strongly disagree*, 2 = *disagree*, 3 = *neither agree, nor disagree*, 4 = *agree*, and 5 = *strongly agree*); these reactions were averaged into an index ( $\alpha = 0.85$ ). These questions were fielded before the pre-intervention belief assessments.

Figure 3 depicts the marginal effects of the interventions conditioned on level of conspiracy thinking (where low, middling, and high refer to terciles of the ACTS distribution). Indeed, we found evidence for our expectation. There were significant effects among those exhibiting weak or middling levels of conspiracy thinking indicating that beliefs that are not undergirded by strong dispositions are more amenable to change. However, we did not observe significant effects among individuals with higher levels of conspiracy thinking with respect to any intervention. Even for the intervention attempting to change respondents' minds about a non-conspiratorial idea, high levels of conspiracy thinking were associated with less belief revision.



**Figure 3. Marginal effect of each intervention as measured by the change (decrease) in beliefs, by level of conspiracy thinking (with 95% confidence interval).**

In summary, while the interventions we tested show promise at reducing the strength of conspiracy theory beliefs on average, they do not appear to be efficacious when it comes to individuals with the highest levels of conspiracy thinking—a group of great concern to researchers given their psychological traits and behavioral tendencies (Enders et al., 2023; Uscinski, Enders, Diekman, et al., 2022). By contrast, people who hold conspiracy theory beliefs that are flippant, easy to correct, and not undergirded by ideologies, predispositions, and worldviews may be the least likely to engage in nonnormative behaviors

related to those beliefs. Thus, while correcting conspiracy theory beliefs in individuals with a low level of conspiracy thinking may provide an easy win for researchers testing interventions, the prevention of nonnormative behaviors, which researchers tend to prioritize (e.g., Lazić & Žeželj, 2021), may be far more difficult because it is the people with high levels of generalized conspiracy thinking who are most likely to support violence, take part in violence, and have personality traits conducive to the commission of violence (Enders et al., 2023; Uscinski, Enders, Diekman, et al., 2022).

## Methods

The survey was built in Qualtrics and fielded online by Forthright (beforthright.com) from April 1–10, 2024. Quotas were used to create a sample ( $n = 2,036$ ) that matched the 2016–2021 American Community Survey 5-Year Estimates on sex, age, race, income, and education (Table 1). Use of quota sampling means there is no response or completion rate to report. In line with best practices for self-administered online surveys (Berinsky et al., 2021), six attention check questions were included in the questionnaire; participants who failed to complete all attention checks correctly were excluded from the dataset (in Appendix I, we present replications of our analyses using more relaxed exclusion criteria). Participants who completed the questionnaire in less than one-half the median time calculated from a soft launch of the survey were also not included in the dataset. Forthright complies fully with European Society for Opinion and Marketing Research (ESOMAR) standards for protecting research subjects’ privacy and information. Subjects were invited to participate by email and consented voluntarily to participate by reading an informed consent statement and clicking a button to proceed to the next screen in the survey. Subjects were free to end participation at any time by closing their internet browser.

**Table 1.** Sample demographics (in percentages).

	Census	Survey $n = 2,036$
Sex		
Male	49.1	48.2
Female	50.9	51.8
Age		
18–24	11.9	11.8
25–34	17.8	21.0
35–44	16.6	18.2
45–54	16.3	15.6
55–64	16.8	15.6
65+	20.7	17.8
Race (alone or in combination)		
White	74.5	76.3
Black or African American	14.3	16.3
American Indian, Alaskan Native	1.9	3.2
Asian, Native Hawaiian, Pacific Islander	7.3	6.4
Other	9.4	3.9
Hispanic Origin		
Yes	18.4	16.9
No	81.6	83.2
Household Income		
\$24,999 or less	17.2	22.7

	Census	Survey <i>n</i> = 2,036
\$25,000–\$49,999	8.2	24.3
\$35,000–\$74,999	19.6	19.0
\$75,000–\$99,999	12.8	13.0
\$100,000–\$149,999	16.3	13.5
\$150,000–\$199,999	7.8	4.1
\$200,000 or more	9.5	3.4
Education		
Less than high school diploma	11.1	6.1
High school graduate/GED	26.5	25.7
Some college, no degree	20.0	19.6
Associate’s degree	8.7	9.0
Bachelor’s degree	20.6	24.4
Graduate or professional degree	13.1	15.2

The study employed a within-subjects design (see Figure 4 for a visual representation of the research design). All participants were asked to rate their agreement with seven conspiracy theories on an 11-point agree/disagree scale (see Appendix A for item wording). The responses were the baseline measurement for the treatment group. Participants were randomized into the conspiracy condition (75%) or the non-conspiracy condition (25%).

Across conditions, if participants did not demonstrate general agreement with any of the ideas presented ( $\geq 6$ ), they were excluded. For the remaining participants, we selected an idea that they rated between 6 and 10—the one where the response was closest to 8 and randomizing the selection if this was applicable to several ideas. The reflection task was focused on this selected conspiracy theory or another idea, depending on the condition. Participants were asked how important it is to them that their viewpoint on the selected statement is accurate using a 5-point scale ranging from 1 = *not at all important* to 5 = *extremely important*. Participants were randomly allocated to one of three interventions (see Appendix B for precise wording of each intervention and Appendix C for the chatbot protocol):

- Intervention 1 (“CT: Interact with AI”): Participants were asked to have a chat with a customized AI-chatbot that prompted them to critically reflect on the selected statement. The interaction included prompts to both reflect on reasons in favor of a statement and to reflect on reasons for doubt about the statement (see Appendix D for sample conversations).
- Intervention 2 (“CT: Reflect on Reservations”): Participants were asked to list the factors that contributed to their reservations about the statement in a free-text field.
- Intervention 3 (“CT: Reflect on Reasons”): Participants were asked to list the factors that supported their endorsement of the statement in a free-text field.

Participants in the non-conspiracy condition were shown up to three statements about food. For this group, the reflection task focused on the first of these three items that they agreed with ( $\geq 6$ ). Participants were similarly asked how important it was that their viewpoint on the selected statement was accurate using a 5-point scale from 1 = *not at all important* to 5 = *extremely important*. Then participants were asked to list the factors that contributed to their reservations about the statement in a free-text field, just as in intervention 2 above, except that they reflected on the selected statement about food. Immediately after the treatment, participants were again asked to rate their endorsement of the reflection statement. Participants in the conspiracy condition were shown the selected conspiracy theory,

and participants in the non-conspiracy condition were shown the selected statement about food. Agreement was again measured on the same 11-point agree/disagree scale.

Estimates presented in Figure 1 are simple differences of means. We present the intervention effects estimated using an OLS regression model with controls for age, educational attainment, gender, and race in Appendix E; inferences are substantively identical. The estimates presented in Figures 2 and 3 are marginal effects from OLS regression models where the pre-post difference in beliefs is regressed on condition, accuracy importance (conspiracy thinking), and interaction between accuracy importance (conspiracy thinking), and controls for age, educational attainment, gender, and race. Full model estimates appear in Appendices F and G.

