Appendix for: Can AI Help Reduce Prejudice? Evaluating the Effectiveness of AI-Powered Personalized Persuasion on Support for Transgender Rights

Charles Crabtree*[1], John Holbein[2], Mitchell Bosley[3], and Semra Sevi[3]

[1]Dartmouth College
[2]University of Virginia
[3]University of Toronto

June 30, 2026

Appendix A: Survey Instrument Design

This appendix provides a detailed overview of the survey instrument used in our study, including the specific sequence of questions, scales, and experimental manipulations.

A.1 Survey Flow and Block Structure

The survey consisted of the following blocks presented in sequence:

1. **Consent Block:** Informed consent agreement describing the study purpose, procedures, risks, benefits, confidentiality measures, and participation rights.
2. **Demographics Block:** Collection of basic demographic information including age, sex assigned at birth, gender identity, education level, household income, race/ethnicity, and geographic location.
3. **Political Demographics Block:** Questions about political engagement, voting behavior, party identification, ideological placement, and ratings of political parties.
4. **Moral Foundations Questionnaire (MFQ30):** The standard 30-item MFQ measuring the importance of five moral foundations: Care/Harm, Fairness/Reciprocity, Loyalty/Betrayal, Authority/Respect, and Purity/Sanctity.
5. **Pre-Treatment Transgender Attitudes Questions Block:** Baseline measures of attitudes toward transgender people, including:
    - A feeling thermometer (0-100) measuring warmth toward transgender people
    - Agreement with statements regarding transgender rights and recognition
    - Self-rated knowledge about transgender issues

*Author order was determined by random draw.

6. **Experimental Manipulation:** Random assignment to either:
    - **Treatment Group:** Interaction with an AI chatbot personalized using MFQ responses
    - **Control Group:** No chatbot interaction
7. **Post-Treatment Questions Block:** Immediate post-treatment measurement of outcomes, including:
    - Feeling thermometer toward transgender people and other social groups
    - Agreement with policy statements regarding transgender rights
    - Behavioral intention measures regarding taking action to support transgender rights
8. **AI Perception Questions:** For treatment group only, questions about perceptions of the conversation and who participants believed they were interacting with.

A.2 Key Measures

A.2.1 Moral Foundations Questionnaire

The MFQ30 consisted of two parts:

1. **Moral Relevance:** 16 items asking "When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking?" rated on a 6-point scale from "Not at all relevant" to "Extremely relevant."
2. **Moral Judgments:** 16 items asking respondents to indicate agreement or disagreement with moral statements on a 6-point scale from "Strongly disagree" to "Strongly agree."

These items were scored according to standard MFQ procedures to create composite scores for each moral foundation.

A.2.2 Transgender Attitudes Measures

Pre- and post-treatment attitudes toward transgender people were measured using:

1. **Feeling Thermometer:** "On a scale from 0 to 100, where 0 is very cold or unfavorable and 100 is very warm or favorable, how do you feel towards transgender people?"
2. **Rights Agreement Scale:** Agreement with statements such as "Transgender people should have the same rights as cisgender people," "Gender identity should be recognized separately from biological sex," and "Transgender individuals should be able to use bathrooms matching their gender identity." Responses were recorded on a 6-point scale from "Strongly disagree" to "Strongly agree."
3. **Policy Support Items:** Agreement with statements such as "Transgender people deserve the same rights and protections as other Americans" and "Congress should pass laws to protect transgender people from job discrimination." Responses were recorded on a 5-point scale from "Strongly disagree" to "Strongly agree."

A.2.3 Behavioral Intentions

Behavioral intentions were measured using a 7-item scale asking about likelihood of engaging in various activities to support transgender rights, including:

1. Attending a rally or protest supporting transgender rights
2. Signing a petition advocating for transgender rights
3. Donating to organizations that support transgender rights
4. Sharing information or advocacy messages about transgender rights on social media
5. Volunteering with groups or events that promote transgender rights
6. Discussing transgender rights issues with friends or family to raise awareness
7. Contacting elected representatives to advocate for policies supporting transgender rights

Each item was rated on a 5-point scale from "Very Unlikely" to "Very Likely."

A.3 Randomization and Attrition

Participants were randomly assigned to treatment or control conditions using Qualtrics' block randomizer with equal probability. Attrition was monitored and recorded at each stage of the survey:

- Initial dropout before completing the pre-treatment measures
- Differential attrition during the treatment phase (higher in the treatment group, likely due to the extended nature of the chatbot interaction)
- Attrition between Wave 1 and Wave 2

Our analysis employs various standard best practice weighting strategies to address potential bias from differential attrition, as detailed in Appendix B.

A.4 AI Perception Measures

To assess how participants perceived the chatbot interaction, we included the following questions for the treatment group:

1. "In the exchange you just had, who do you think was your conversation partner?" with response options: "A human," "A bot or AI assistant (like ChatGPT)," "Unsure / I don't know."
2. "How do you feel about the conversation you had?" (open-ended response)

These measures helped us understand if participants perceived the interaction as authentic and how their perceptions might have influenced treatment effects.

