GPT-3 Motivated Reasoning (Study 5) (#147883)

Author(s)
Yamil Velez (Columbia University) - yrv2004@columbia.edu
Patrick Liu (Columbia University) - ppl2115@columbia.edu

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?
Do people exposed to counter-attitudinal arguments polarize or moderate their attitudes? Is polarization's occurrence contingent upon how strongly worded the counter-arguments are? We use GPT-3.5 to generate tailored counter-arguments for each respondent to assess if strong attitudes enhance attitude polarization in the face of counter-attitudinal content.

3) Describe the key dependent variable(s) specifying how they will be measured.
As in Study 4, we plan to measure attitude strength using two new measures we developed: an attitudinal defense scale capturing one's willingness to defend an attitude across different social contexts and an attitude extremity scale that assesses the level of commitment individuals are willing to invest in supporting their attitude.

4) How many and which conditions will participants be assigned to?
Three conditions: placebo condition (random pseudo-news article drawn from placebo corpus), weakly worded counter-attitudinal condition, and strongly worded counterattitudinal condition. The third condition differs from the second in that the counterarguments will directly and harshly attack the respondent's issue position (as in Study 4).

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
Main analysis: We plan to regress all outcomes on pre-treatment measures of age, political sophistication, gender, education, ideology, political party, attitude certainty, 7-point attitude strength, self-reported political behavior, and race (white =1) with HC2 robust standard errors (using OLS regression). If ATE > 0 (ATE < 0), this is evidence of attitude polarization (moderation).

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We plan to analyze the subset of respondents who provided a valid prompt and obtained tailored responses. Qualtrics will flag if the respondent is shown generic responses (if GPT-3 could not produce output in response to their prompt), returning generic_flag = 1. We will analyze cases where generic_flag does not equal one.

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.
2000

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We will use the 7-point Likert scale and 0-100 certainty scale to verify that we have recovered strong attitudes and as pre-treatment covariates, rather than analyzing them as outcome variables. We specifically developed the two scales above – attitude defense and attitude extremity – to address potential ceiling effects affecting the more traditional measures.

We will also perform a conditional test and perform the main analysis while limiting our sample to people who score a 7 on the 7-point attitude strength measure. This removes potential cases in which the AI model made an error while summarizing the subject's strongly held attitude.

After answering outcome questions, respondents will be shown both the weakly worded and strongly worded counterarguments and will be asked to provide open-ended thought listings in response to each. To explore the mechanisms by which harsh wording may contribute to attitude polarization, we will conduct exploratory analyses of thought listings, comparing respondents' thought listing text on various measures: level of valence/affect (using VADER sentiment analysis), amount of defensive language (coded by GPT), effort (time spent and length), and share of words in thought listing denoting a feeling of being attacked by the counterarguments ("insult", "attack", "hurt", "offend", "disrespect").

Available at https://aspredicted.org/34pa5.pdf