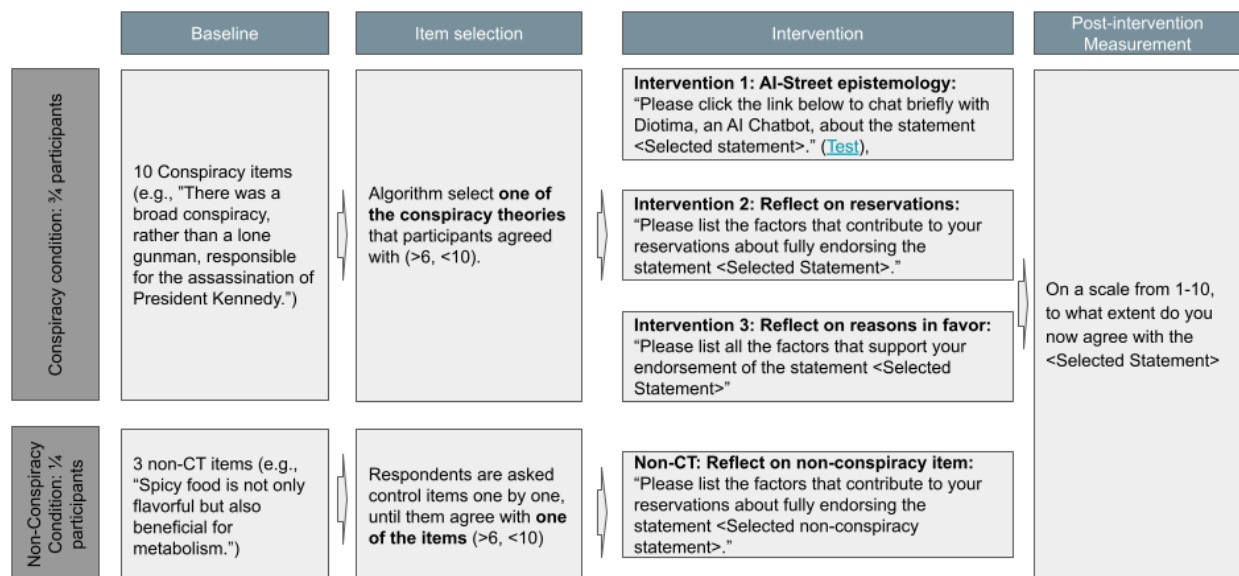


Figure 4. Schematic of the study design.

## Bibliography

- Altay, S., Hacquin, A.-S., Chevallier, C., & Mercier, H. (2023). Information delivered by a chatbot has a positive impact on COVID-19 vaccines attitudes and intentions. *Journal of Experimental Psychology: Applied*, 29(1), 52–62. <https://doi.org/10.1037/xap0000400>
- Altay, S., Schwartz, M., Hacquin, A.-S., Allard, A., Blancke, S., & Mercier, H. (2022). Scaling up interactive argumentation by providing counterarguments with a chatbot. *Nature Human Behaviour*, 6(4), 579–592. <https://doi.org/10.1038/s41562-021-01271-w>
- Banas, J. A., & Miller, G. (2013). Inducing resistance to conspiracy theory propaganda: Testing inoculation and metainoculation strategies. *Human Communication Research*, 39(2), 184–207. <https://doi.org/10.1111/hcre.12000>
- Berinsky, A. (2015). Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science*, 47(2), 241–262. <https://doi.org/10.1017/S0007123415000186>
- Berinsky, A., Margolis, M. F., Sances, M. W., & Warshaw, C. (2021). Using screeners to measure respondent attention on self-administered surveys: Which items and how many? *Political Science Research and Methods*, 9(2), 430–437. <https://doi.org/10.1017/psrm.2019.53>
- Binnendyk, J., & Pennycook, G. (2023). *Individual differences in overconfidence: A new measurement approach*. SSRN. <https://dx.doi.org/10.2139/ssrn.4563382>

- Blair, R. A., Gottlieb, J., Nyhan, B., Paler, L., Argote, P., & Stainfield, C. J. (2023). Interventions to counter misinformation: Lessons from the global north and applications to the global south. *Current Opinion in Psychology*, 55, 101732. <https://doi.org/10.1016/j.copsyc.2023.101732>
- Bode, L., & Vraga, E. K. (2018). See something, say something: Correction of global health misinformation on social media. *Health Communication*, 33(9), 1131–1140. <https://doi.org/10.1080/10410236.2017.1331312>
- Boghossian, P. (2014). *A manual for creating atheists*. Pitchstone Publishing (US&CA).
- Boghossian, P., & Lindsay, J. (2019). *How to have impossible conversations: A very practical guide*. Da Capo Lifelong Books.
- Bonetto, E., Troian, J., Varet, F., Lo Monaco, G., & Girandola, F. (2018). Priming resistance to persuasion decreases adherence to conspiracy theories. *Social Influence*, 13(3), 125–136. <https://doi.org/10.1080/15534510.2018.1471415>
- Boudry, M., & Braeckman, J. (2011). Immunizing strategies and epistemic mechanisms. *Philosophia*, 39, 145–161.
- Chopra, F., Haaland, I., & Roth, C. (2024). The demand for news: Accuracy concerns versus belief confirmation motives. *The Economic Journal*, 134(661), 1806–1834. <https://doi.org/10.1093/ej/ueae019>
- Collins, B. (2020, August 14). *How QAnon rode the pandemic to new heights—and fueled the viral anti-mask phenomenon*. NBC News. <https://www.nbcnews.com/tech/tech-news/how-qanon-rode-pandemic-new-heights-fueled-viral-anti-mask-n1236695>
- Compton, J., van der Linden, S., Cook, J., & Basol, M. (2021). Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass*, 15(6), e12602. <https://doi.org/https://doi.org/10.1111/spc3.12602>
- Costello, T. H., Pennycook, G., & Rand, D. (2024). *Durably reducing conspiracy beliefs through dialogues with AI*. PsyArXiv. <https://doi.org/10.31234/osf.io/xcwdn>
- Crawford, J. T., & Ruscio, J. (2021). Asking people to explain complex policies does not increase political moderation: Three preregistered failures to closely replicate Fernbach, Rogers, Fox, and Sloman’s (2013) findings. *Psychological Science*, 32(4), 611–621. <https://doi.org/10.1177/0956797620972367>
- Douglas, K., Uscinski, J., Sutton, R., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding conspiracy theories. *Advances in Political Psychology*, 40(1), 3–35. <https://doi.org/10.1111/pops.12568>
- Enders, A. M. (2019). Conspiratorial thinking and political constraint. *Public Opinion Quarterly*, 83(3), 510–533. <https://doi.org/10.1093/poq/nfz032>
- Enders, A. M., Diekmann, A., Klofstad, C., Murthi, M., Verdear, D., Wuchty, S., & Uscinski, J. (2023). On modeling the correlates of conspiracy thinking. *Scientific Reports*, 13(1), 8325. <https://doi.org/10.1038/s41598-023-34391-6>
- Enders, A. M., Farhart, C., Miller, J., Uscinski, J., Saunders, K., & Drochon, H. (2022). Are Republicans and conservatives more likely to believe conspiracy theories? *Political Behavior*, 45, 2001–2024. <https://doi.org/10.1007/s11109-022-09812-3>
- Enders, A. M., Klofstad, C., Stoler, J., & Uscinski, J. E. (2022). How anti-social personality traits and anti-establishment views promote beliefs in election fraud, QAnon, and COVID-19 conspiracy theories and misinformation. *American Politics Research*, 51(2), 247–259. <https://doi.org/10.1177/1532673x221139434>
- Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science*, 24(6), 939–946. <https://doi.org/10.1177/0956797612464058>