A.5 Wave 2 Follow-up

The Wave 2 survey repeated key outcome measures from Wave 1 to assess durability of effects. It was administered approximately one week after the initial survey and included:

- Feeling thermometer toward transgender people
- Policy support items
- Behavioral intention measures

The full survey instrument, including exact question wording, response options, and programming logic, is available in the study's preregistration materials and data repository.

Appendix B: Attrition, Selection, and Robustness Analysis

B.1 Differential Attrition in Wave 1

The experimental design of our study included random assignment to treatment and control conditions. However, as is common in online survey experiments, we observed differential attrition between the treatment and control groups. In this section, we examine the extent of this differential attrition, its potential causes, and its implications for our results.

Extent of Differential Attrition

Overall, we observed higher attrition in the treatment group compared to the control group. Specifically, 87.8% of participants assigned to the treatment condition completed Wave 1, compared to 98.5% of participants in the control condition, representing a difference of 10.7 percentage points.

Covariate Balance Despite Differential Attrition

Despite the differential attrition, we found that covariates remained generally well-balanced between treatment and control groups among participants who completed Wave 1. Table S3 presents standardized mean differences for key covariates, with most falling below the standard threshold of 0.1, indicating good balance. Balance for all covariates for each wave is also provided in Table S1.

Table S1: Covariate Balance Across Waves

	Wave 1			Wave 2		
Variable	**Control**	**Treatment**	**SMD**	**Control**	**Treatment**	**SMD**
Pre-Treatment Transgender Thermometer	60.004	61.500	0.048	58.545	61.625	0.098
age_35-44 years old	0.187	0.190	0.006	0.187	0.161	−0.068
age_25-34 years old	0.175	0.171	−0.009	0.140	0.138	−0.006
age_45-54 years old	0.162	0.162	−0.001	0.166	0.180	0.037
age_55-64 years old	0.121	0.118	−0.008	0.156	0.150	−0.019
age_18-24 years old	0.138	0.146	0.022	0.062	0.076	0.054
age_65+ years old	0.217	0.214	−0.009	0.289	0.295	0.015
political_attention_Some of the time	0.328	0.342	0.029	0.270	0.318	**0.107**
political_attention_Most of the time	0.544	0.529	−0.030	0.625	0.574	**-0.105**
political_attention_Hardly at all	0.046	0.042	−0.016	0.037	0.034	−0.014
political_attention_Only now and then	0.082	0.087	0.017	0.069	0.074	0.021
ideological_placement	4.470	4.387	−0.046	4.354	4.282	−0.039
sex_Male	0.501	0.505	0.008	0.544	0.528	−0.031
sex_Female	0.497	0.492	−0.010	0.455	0.470	0.030
sex_Other	0.001	0.002	0.023	0.002	0.002	0.007
gender_Man	0.501	0.503	0.003	0.544	0.527	−0.035
gender_Woman	0.489	0.487	−0.004	0.445	0.466	0.042
gender_Non-binary	0.009	0.009	−0.004	0.011	0.008	−0.037
gender_A gender not listed here (please specify)	0.001	0.002	0.026	NA	NA	NA
education_High school diploma or GED	0.231	0.232	0.003	0.225	0.195	−0.073
education_Bachelor's degree	0.249	0.255	0.014	0.262	0.277	0.034
education_Some college, but no degree	0.199	0.218	0.047	0.206	0.241	0.084
education_Graduate or professional degree...	0.161	0.138	−0.066	0.166	0.146	−0.055
education_Associates or technical degree	0.137	0.133	−0.013	0.129	0.131	0.004
education_Some high school or less	0.022	0.022	0.004	0.013	0.011	−0.013
education_Prefer not to say	0.001	0.002	0.026	NA	NA	NA
income_$50,000-$99,999	0.315	0.323	0.017	0.340	0.333	−0.014
income_Less than $25,000	0.113	0.121	0.026	0.080	0.087	0.027
income_$25,000-$49,999	0.183	0.184	0.003	0.159	0.197	0.098
income_More than $200,000	0.066	0.046	−0.086	0.065	0.045	−0.087

