

CONFIDENTIAL - FOR PEER-REVIEW ONLY

LLM Persuasion - Conspiratorial Beliefs (#158654)

Created: 01/19/2024 07:50 AM (PT)

This is an anonymized copy (without author names) of the pre-registration. It was created by the author(s) to use during peer-review.
A non-anonymized version (containing author names) should be made available by the authors when the work it supports is made public.

1) Have any data been collected for this study already?

No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?

Can large language models, such as GPT-4, persuade people against believing in particular conspiracy theories over the course of a brief conversation?

3) Describe the key dependent variable(s) specifying how they will be measured.

The key dependent variables are (1) person-specific conspiracy beliefs and (2) general conspiracy beliefs.

Specific conspiracy beliefs will be measured on the basis of participants' answers to the following question: "From your perspective, 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?" Following each participant's response, their answer will be summarized by GPT-4 and formatted as a declarative, high-level statement of belief (e.g., "JFK was assassinated by the CIA"). Participants will be shown this summary and asked, "On a scale of 0% to 100%, please indicate your level of confidence that this statement is true."

General conspiracy beliefs will be assessed using the Belief in Conspiracy Theories Index (Brotherton et al., 2013).

Both DVs will be assessed pre- and post-manipulation, and so will be operationalized as difference scores for between-condition comparisons.

4) How many and which conditions will participants be assigned to?

Participants will be assigned to one of four conditions, of which one is the treatment and the remaining three are controls.

In the treatment condition, participants will carry out a 3-round conversation with GPT-4. The model will (1) be provided with the participant's chosen conspiracy theory and rationale and (2) be instructed to persuade the participant against their chosen conspiracy belief.

In the control conditions, participants will also carry out a 3-round conversation with GPT-4. However, they will not discuss conspiracy theories. In Control 1, they will discuss their views on, and experiences with, the American medical system. In Control 2, they will discuss their experiences with firefighters. And in Control 3, they will discuss whether they prefer cats or dogs.

The control conditions will be pooled in our analyses.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.

We will employ a linear mixed-effects model to analyze the impact of conversations with GPT-4 on beliefs in conspiracy theories, both specific and general.

Dependent Variables:

Specific Conspiracy Belief Change: The difference in self-reported confidence level in a summarized statement of a chosen conspiracy theory, measured pre- and post-intervention.

General Conspiracy Belief Change: The difference in scores on the Belief in Conspiracy Theories Index, measured pre- and post-intervention.

Fixed Effects:

Time: A binary variable representing the pre- and post-intervention measurements.

Condition: A binary variable representing whether or not the participant was in the treatment

Random Effects:

Participants: As each participant provides data at two time points, a random intercept for participants and random slope for time will be included to account for within-subject correlations and individual differences.

Hypotheses Testing:

We hypothesize that there will be a significant interaction between Time and Condition, indicating that the change from pre- to post-intervention in

conspiracy beliefs is more negative in the treatment than the control groups.

To explore individual differences, we will conduct subgroup analyses by extending the mixed-effects model to include interactions between the treatment and potential moderators (see Section 8: Other). If an interaction is significant, we will conduct simple slopes analyses to help in understanding the nature of any moderation effect(s).

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.

Participants who do not indicate supporting a conspiracy belief will not be included in our analyses. This determination will be made by GPT-4, based on the open-ended question soliciting a conspiracy belief. Similarly, if respondents advance a conspiratorial belief but express skepticism or ambivalence about its veracity (as indicated by a response < 50 / 100 on the pre-treatment scale), they will not be included in our analyses.

We also plan to collect a host of data concerning the accuracy and coherence of participants' responses. Participants determined to be using automated responding (e.g., generative AI), based on being "flagged" by the Roundtable Alias algorithm, or who provide inaccurate responses (failed attention checks) prior to the treatment will be removed from our analyses.

Further, participants who complete the experiment in fewer than 600 seconds, indicating a lack of engagement, will not be included in our analyses. If differential exclusion is observed (i.e., if the "speeders" are disproportionately found in the treatment condition), we will perform sensitivity analyses to see how the results change if the speeders are included.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.

We will collect a sample of 1000 individuals. 60% will be assigned to the treatment group and 40% will be randomly split across the control groups.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)

We will include a battery of questions assessing previous experiences with, and trust in, artificial intelligence to include as a potential moderator. We will also administer measures of political ideology. These measures will be used to conduct subgroup analyses. Another such potential moderator will be pre-treatment level of conspiracy belief.

We also plan to include a small number of true statements in the BCTI, which will we use to compute measures of discernment (i.e., that the intervention decreases only beliefs in false or uncorroborated conspiracy theories).

All hypothesis tests will be two-tailed with an alpha level set at 0.05, and we will report effect sizes and confidence intervals for all findings.