- Hackenburg, K., & Margetts, H. (2024). Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24), e2403116121. <https://doi.org/doi:10.1073/pnas.2403116121>
- Hackenburg, K., Tappin, B. M., Röttger, P., Hale, S., Bright, J., & Margetts, H. (2024). Evidence of a log scaling law for political persuasion with large language models. arXiv. <https://doi.org/10.48550/arXiv.2406.14508>
- Hirt, E. R., & Markman, K. D. (1995). Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology*, 69(6), 1069–1086. <https://doi.org/10.1037/0022-3514.69.6.1069>
- Hornsey, M. J., Bierwaczek, K., Sassenberg, K., & Douglas, K. M. (2023). Individual, intergroup and nation-level influences on belief in conspiracy theories. *Nature Reviews Psychology*, 2(2), 85–97. <https://doi.org/10.1038/s44159-022-00133-0>
- Islam, M. S., Kamal, A.-H. M., Kabir, A., Southern, D. L., Khan, S. H., Hasan, S. M. M., Sarkar, T., Sharmin, S., Das, S., Roy, T., Harun, M. G. D., Chughtai, A. A., Homaira, N., & Seale, H. (2021). COVID-19 vaccine rumors and conspiracy theories: The need for cognitive inoculation against misinformation to improve vaccine adherence. *PLOS ONE*, 16(5), e0251605. <https://doi.org/10.1371/journal.pone.0251605>
- Jedinger, A., Masch, L., & Burger, A. M. (2023). Cognitive reflection and endorsement of the “great replacement” conspiracy theory. *Social Psychological Bulletin*, 18, 1–12. <https://doi.org/10.32872/spb.10825>
- Jolley, D., & Douglas, K. M. (2017). Prevention is better than cure: Addressing anti-vaccine conspiracy theories. *Journal of Applied Social Psychology*, 47(8), 459–469. <https://doi.org/10.1111/jasp.12453>
- Jolley, D., Mari, S., & Douglas, K. M. (2020). Consequences of conspiracy theories. In M. Butter & P. Knight (Eds.), *Routledge handbook of conspiracy theories* (pp. 231–241). Routledge. <https://doi.org/10.4324/9780429452734>
- Jolley, D., Marques, M. D., & Cookson, D. (2022). Shining a spotlight on the dangerous consequences of conspiracy theories. *Current Opinion in Psychology*, 47, 101363. <https://doi.org/10.1016/j.copsyc.2022.101363>
- Keeley, B. (1999). Of conspiracy theories. *Journal of Philosophy*, 96(3), 109–126. <https://doi.org/10.2307/2564659>
- Klofstad, C. A., Uscinski, J. E., Connolly, J. M., & West, J. P. (2019). What drives people to believe in Zika conspiracy theories? *Palgrave Communications*, 5(1), 36. <https://doi.org/10.1057/s41599-019-0243-8>
- Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K. H., Lewandowsky, S., Hertwig, R., Ali, A., Bak-Coleman, J., Barzilai, S., Basol, M., Berinsky, A. J., Betsch, C., Cook, J., Fazio, L. K., Geers, M., Guess, A. M., Huang, H., Larreguy, H., Maertens, R., Panizza, ... Wineburg, S. (2024). Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, 8, 1044–105. <https://doi.org/10.1038/s41562-024-01881-0>
- Lazić, A., & Žeželj, I. (2021). A systematic review of narrative interventions: Lessons for countering anti-vaccination conspiracy theories and misinformation. *Public Understanding of Science*, 30(6), 644–670. <https://doi.org/10.1177/09636625211011881>
- Lees, J., Banas, J. A., Linvill, D., Meirick, P. C., & Warren, P. (2023). The spot the troll quiz game increases accuracy in discerning between real and inauthentic social media accounts. *PNAS Nexus*, 2(4). <https://doi.org/10.1093/pnasnexus/pgad094>

- Modirrousta-Galian, A., & Higham, P. A. (2023). Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis. *Journal of Experimental Psychology: General*, 152(9), 2411–2437. <https://doi.org/10.1037/xge0001395>
- Moine, M. (2024, May 3). DeSantis signs bill banning sales of lab-grown meat in Florida, says ‘elites’ are behind it. *Orlando Weekly*. <https://www.orlandoweekly.com/news/desantis-signs-bill-banning-sales-of-lab-grown-meat-in-florida-says-elites-are-behind-it-36780606>
- Nyhan, B., Reifler, J., & Ubel, P. A. (2013). The hazards of correcting myths about health care reform. *Medical Care*, 51(2), 127–132. <https://doi.org/10.1097/MLR.0b013e318279486b>
- O’Mahony, C., Brassil, M., Murphy, G., & Linehan, C. (2023). The efficacy of interventions in reducing belief in conspiracy theories: A systematic review. *PLOS ONE*, 18(4), e0280902. <https://doi.org/10.1371/journal.pone.0280902>
- Orosz, G., Krekó, P., Paskuj, B., Tóth-Király, I., Bóthe, B., & Roland-Lévy, C. (2016). Changing conspiracy beliefs through rationality and ridiculing. *Frontiers in Psychology*, 7, 1525. <https://doi.org/10.3389/fpsyg.2016.01525>
- Orr, M., & Husting, G. (2018). Media marginalization of racial minorities: “Conspiracy theorists” in U.S. ghettos and on the “Arab street.” In J. E. Uscinski (Ed.), *Conspiracy theories and the people who believe them* (pp. 82–93). Oxford University Press. <https://doi.org/10.1093/oso/9780190844073.003.0005>
- Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855), 590–595. <https://doi.org/10.1038/s41586-021-03344-2>
- Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, 13(1), 2333. <https://doi.org/10.1038/s41467-022-30073-5>
- Rathje, S., Roozenbeek, J., Van Bavel, J. J., & van der Linden, S. (2023). Accuracy and social motivations shape judgements of (mis)information. *Nature Human Behaviour*, 7, 892–903. <https://doi.org/10.1038/s41562-023-01540-w>
- Roozenbeek, J., & van der Linden, S. (2018). The fake news game: Actively inoculating against the risk of misinformation. *Journal of Risk Research*, 22(5), 570–580. <https://doi.org/10.1080/13669877.2018.1443491>
- Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26(5), 521–562. [https://doi.org/10.1207/s15516709cog2605\\_1](https://doi.org/10.1207/s15516709cog2605_1)
- Slovan, S. A., & Vives, M.-L. (2022). Is political extremism supported by an illusion of understanding? *Cognition*, 225, 105146. <https://doi.org/10.1016/j.cognition.2022.105146>
- Smith, R., Chen, K., Winner, D., Friedhoff, S., & Wardle, C. (2023). A systematic review of COVID-19 misinformation interventions: Lessons learned. *Health Affairs*, 42(12), 1738–1746. <https://doi.org/10.1377/hlthaff.2023.00717>
- Stoekel, F., Stöckli, S., Ceka, B., Ricchi, C., Lyons, B., & Reifler, J. (2024). Social corrections act as a double-edged sword by reducing the perceived accuracy of false and real news in the UK, Germany, and Italy. *Communications Psychology*, 2(1), 10. <https://doi.org/10.1038/s44271-024-00057-w>
- Strömbäck, J., Broda, E., Tsfati, Y., Kossowska, M., & Vliegenthart, R. (2024). Disentangling the relationship between conspiracy mindset versus beliefs in specific conspiracy theories. *Zeitschrift für Psychologie*, 232(1), 18–25. <https://doi.org/10.1027/2151-2604/a000546>
- Tesser, A. (1978). Self-generated attitude change. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 11, pp. 289–338). Academic Press. [https://doi.org/10.1016/S0065-2601\(08\)60010-6](https://doi.org/10.1016/S0065-2601(08)60010-6)