Continued on next page

Table S1: Covariate Balance Across Waves – Continued

	Wave 1			Wave 2		
Variable	**Control**	**Treatment**	**SMD**	**Control**	**Treatment**	**SMD**
income_$100,000-$199,999	0.323	0.326	0.006	0.356	0.337	−0.039
race_African-American	0.123	0.112	−0.035	0.096	0.112	0.053
race_White, non Hispanic	0.551	0.564	0.026	0.616	0.634	0.039
race_Hispanic	0.128	0.120	−0.022	0.097	0.095	−0.009
race_White, non Hispanic,Hispanic	0.030	0.029	−0.004	0.040	0.034	−0.031
race_Hispanic,Native American	0.005	0.005	−0.002	0.005	0.004	−0.015
race_White, non Hispanic,Hispanic,Asian/ Pac Isl	0.000	0.001	0.040	NA	NA	NA
race_African-American,Hispanic	0.011	0.013	0.014	0.016	0.013	−0.022
race_Asian/ Pacific Islander	0.078	0.077	−0.002	0.065	0.042	**-0.106**
race_Native American	0.020	0.022	0.013	0.026	0.019	−0.045
race_White, non Hispanic,African-American	0.005	0.008	0.038	0.003	0.002	−0.026
race_Other	0.015	0.018	0.023	0.005	0.017	**0.118**
race_WnH,AfAm,Hisp,API,NA	0.001	0.000	−0.053	NA	NA	NA
race_WnH,AfAm,NA	0.002	0.001	−0.034	0.003	0.002	−0.026
race_WnH,API	0.001	0.005	0.061	0.000	0.006	**0.107**
race_Hispanic,API	0.003	0.001	−0.047	0.002	0.000	−0.057
race_WnH,NA	0.007	0.008	0.011	0.008	0.011	0.035
race_AfAm,API	0.004	0.001	−0.059	0.003	0.000	−0.080
race_WnH,Hisp,NA	0.004	0.002	−0.021	0.006	0.002	−0.070
race_WnH,AfAm,Hisp	0.002	0.002	−0.012	0.000	0.002	0.062
race_API,NA	0.001	0.001	0.003	0.002	0.000	−0.057
race_Prefer not to say	0.001	0.002	0.005	NA	NA	NA
race_AfAm,NA	0.002	0.003	0.021	0.003	0.000	−0.080
race_WnH,AfAm,API	0.001	0.001	−0.018	NA	NA	NA
race_WnH,AfAm,Hisp,NA	0.001	0.000	−0.037	NA	NA	NA
race_AfAm,Hisp,NA	0.001	0.001	0.003	0.002	0.002	0.007
race_Hispanic,Other	0.000	0.002	0.057	NA	NA	NA
race_AfAm,API,Other	0.001	0.000	−0.037	0.002	0.000	−0.057
race_WnH,AfAm,Hisp,API,NA,Other	0.000	0.001	0.040	0.000	0.002	0.062
race_Other,Prefer not to say	0.001	0.000	−0.037	NA	NA	NA
race_API,Other	0.001	0.001	0.003	0.002	0.000	−0.057

Continued on next page

Table S1: Covariate Balance Across Waves – Continued

	Wave 1			Wave 2		
Variable	**Control**	**Treatment**	**SMD**	**Control**	**Treatment**	**SMD**
race_WnH,AfAm,API,NA	0.000	0.001	0.040	0.000	0.002	0.062
race_AfAm,Other	0.000	0.001	0.040	NA	NA	NA
race_AfAm,API,NA,Other	0.001	0.000	−0.037	NA	NA	NA
race_WnH,Other	0.001	0.000	−0.037	NA	NA	NA
chatgpt_use_No	0.489	0.477	−0.024	0.509	0.494	−0.029
chatgpt_use_Yes	0.473	0.474	0.002	0.466	0.451	−0.030
chatgpt_use_I do not know	0.037	0.049	0.056	0.026	0.055	**0.150**
chatgpt_accuracy_Strongly agree	0.198	0.197	−0.004	0.190	0.182	−0.021
chatgpt_accuracy_Strongly disagree	0.034	0.033	−0.001	0.038	0.032	−0.033
chatgpt_accuracy_Agree	0.376	0.405	0.059	0.360	0.407	0.096
chatgpt_accuracy_Neither agree nor disagree	0.322	0.308	−0.029	0.337	0.314	−0.047
chatgpt_accuracy_Disagree	0.070	0.057	−0.056	0.075	0.064	−0.042

Note: SMD = Standardized Mean Difference. Values are means for continuous variables and proportions for binary/categorical variables. Cells shaded pink indicate an absolute SMD ≥ 0.1, a common threshold for potential imbalance. *NA* indicates data not available or category not present in Wave 2.

Table S2: Attrition Rates by Treatment Assignment (Wave 1)

Group	Assigned	Completed	Completion Rate
Control	1,447	1,426	98.5%
Treatment	1,430	1,255	87.8%
Difference			10.7%

Table S3: Covariate Balance Between Treatment and Control Groups (Wave 1 Completers)

Covariate	Standardized Mean Difference
Pre-treatment Transgender rating (q24_1)	0.048
Age	-0.009 to 0.022
Political attention	-0.030 to 0.029
Ideological placement	-0.046
Sex	-0.010 to 0.023
Gender	-0.004 to 0.026
Education	-0.066 to 0.047
Income	-0.086 to 0.026
Prior ChatGPT use	-0.024 to 0.056
Perceived ChatGPT accuracy	-0.056 to 0.059

Note: For categorical variables with multiple levels, the range of SMDs across levels is presented.

Drivers of Differential Attrition

To understand the factors associated with differential attrition, we estimated a series of logistic regression models predicting Wave 1 completion with interactions between treatment assignment and pre-treatment covariates. This analysis helps identify which subgroups were more or less likely to drop out differentially in response to treatment.

Table S4 presents the most significant interaction terms from these models.

Table S4: Significant Predictors of Differential Attrition (Wave 1)

Interaction Term	Coefficient	p-value
Treatment × ChatGPT Accuracy (Disagree)	-0.135	< 0.001
Treatment × Sex (Male)	0.043	0.018
Treatment × Prior Transgender Rating (q24_1)	0.001	0.019
Treatment × ChatGPT Accuracy (Strongly disagree)	-0.116	0.023
Treatment × Income ($25,000-$49,999)	-0.052	0.050

The most pronounced differential attrition occurred among respondents who disagreed with the statement that ChatGPT is accurate. These respondents were significantly more likely to drop out if assigned to treatment (negative coefficient of -0.135, $p<0.001$). Similar patterns were observed for those who strongly disagreed about ChatGPT's accuracy. This suggests that participants with pre-existing skepticism toward ChatGPT were particularly likely to discontinue the survey when exposed to the treatment. (As we mention in the text, our paper reveals that in scaling any AI-based treatment intervention targeting this group of AI-skeptics is a notable challenge.)