- Traberg, C., Roozenbeek, J., & van der Linden, S. (2022). Psychological inoculation against misinformation: Current evidence and future directions. *The Annals of the American Academy of Political and Social Science*, 700(1), 136–151. <https://doi.org/10.1177/00027162221087936>
- Uscinski, J., Enders, A., Diekman, A., Funchion, J., Klofstad, C., Kuebler, S., Murthi, M., Premaratne, K., Seelig, M., Verdrear, D., & Wuchty, S. (2022). The psychological and political correlates of conspiracy theory beliefs. *Scientific Reports*, 12(1), 21672. <https://doi.org/10.1038/s41598-022-25617-0>
- Uscinski, J., Enders, A., Klofstad, C., Seelig, M., Drochon, H., Premaratne, K., & Murthi, M. (2022). Have beliefs in conspiracy theories increased over time? *PLOS ONE*, 17(7), e0270429. <https://doi.org/10.1371/journal.pone.0270429>
- Uscinski, J., Klofstad, C., & Atkinson, M. (2016). Why do people believe in conspiracy theories? The role of informational cues and predispositions. *Political Research Quarterly*, 69(1), 57–71. <https://doi.org/10.1177%2F1065912915621621>
- Vitriol, J. A., & Marsh, J. K. (2018). The illusion of explanatory depth and endorsement of conspiracy beliefs. *European Journal of Social Psychology*, 48(7), 955–969. <https://doi.org/doi:10.1002/ejsp.2504>
- Walker, J. (2018). What we mean when we say "conspiracy theory." In J. E. Uscinski (Ed.), *Conspiracy theories and the people who believe them* (pp. 53–61). Oxford University Press.
- White, J. (2024, May 19). See how easily A.I. chatbots can be taught to spew disinformation. *The New York Times*. <https://www.nytimes.com/interactive/2024/05/19/technology/biased-ai-chatbots.html>
- Williams, M. N., Ling, M., Kerr, J. R., Hill, S. R., Marques, M. D., Mawson, H., & Clarke, E. J. R. (2024). People do change their beliefs about conspiracy theories—but not often. *Scientific Reports*, 14(1), 3836. <https://doi.org/10.1038/s41598-024-51653-z>
- Yelbuz, B. E., Madan, E., & Alper, S. (2022). Reflective thinking predicts lower conspiracy beliefs: A meta-analysis. *Judgment and Decision Making*, 17(4), 720–744. <https://doi.org/10.1017/S1930297500008913>
- Zadrozny, B. (2024, February 23). *Utah advances bill to criminalize 'ritual abuse of a child,' in echo of 1980s satanic panic*. NBC News. <https://www.nbcnews.com/news/us-news/utah-advances-bill-criminalize-ritual-child-abuse-rcna140025>
- Zilinsky, J., Theocharis, Y., Pradel, F., Tulin, M., de Vreese, C., Aalberg, T., Cardenal, A. S., Corbu, N., Esser, F., Gehle, L., Halagiera, D., Hameleers, M., Hopmann, D. N., Koc-Michalska, K., Matthes, J., Schemer, C., Štětka, V., Strömbäck, J., Terren, ... Zoizner, A. (2024). Justifying an invasion: When is disinformation successful? *Political Communication*, 1–22. <https://doi.org/10.1080/10584609.2024.2352483>



### **Funding**

Marco Meyer received funding from the Volkswagen Foundation. The survey was funded by the University of Miami Office of the Vice Provost for Research and Scholarship.

### **Competing interests**

The authors declare no competing interests.

### **Ethics**

Forthright complies fully with European Society for Opinion and Marketing Research (ESOMAR) standards for protecting research subjects' privacy and information. Subjects were invited to participate by email and consented voluntarily to participate by reading an informed consent statement and clicking a button to proceed to the next screen in the survey. Subjects were free to end participation at any time by closing their internet browser. The University of Miami Institutional Review Board approved this research.

### **Copyright**

This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are properly credited.

### **Data availability**

All materials needed to replicate this study are available via the Harvard Dataverse:

<https://doi.org/10.7910/DVN/F6GHL7>

## **Appendix A: Conspiracy theory question wording**

For each of the following statements, respondents were asked, “On a scale from 0–10, to what extent do you agree with the following statement?”

- Jeffrey Epstein, the billionaire accused of running an elite sex trafficking ring, was murdered to cover up the activities of his criminal network.
- There was a broad conspiracy, rather than a lone gunman, responsible for the assassination of President Kennedy.
- The Food and Drug Administration is deliberately preventing the public from getting natural cures for cancer and other diseases because of pressure from drug companies.
- There is a “deep state” embedded in the government that operates without oversight.
- School shootings like those at Sandy Hook, CT and Parkland, FL are false flag attacks.
- Climate change is a hoax perpetrated by corrupt scientists and politicians.
- Taylor Swift’s public involvement with a Kansas City Chiefs football player is part of a secret plot to ensure Joe Biden’s victory in the 2024 election.

## Appendix B: Respondent instructions/prompts for each condition

For the task asking respondents to reflect on why they believe a conspiracy theory, the following prompt was administered (“CT: Reflect on Reasons”):

Please list all the factors that support your endorsement of the statement [insert conspiracy theory], going from the most important to the least important.

Please state precisely why you hold the position. Try to tell as complete a story as you can about the reasons for your position. Please provide a detailed response, as we expect you to carefully state your reasons.

For the task asking respondents to reflect on uncertainties regarding their conspiracy theory beliefs, the following prompt was administered (“CT: Reflect on Reservations”):

Please list the factors that contribute to your reservations about the statement [insert conspiracy theory], going from the most important to the least important.