Male respondents showed lower differential attrition, with males in the treatment group being less likely to drop out compared to females (positive coefficient of 0.043, $p=0.018$). Respondents with more positive pre-treatment ratings of transgender people (q24_1) were slightly less prone to treatment-induced attrition,

though the coefficient is small (0.001, p=0.019). (To interpret this coefficient more meaningfully, consider that a 10-point increase on the feeling thermometer (representing a 10% difference in warmth toward transgender people) would be associated with a 0.1 percentage point reduction in differential attrition.) Income also played a role, with participants in the $25,000-$49,999 bracket showing increased differential attrition.

B.2 Selection into Wave 2

Extent of Selection into Wave 2

Among participants who completed Wave 1, we observed additional attrition before Wave 2, with some differential patterns between treatment and control groups. Table S5 presents the completion rates for Wave 2 conditional on having completed Wave 1.

Group	Wave 1 Completers	Wave 2 Completers	Completion Rate
Control	1,426	627	44.0%
Treatment	1,255	528	42.1%
Difference			1.9%

Table S5: Attrition Rates from Wave 1 to Wave 2

Overall, approximately 43.1% of Wave 1 participants returned to complete Wave 2. The difference in completion rates between treatment and control groups was smaller for Wave 2 (1.9 percentage points) compared to the more substantial differential attrition observed in Wave 1 (10.7 percentage points). However, the cumulative differential attrition from initial assignment to Wave 2 completion remained substantial, with 36.9% of the treatment group completing both waves compared to 43.3% of the control group.

Covariate Balance in Wave 2

We assessed covariate balance between treatment and control groups among participants who completed both waves. Table S6 presents standardized mean differences for key covariates among Wave 2 completers.

Table S6: Covariate Balance Between Treatment and Control Groups (Wave 2 Completers)

Covariate	Standardized Mean Difference
Pre-treatment Transgender rating (q24_1)	0.098
Age	-0.068 to 0.054
Political attention	-0.105 to 0.107
Ideological placement	-0.039
Sex	-0.031 to 0.030
Gender	-0.037 to 0.042
Education	-0.073 to 0.084
Income	-0.087 to 0.098
Prior ChatGPT use	-0.030 to 0.150
Perceived ChatGPT accuracy	-0.047 to 0.096

Note: For categorical variables with multiple levels, the range of SMDs across levels is presented.

While most covariates maintained acceptable balance in Wave 2, some showed standardized mean differences approaching or slightly exceeding the conventional 0.1 threshold. In particular, there were modest imbalances in political attention, pre-treatment transgender ratings, education, and ChatGPT use. These imbalances, while not severe, highlight the importance of controlling for these variables in our analyses to account for potential selection bias.

Drivers of Selection into Wave 2

To identify factors associated with differential attrition between Waves 1 and 2, we estimated regression models with interactions between treatment and covariates measured at Wave 1. Overall, we found fewer strong predictors of differential attrition for the transition from Wave 1 to Wave 2 compared to what we observed for initial completion of Wave 1.

The most notable pattern was that participants who used ChatGPT (particularly those who responded "I do not know" to prior ChatGPT use) showed a higher likelihood of completing Wave 2 if assigned to treatment. This suggests that the treatment experience may have been more engaging for participants less familiar with ChatGPT, perhaps because it provided them with a novel experience that increased their interest in the follow-up survey.

Age also played a role, with older participants (especially those 65+ years old) being somewhat more likely to complete Wave 2 compared to younger participants regardless of treatment assignment. Additionally, participants with higher pre-treatment interest in AI systems and those with more positive views about ChatGPT's accuracy were generally more likely to complete Wave 2.

B.3 Robustness Analyses

Given the evidence of differential attrition, we employ multiple approaches to assess the robustness of our findings. We triangulate between various estimation strategies to determine the sensitivity of our results to different assumptions about the attrition process. Table S7 shows treatment effect estimates across various models for each primary outcome in both Wave 1 and Wave 2.

Multiple Weighting and Adjustment Approaches

Given the complex attrition patterns in our study, we implemented a triangulation strategy with multiple standard best-practice estimation approaches to assess the robustness of our findings. This multi-model approach allows us to evaluate whether treatment effect estimates are sensitive to different statistical assumptions and selection bias adjustments.

Inverse Probability Weighting (IPW) Methodology To address potential bias from differential attrition, we employed inverse probability weighting (IPW), a technique that reweights observed data to recreate a pseudo-population where selection is independent of measured confounders (Rosenbaum and Rubin, 1983; Hirano et al., 2003).

In the context of our experiment with differential attrition, IPW works by giving greater weight to underrepresented groups in the observed sample. Mathematically, for each subject i with covariates X_i, we first estimate the probability of completing the survey (selection) conditional on treatment assignment and pre-treatment covariates:

$$e_i = P(S_i = 1|T_i, X_i) \tag{1}$$

where S_i is an indicator for survey completion, T_i is the treatment assignment, and X_i is a vector of pre-treatment covariates. We estimated these probabilities using logistic regression models with a rich set of covariates including demographics, political variables, and pre-treatment attitudes.

For ATE estimation, weights are calculated as:

$$w_i^{ATE} = \frac{S_i}{e_i} \tag{2}$$

This weighting creates a pseudo-population where treatment and control groups are balanced on observed covariates. For the average treatment effect on the treated (ATT), we modified the weights to focus on the treatment group's characteristics:

$$w_i^{ATT} = S_i \left(T_i + (1 - T_i) \frac{e_i}{1 - e_i} \frac{P(T_i = 1)}{P(T_i = 0)} \right) \tag{3}$$

For Wave 2 analysis, we implemented a sequential weighting approach to account for the two-stage selection process (selection into Wave 1 followed by selection into Wave 2). This involved:

$$w_i^{seq} = \frac{S_{i1}}{e_{i1}} \times \frac{S_{i2}}{e_{i2}} \tag{4}$$

where S_{i1} and S_{i2} are indicators for completing Waves 1 and 2, and e_{i1} and e_{i2} are the respective selection probabilities.

Model Specifications We implemented a wide range of model specifications to systematically explore how sensitive our results were to different adjustments and assumptions:

1. **Unadjusted models** provide a baseline reference point:
    - **Unweighted without adjustment**: Simple OLS regression without any corrections for attrition or covariate imbalance.
2. **Covariate adjustment models** control for pre-treatment variables:
    - **Lin adjustment** (Lin, 2013): Inclusion of pre-treatment covariates (particularly the pre-treatment transgender feeling thermometer) to improve precision without imposing functional form assumptions.
    - **Full covariate adjustment**: Including the complete set of pre-treatment covariates.
3. **Weighting approaches** address differential attrition by reweighting:
    - **ATE weighting**: Estimates effects for the entire initial sample using inverse probability weights.
    - **ATT weighting**: Focuses on estimating effects for participants in the treatment group.
    - **Attrition-specific weights**: Specially constructed to address the specific differential attrition patterns observed in our study.
4. **Combined approaches** provide additional robustness:
    - **Weighted models with Lin adjustment**: Combining weighting with minimal covariate adjustment.
    - **Doubly robust estimation**: Combining weighting with full covariate adjustment for protection against misspecification (Bang and Robins, 2005).
5. **Wave 2-specific adjustments** address the additional selection process:
    - **Wave 2-only weights**: Accounting only for selection between Waves 1 and 2.
    - **Sequential weights**: Multi-stage weights accounting for both selection into Wave 1 and subsequent selection into Wave 2.

By systematically comparing results across these models, we can identify robust treatment effects that persist across specifications as well as more sensitive findings that vary depending on modeling choices. Consistency across different adjustment methods provides evidence that results are not artifacts of any particular statistical approach.

For all weighting approaches, we implemented specific procedures to ensure robust and efficient estimation:

1. **ATE and ATT weights**: We estimated both Average Treatment Effect (ATE) and Average Treatment Effect on the Treated (ATT) weights using the covariate balancing propensity score (CBPS) method (Imai and Ratkovic, 2014), which directly optimizes the covariate balance between treatment and control groups while estimating the propensity scores.

2. **Attrition-specific weights**: For addressing differential attrition in Wave 1, we constructed specialized weights by estimating propensity models that included treatment-by-covariate interactions. This approach allowed us to account for heterogeneous selection processes where certain subgroups were more prone to differential attrition than others.

3. **Weight trimming**: To reduce the influence of extreme weights and improve statistical efficiency, we trimmed all estimated weights to the 1st and 99th percentiles of their respective distributions (Seaman and White, 2013). Specifically, we replaced any weight below the 1st percentile with the 1st percentile value, and any weight above the 99th percentile with the 99th percentile value. This approach strikes a balance between bias reduction and variance control.

These procedures were applied consistently across all weighting scenarios to ensure robust estimation without allowing extreme weights to dominate the analysis.

Doubly Robust Estimation We implemented doubly robust estimation to gain additional protection against model misspecification. The doubly robust approach combines inverse probability weighting with outcome regression modeling, providing consistent estimates if *either* the propensity score model *or* the outcome regression model is correctly specified (Bang and Robins, 2005).

For our implementation, we used a two-step procedure:

1. We first estimated the IPW propensity scores using logistic regression with pre-treatment covariates.

2. We then fitted weighted outcome models with the same covariates, effectively combining both approaches.

This approach offers "two chances" to get the model right: if the propensity score model correctly accounts for selection, the estimates will be consistent even if the outcome model is misspecified; conversely, if the outcome model is correct, the estimates will be consistent even if the weights are imperfect.

Sequential Weighting for Two-Wave Analysis For the Wave 2 analysis, we needed to account for attrition at two different stages: between assignment and Wave 1, and between Waves 1 and 2. We implemented a sequential weighting approach (Hong, 2010) that factorizes the selection process into two components and calculates weights for each stage:

1. First-stage weights (w_{i1}): Account for selection into Wave 1, based on treatment and pre-treatment covariates

2. Second-stage weights ($w_{i2|1}$): Account for selection into Wave 2 conditional on having completed Wave 1

The final weights were calculated as the product of these two component weights: $w_i^{seq} = w_{i1} \times w_{i2|1}$. This sequential approach allowed us to model the distinct selection processes at each stage while maintaining a coherent weighting strategy for the final analysis.