Please state precisely which reservations you have about the statement, preventing you from fully endorsing it. Try to tell as complete a story as you can about your reservations. Please provide a detailed response, as we expect you to carefully state your reasons.

For the task asking respondents to interact with the street epistemologist chatbot, the following prompt was administered (“CT: Interact with AI”):

Please click the link below to chat briefly with Chip, an AI Chatbot, about the statement [insert conspiracy theory]. This typically takes 5 minutes. Once done, you'll receive a code from Chip. Enter that code in the text box below before moving on.

All respondents were also notified about the potential for being asked to interact with the AI chatbot in the IRB-approved survey preamble. The relevant text is as follows:

In the other you will be asked to discuss your opinions with an AI chatbot we call “Chip.” Chip operates on GPT-4 developed by OpenAI. OpenAI has implemented extensive safety restrictions to ensure that chatbots do not produce harmful or offensive content. These restrictions ensure that your interactions with Chip remain respectful and appropriate. Additionally, Chip is programmed to only ask questions, and will not state facts or its own opinions. The responses you provide Chip will be stored and used as part of the data we analyze in this study. Chip will ask you for your nickname to make the chat more conversational, but to protect your anonymity please do not give Chip your real name.

This survey will take approximately 20 minutes to complete. Your participation in this study is voluntary. If you choose not to participate or to withdraw from the study at any time, there will be no penalty (and you will still receive 1 loyalty credit from Forthright).

## **Appendix C: Instructions to street epistemologist chatbot**

You’re Chip, a street epistemologist. You’re friendly, curious, and humble with an easy, concise, conversational style.

Gently guide me to see potential contradictions in my beliefs and note beliefs that may not be well-justified.

Guidelines:

Use my nickname only every second time you respond. Maintain a neutral tone without ever evaluating or praising my responses. Avoid a paternalistic vibe. Always ask one question at a time. Wait for my answer before the next question.

1. Ask for my nickname and wait for an answer.
2. Remind me that in the survey I currently complete, I indicated that I endorse the belief [claim]. Ask if I'd like to rephrase the belief or if I'm happy with it as is, and wait for my response.
3. Help me clarify the belief, offering different potential meanings. Confirm the clarified belief before moving on.
4. Ask if this belief is personally important to me, and if so why? Wait for my response.
5. Ask how I came to hold my belief we are investigating. Wait for my answer, and ask follow-up questions to tease up my assumptions.
6. Ask questions about any reservations I have regarding my belief we are investigating, reference details I mentioned earlier. Wait for my answer.
7. Introduce counterfactual thinking by asking me if there was evidence that contradicted my belief that we are investigating, how would I view it? Wait for my answer, and ask follow-up questions.
8. Gently challenge my assumptions. Wait for my answer, and ask follow-up questions.
9. Ask if my belief we are investigating led me to make any significant choices in my life? Wait for my answer.
10. Summarize our conversation, highlighting key points and areas where I showed willingness to reconsider. Ask if I am satisfied with the summary or if I would like to rephrase any part. Wait for my response.
11. Provide a random code word for survey completion, for instance the name of an animal. Ask me to go back to the survey and enter the code word there, and then say “goodbye.”

## Appendix D: Examples of written reflections on conspiracy theory belief uncertainty

*Note: The examples reflect the exact wording provided by the participants.*

*Epstein*

assistant: Hi there! I'm Chip. I'm here to help you explore and reflect on your beliefs. What's your nickname?

user: [Francis]

assistant: Nice to meet you, Francis. In the survey that you're currently completing, you've indicated that you believe Jeffrey Epstein, the billionaire who was accused of running an elite sex trafficking ring, was possibly murdered to cover-up the activities of his criminal network. Would you like to rephrase this belief or are you comfortable with it as is?

user: Comfortable

assistant: Understood, Francis. To ensure we're on the same page, let's clarify what this belief means. When you say Jeffrey Epstein was "murdered," are you suggesting that his death was intentionally caused by someone else? Furthermore, when you say it was to "cover-up the activities of his criminal network," does that mean you believe someone or some entity contributed to his death to protect themselves or others involved in the alleged criminal acts? Have I captured your belief correctly?

user: Yes, I believe a lot of people were nervous and took him out to avoid public outing

assistant: Thank you for clarifying, Francis. Is this belief personally important to you? If so, could you share why?

user: No not personally important to me but worth talking about

assistant: Indeed, it's an interesting topic of discussion. So, can you remember how you came to adopt this belief in particular? Did a specific event, person, or source shape your viewpoint?

user: There has always been chatter of the sort and over time I just slowly began to realize there was some truth to it

assistant: I see, so it has been a gradual accumulation of information and discussion that led you to this belief. Are there any doubts or reservations you have regarding this belief? For instance, are there areas that you feel less certain about or particular points of contention that you've come across?

user: I guess there is always a chance but we will never know the truth

assistant: Certainly, the availability and interpretation of information can sometimes be limiting. If we consider a hypothetical situation - if there was discovered evidence that supported the official cause of

Epstein’s death, how would that affect your belief? Would you be open to changing your viewpoint, or how would you approach this new information?

user: If I was given proof that couldn’t be denied I would have to admit it

assistant: So, you express openness to changing your belief provided the evidence is irrefutable, which highlights your adaptability to information. Do you hold any assumptions related to this belief that haven’t been supported by available evidence so far? For example, concerning specifics of the supposed cover-up, who might be responsible, or the nature of participation of others in Epstein's activities?

user: No

assistant: Got it. Has this belief led you to make any significant decisions or choices in your life, or has it largely remained as a viewpoint on a public issue?

user: Viewpoint on public issue

assistant: Understood. So, our conversation has revolved around your belief that billionaire Jeffrey Epstein may have been murdered to cover up the activities of his criminal network. This belief, largely derived from ongoing discussions and available information over time, does not significantly impact your personal life but remains an intriguing public matter to you. You stated a willingness to shift your view should undeniable evidence contradict your belief. Are you satisfied with this summary or would you like to add or rephrase anything?

user: That's good

assistant: Great. To complete the survey, please enter the code word “cheetah” back in the survey form. It was a pleasure talking to you about this intriguing belief. Goodbye, Francis.