Block Bootstrap for Standard Errors Inverse probability weighting introduces additional uncertainty because the weights themselves are estimated rather than known. Conventional standard errors from weighted regression do not account for this additional source of variability, potentially leading to anti-conservative inference.

Following recommendations from the statistical literature, we implemented a block bootstrap procedure to obtain valid standard errors for our weighted models (Cameron et al., 2008). Our implementation follows this approach:

1. Draw bootstrap samples with replacement from the original data
2. Importantly, we used the fixed, original weights for each observation rather than recalculating them in each bootstrap sample (the "block bootstrap" approach)
3. Re-estimate the treatment effect in each bootstrap sample using the original weights
4. Calculate standard errors as the standard deviation of the bootstrap distribution

This approach accounts for the complex data structure while avoiding the computational instability that can arise from re-estimating propensity scores in each bootstrap replicate. For hypothesis testing and confidence intervals, we used the percentile method based on the bootstrap distribution.

Multiple Comparison Adjustment: Benjamini-Hochberg Procedure

Given the large number of statistical models estimated for each outcome variable in our robustness checks (as shown in Table S7), there is an increased risk of finding statistically significant results purely by chance (Type I errors). To address this potential issue of multiple comparisons, we applied the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995). The BH method controls the False Discovery Rate (FDR), which is the expected proportion of rejected null hypotheses that are actually false positives, offering a less conservative approach than methods controlling the Family-Wise Error Rate (FWER) like Bonferroni correction, while still providing strong control over false positives in exploratory settings with many tests.

Importantly, we applied this adjustment *within* each primary outcome measure (Feeling Thermometer, Behavioral Intentions, Attitudes, Latent Support) across the various model specifications presented for that specific outcome in Table S7. For instance, the BH correction for the Feeling Thermometer outcome considers all p-values from the different models run for that thermometer variable in a given wave, but not p-values from models run for Behavioral Intentions in that same wave. This within-outcome approach acknowledges that the models for a specific DV are related tests of the same underlying hypothesis under different assumptions, warranting adjustment, while treating the different DVs as conceptually distinct families of tests. The resulting BH-adjusted p-values, reported alongside the unadjusted p-values in the table, provide a more rigorous assessment of statistical significance in the context of our multi-model robustness analysis. We indicate significance levels based on these BH-adjusted p-values in Table S7.

Examination of Main Effects Across Specifications

Table S7 provides a detailed overview of the treatment effect estimates under various model specifications.

Wave 1 Outcomes: As shown in the top section of Table S7, the treatment effects in Wave 1 demonstrate considerable robustness across multiple metrics.

- **Feeling Thermometer (W1):** The unadjusted estimate shows a large effect (5.19 points, unadjusted $p<0.001$). Applying Lin adjustment reduces the estimate (4.09 points, unadjusted $p<0.001$) but it remains highly significant. Across all weighting schemes and doubly robust specifications, the

estimated effect remains positive, statistically significant (all unadjusted and BH-corrected $p<0.001$), and clusters around 3.8 to 4.1 points (excluding the unweighted and Attrition ATE weighted without Lin adjustment). The E-values for the point estimates in adjusted/weighted models are consistently around 1.5 (e.g., 1.50 for Lin adjusted), indicating moderate robustness; an unmeasured confounder associated with both treatment and outcome by risk ratios of 1.5 would be needed to explain away the effect. The E-values for the confidence interval lower bounds are also robust, around 1.35-1.36, suggesting fair resistance against confounding that could render the result non-significant.

- **Behavioral Intentions (W1):** Similar robustness is observed. The unadjusted effect (0.277, unadjusted $p<0.001$) is reduced by adjustment and weighting (to approx. 0.23-0.25), but remains highly statistically significant (unadjusted and BH-corrected $p<0.001$) across all specifications. The E-values are strong, with point estimate E-values around 1.6-1.7 and CI E-values around 1.45-1.49 for adjusted/weighted models, indicating substantial robustness to unmeasured confounding.

- **Attitudes (W1):** The unadjusted effect is 0.127 (unadjusted $p=0.003$). Adjustment and weighting reduce the estimate to approximately 0.09. While smaller, the effect remains statistically significant across all adjusted and weighted specifications after BH correction (all BH-corrected $p<0.01$ or $p<0.05$). The E-values for adjusted point estimates are around 1.36-1.38, suggesting some robustness, but the CI E-values are lower (around 1.15-1.18), indicating more vulnerability to confounding pushing the result to non-significance compared to other Wave 1 outcomes.

- **Latent Support (W1):** The unadjusted effect (0.285, unadjusted $p<0.001$) is attenuated by adjustment and weighting (to approx. 0.22-0.25) but remains positive and highly significant (unadjusted and BH-corrected $p<0.001$) across all models. E-values are robust, with point estimate E-values around 1.55-1.6 and CI E-values around 1.41-1.43 in adjusted/weighted specifications.

Overall, the Wave 1 results in Table S7 consistently indicate positive and statistically significant treatment effects across all four primary outcomes, supported by significant BH-corrected p-values and generally moderate to strong E-values, especially for Behavioral Intentions and Latent Support.

Wave 2 Outcomes: The bottom section of Table S7 reveals that Wave 2 effects are considerably less robust, as indicated by non-significant BH-corrected p-values and low E-values for most outcomes and models.

- **Feeling Thermometer (W2):** The unadjusted estimate is positive and nominally significant (3.78, unadjusted $p=0.040$), but this significance does not survive BH correction (BH $p=0.416$). Applying Lin adjustment (1.62, $p=0.214$) or any form of weighting renders the effect statistically non-significant by both unadjusted and BH-corrected standards (all BH $p>0.4$). Correspondingly, the E-values are low; for the Lin-adjusted model, the point estimate E-value is only 1.27, and critically, the E-value for the confidence interval is 1.00, indicating that even very weak unmeasured confounding could explain away the point estimate and shift the confidence interval to include zero.

- **Behavioral Intentions (W2):** The unadjusted estimate is marginal (0.144, unadjusted $p=0.062$) and non-significant after BH correction (BH $p=0.603$). After Lin adjustment (0.063, $p=0.292$) or applying various weighting schemes (estimates around 0.04-0.07), the effect becomes clearly non-significant by any standard (all BH $p>0.6$). E-values are low, with point estimate E-values around 1.2-1.28 and CI E-values uniformly at 1.00 for adjusted/weighted models, indicating high sensitivity to unmeasured confounding.

- **Attitudes (W2):** This outcome shows the most persistence, although still weakened. The unadjusted effect is significant (0.186, unadjusted $p=0.004$, BH $p=0.072$). With Lin adjustment, the effect remains nominally significant (0.118, unadjusted $p=0.021$) but is marginal after BH correction

(BH p=0.147). Some weighted specifications also retain nominal significance (e.g., Regression adjustment only: 0.100, unadjusted p=0.046, BH p=0.147). Sequential weighting models yield slightly smaller estimates (around 0.08-0.10) that are generally non-significant (BH p>0.16). Point estimate E-values for adjusted models are around 1.4, but CI E-values are lower (around 1.0-1.14), suggesting the significance is fragile.

- **Latent Support (W2):** The unadjusted estimate is significant (0.230, unadjusted p=0.010) but marginal after BH correction (BH p=0.160). Lin adjustment reduces the estimate but it remains nominally significant (0.119, unadjusted p=0.045, BH p=0.192). However, most weighted models (especially sequential ones) produce smaller estimates (around 0.09-0.12) that are generally non-significant (BH p>0.19). Point estimate E-values are around 1.3-1.36 for adjusted models, while CI E-values are near 1.0 (1.00-1.04), indicating vulnerability.

In summary, Table S7 highlights that the apparent treatment effects observed in Wave 2 are sensitive to model specification and multiple comparison adjustments. For most outcomes, weighting for attrition and applying BH corrections attenuates effects below significance thresholds, and low E-values (especially for the CI) suggest vulnerability to unmeasured confounding. Only the effect on Attitudes shows some limited evidence of robustness, though even this finding is borderline in several specifications and less robust than Wave 1 effects.