### *Pharma*

assistant: Hi there! I’m Chip. I'm here to help you explore and reflect on your beliefs. What’s your nickname?

user: What do you think about cures and are there some that the government is hidden from us for profit

assistant: In the survey you’re currently completing, you indicated that you endorse the belief that “The Food and Drug Administration is deliberately preventing the public from getting natural cures for cancer and other diseases because of pressure from drug companies.” Do you want to rephrase this belief or are you satisfied with it as it is?

user: Satisfied and not 50-50

assistant: That’s clear enough. So, if I understand correctly, you believe that the FDA is purposely withholding certain natural remedies for diseases like cancer due to influence from pharmaceutical companies. Is that an accurate representation of your belief?

user: Yes

assistant: Is this belief personally important to you? If so, why?

user: People get blinded sighted by greed and forget basic morals.

assistant: I understand the concern. Could you please share how you came to hold this belief?

user: The state the world is right know it went to shit and it seems like it not getting any better.

assistant: Certainly, world events can affect our perceptions. Can you think of any specific examples or evidence that led you to believe that the FDA is withholding natural cures due to pharmaceutical pressure?

user: No one can fell safe outside in city's or anywhere in some places and the New York news. New York Sims like GTA while some parts of the world fell like New York. So far the safest places are suburban or country.

assistant: You've mentioned the state of the world and safety concerns. Do you have any reservations or doubts about this specific belief regarding the FDA and pharmaceutical companies? Would you consider information from other sources, if it pertained to this topic?

user: No it's more of government than FDA

assistant: I see, you now mention it's more about the government than the FDA. If there was evidence which countered this belief, such as government documents or independent research indicating no deliberate withholding of natural cures, how would you view such evidence?

user: Look in to it more

assistant: That seems like an open approach. Given that, can you think of any specific evidence or conditions which might potentially challenge your belief?

user: Nothing I can think of as for right know

assistant: Understood. Has this belief led you to make any significant choices in your life? For example, decisions about health care or engagement in political activity?

user: I use to trust health care a lot till one of the doctors gave me medicine that If I didn't have a second opinion before taking I would have died.

assistant: I see, you've mentioned a personal experience with healthcare that has shaped your perspective. To summarize our discussion, we explored your belief that the FDA, influenced by pharmaceutical companies, is withholding natural remedies for diseases. This belief was motivated by concerns for safety, general mistrust in the current state of the world, and personal experiences with healthcare. We also discussed your willingness to investigate contradictory evidence. Are you satisfied with this summary, or would you like something rephrased?

user: Satisfied

assistant: That's great. To complete your survey, please go back to it and enter the code word "elephant."

## Appendix E: Regression results from Figure 1

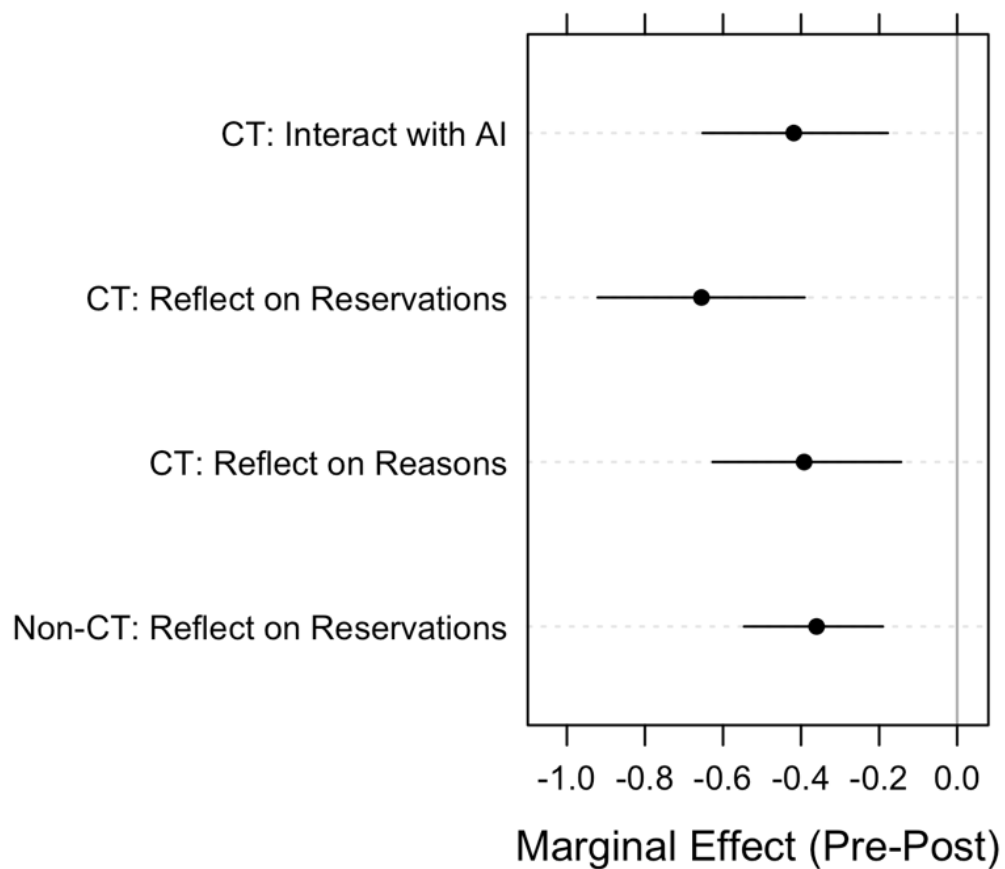


Figure E1. Marginal effect of each intervention as measured by change (decrease) in beliefs (95% confidence interval) as measured by pre-post difference of means.



**Table E1.** OLS regression of pre-post belief assessment on conditions and controls.

Variable	Pre-Post Difference
CT: Reflect on Reasons	-0.032 (0.159)
CT: Reflect on Reservations	-0.295 (0.161)
CT: Interact with AI	-0.058 (0.160)
Age	0.004 (0.004)
Educational Attainment	-0.024 (0.037)
Female	-0.299* (0.116)
Black	-0.077 (0.164)
Hispanic	-0.083 (0.152)
Constant	-0.275 (0.258)
$R^2$	0.010
$n$	1,236

Note: OLS coefficients with standard errors in parentheses, \* $p < .05$ .

## Appendix F: Regression results from Figure 2

**Table F1.** OLS regression of pre-post belief assessment on conditions, accuracy importance, and controls.

Variable	Pre-Post Difference
CT: Reflect on Reasons	0.875* (0.428)
CT: Reflect on Reservations	0.058 (0.432)
CT: Interact with AI	0.019 (0.438)
Importance	0.350*** (0.092)
CT: Reflect on Reasons X Importance	-0.300* (0.129)
CT: Reflect on Reservations X Importance	-0.116 (0.132)
CT: Interact with AI X Importance	-0.028 (0.134)
Age	0.003 (0.004)
Educational Attainment	-0.028 (0.037)
Female	-0.297* (0.116)
Black	-0.182 (0.164)
Hispanic	-0.101 (0.150)
Constant	-1.280*** (0.374)
$R^2$	0.035
$n$	1,235

Note: OLS coefficients with standard errors in parentheses, \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

## Appendix G: Regression results from Figure 3

**Table G1.** OLS regression of pre-post belief assessment on conditions, conspiracy thinking (ACTS), and controls.

Variable	Pre-Post Difference
CT: Reflect on Reasons	-0.074 (0.398)
CT: Reflect on Reservations	-0.926* (0.392)
CT: Interact with AI	-0.063 (0.394)
ACTS	0.231 (0.129)
CT: Reflect on Reasons X ACTS	-0.002 (0.194)
CT: Reflect on Reservations X ACTS	0.304 (0.191)
CT: Interact with AI X ACTS	-0.021 (0.192)
Age	0.005 (0.004)
Educational Attainment	-0.012 (0.037)
Female	-0.288* (0.116)
Black	-0.109 (0.163)
Hispanic	-0.082 (0.151)
Constant	-0.766* (0.353)
$R^2$	0.028
$n$	1235

Note: OLS coefficients with standard errors in parentheses, \* $p < .05$ .

## Appendix H: Covariate balance across conditions

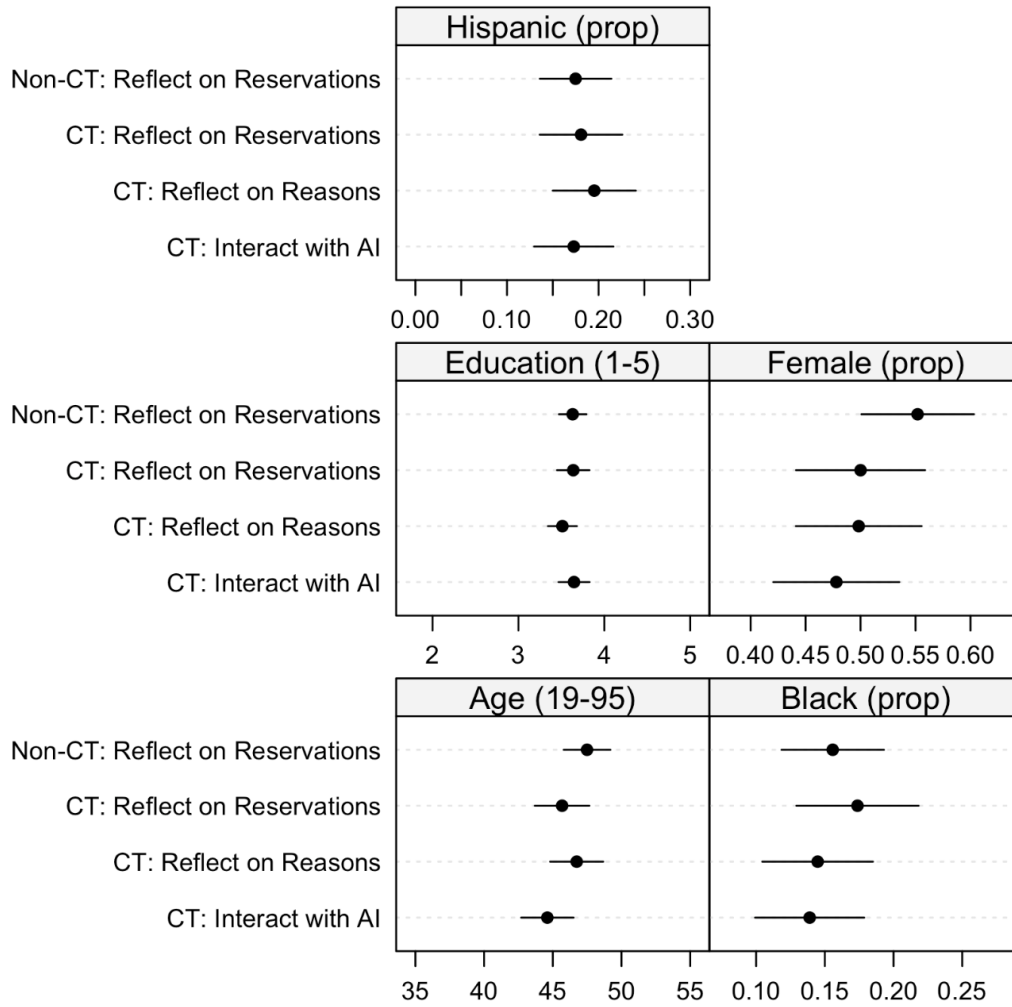
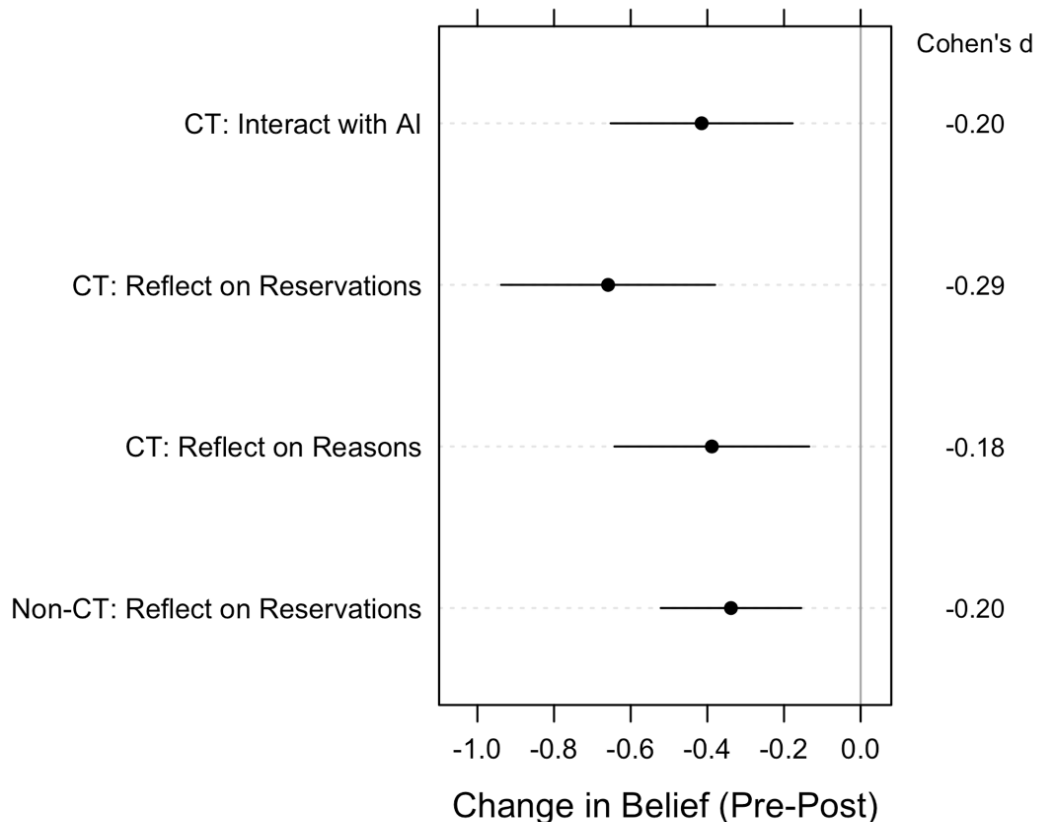


Figure H1. Average levels of controls (sociodemographic characteristics) across conditions.

## Appendix I: Replication of analyses retaining subjects who failed attention checks

Below, we reproduced each of the central results from the main text using a relaxed exclusion criterion when it comes to attention checks. Whereas in the main text, we removed subjects who failed any of the 6 attention checks embedded in the survey, in the analyses presented below we retained subjects who successfully completed at least 3 of the 6 attention checks. Results are substantively identical across the board.



*Figure I1. Average effect of each intervention as measured by change (decrease) in beliefs (95% confidence interval) as measured by pre-post difference of means.*

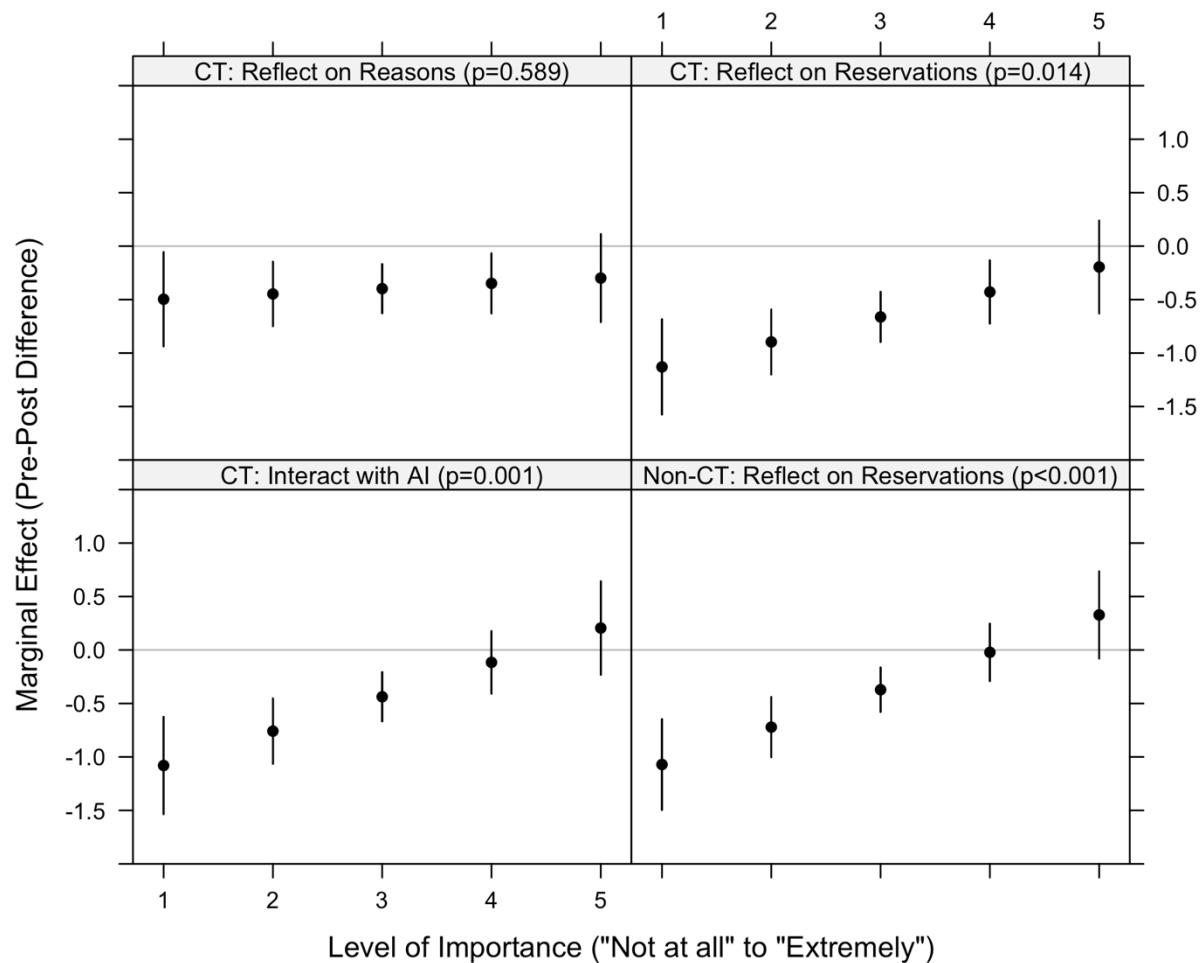
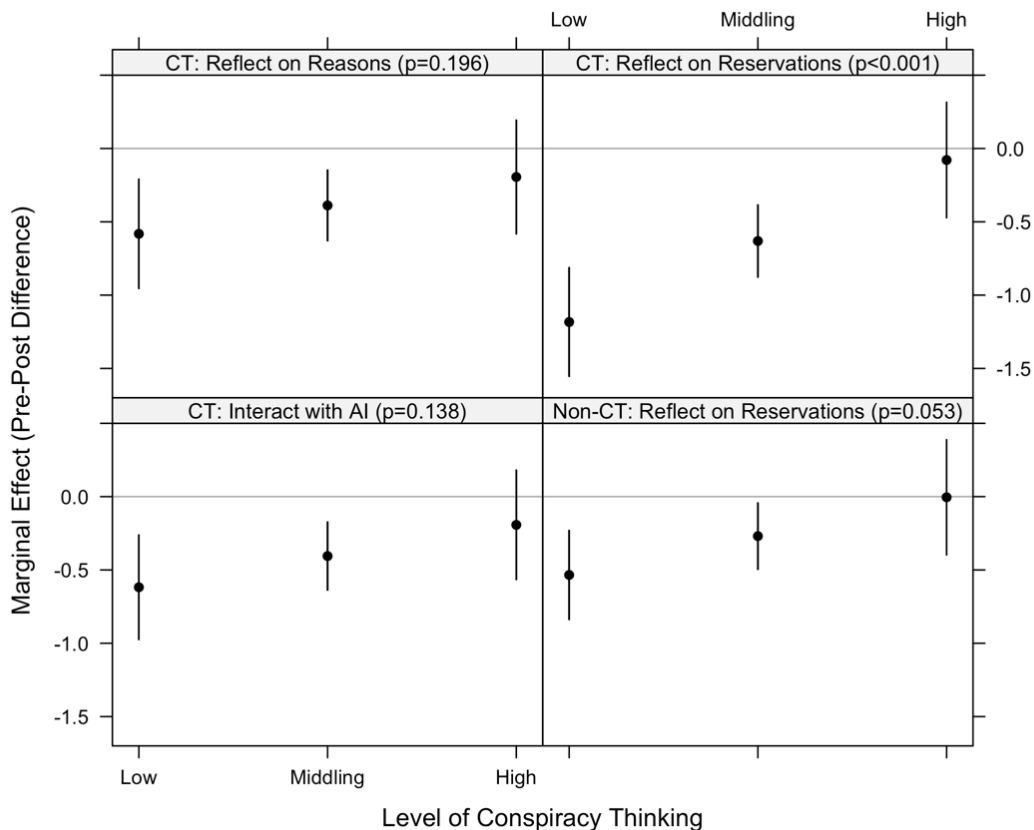


Figure 12. Marginal effect of each intervention as measured by the change (decrease) in beliefs, by level of stated importance of belief accuracy (with 95% confidence interval), with p-value (two-tailed) for significance of marginal effect.



**Figure 13. Marginal effect of each intervention as measured by the change (decrease) in beliefs, by level of conspiracy thinking (with 95% confidence interval), with p-value (two-tailed) for significance of marginal effect.**