Table S7: Effects with BH-Adjusted p-values and E-values

Wave	Model	Est.	SE	p-val.[1]	p-val. (BH)[1,2]	E-val.[3]	E-val. (CI)[4]	95% CI	N
	Feeling Thermometer (W1)								
Wave 1	Unweighted (no adjustment)	**5.194**	**1.206**	1.706102e-05	**0.000*** **	1.60	1.39	**[2.831, 7.557]**	2,681
Wave 1	Unweighted with Lin adjustment	**4.091**	**0.817**	5.786268e-07	**0.000*** **	1.50	1.36	**[2.491, 5.692]**	2,681
Wave 1	ATT weighted	**3.930**	**1.172**	0.000000e+00	**0.000*** **	1.49	1.27	**[1.634, 6.275]**	2,681
Wave 1	ATT weighted with Lin adjustment	**3.864**	**0.798**	0.000000e+00	**0.000*** **	1.48	1.34	**[2.246, 5.443]**	2,681
Wave 1	ATE weighted	**4.033**	**1.181**	0.000000e+00	**0.000*** **	1.50	1.28	**[1.697, 6.413]**	2,681
Wave 1	ATE weighted with Lin adjustment	**3.981**	**0.796**	0.000000e+00	**0.000*** **	1.49	1.35	**[2.388, 5.547]**	2,681
Wave 1	Attrition ATE weighted	**4.706**	**1.192**	0.000000e+00	**0.000*** **	1.56	1.35	**[2.376, 7.086]**	2,681
Wave 1	Attrition ATE weighted with Lin adjustment	**4.084**	**0.794**	0.000000e+00	**0.000*** **	1.50	1.36	**[2.498, 5.659]**	2,681
Wave 1	Double robust (ATT weighted with all covariates)	**3.805**	**0.795**	0.000000e+00	**0.000*** **	1.48	1.34	**[2.233, 5.436]**	2,679
Wave 1	Double robust (ATE weighted with all covariates)	**3.936**	**0.795**	0.000000e+00	**0.000*** **	1.49	1.35	**[2.414, 5.529]**	2,679
Wave 1	Double robust (Attrition ATE weighted with all covariates)	**3.927**	**0.793**	0.000000e+00	**0.000*** **	1.49	1.35	**[2.394, 5.516]**	2,679
Wave 1	Regression adjustment only (no weights)	**3.929**	**0.807**	1.201535e-06	**0.000*** **	1.49	1.35	**[2.347, 5.512]**	2,679
	Behavioral Intentions (W1)								
Wave 1	Unweighted (no adjustment)	**0.277**	**0.050**	3.038025e-08	**0.000*** **	1.73	1.53	**[0.179, 0.375]**	2,681
Wave 1	Unweighted with Lin adjustment	**0.238**	**0.039**	8.461987e-10	**0.000*** **	1.65	1.49	**[0.162, 0.314]**	2,681
Wave 1	ATT weighted	**0.240**	**0.050**	0.000000e+00	**0.000*** **	1.65	1.45	**[0.138, 0.33]**	2,681
Wave 1	ATT weighted with Lin adjustment	**0.238**	**0.040**	0.000000e+00	**0.000*** **	1.65	1.48	**[0.154, 0.313]**	2,681
Wave 1	ATE weighted	**0.234**	**0.050**	0.000000e+00	**0.000*** **	1.64	1.43	**[0.13, 0.326]**	2,681
Wave 1	ATE weighted with Lin adjustment	**0.232**	**0.040**	0.000000e+00	**0.000*** **	1.63	1.47	**[0.15, 0.308]**	2,681
Wave 1	Attrition ATE weighted	**0.254**	**0.049**	0.000000e+00	**0.000*** **	1.68	1.48	**[0.151, 0.346]**	2,681
Wave 1	Attrition ATE weighted with Lin adjustment	**0.232**	**0.039**	0.000000e+00	**0.000*** **	1.63	1.47	**[0.15, 0.307]**	2,681
Wave 1	Double robust (ATT weighted with all covariates)	**0.234**	**0.039**	0.000000e+00	**0.000*** **	1.64	1.48	**[0.153, 0.309]**	2,679
Wave 1	Double robust (ATE weighted with all covariates)	**0.231**	**0.039**	0.000000e+00	**0.000*** **	1.63	1.47	**[0.151, 0.306]**	2,679
Wave 1	Double robust (Attrition ATE weighted with all covariates)	**0.226**	**0.038**	0.000000e+00	**0.000*** **	1.62	1.47	**[0.148, 0.3]**	2,679
Wave 1	Regression adjustment only (no weights)	**0.226**	**0.037**	1.070375e-09	**0.000*** **	1.62	1.47	**[0.154, 0.299]**	2,679
	Attitudes (W1)								
Wave 1	Unweighted (no adjustment)	**0.127**	**0.043**	2.994494e-03	**0.007****	1.46	1.23	**[0.043, 0.21]**	2,681
Wave 1	Unweighted with Lin adjustment	**0.095**	**0.034**	5.517471e-03	**0.007****	1.38	1.18	**[0.028, 0.162]**	2,681
Wave 1	ATT weighted	**0.090**	**0.043**	3.400000e-02	**0.036***	1.37	1.08	**[0.005, 0.172]**	2,681
Wave 1	ATT weighted with Lin adjustment	**0.088**	**0.035**	6.000000e-03	**0.007****	1.36	1.15	**[0.021, 0.159]**	2,681
Wave 1	ATE weighted	**0.090**	**0.043**	3.600000e-02	**0.036***	1.36	1.08	**[0.005, 0.172]**	2,681
Wave 1	ATE weighted with Lin adjustment	**0.088**	**0.035**	6.000000e-03	**0.007****	1.36	1.15	**[0.021, 0.16]**	2,681
Wave 1	Attrition ATE weighted	**0.108**	**0.043**	4.000000e-03	**0.007****	1.41	1.16	**[0.022, 0.191]**	2,681
Wave 1	Attrition ATE weighted with Lin adjustment	**0.091**	**0.035**	4.000000e-03	**0.007****	1.37	1.16	**[0.024, 0.163]**	2,681
Wave 1	Double robust (ATT weighted with all covariates)	**0.089**	**0.033**	2.000000e-03	**0.007****	1.36	1.16	**[0.026, 0.156]**	2,679

Continued on next page

Wave	Model	Est.	SE	p-val.[1]	p-val. (BH)[1,2]	E-val.[3]	E-val. (CI)[4]	95% CI	N
Wave 1	Double robust (ATE weighted with all covariates)	**0.088**	**0.033**	0.000000e+00	**0.000*** **	1.36	1.16	**[0.025, 0.157]**	2,679
Wave 1	Double robust (Attrition ATE weighted with all covariates)	**0.089**	**0.033**	2.000000e-03	**0.007** **	1.36	1.16	**[0.023, 0.156]**	2,679
Wave 1	Regression adjustment only (no weights)	**0.092**	**0.032**	4.412885e-03	**0.007** **	1.37	1.18	**[0.029, 0.155]**	2,679
	Latent Support (W1)								
Wave 1	Unweighted (no adjustment)	**0.285**	**0.058**	7.880294e-07	**0.000*** **	1.66	1.46	**[0.172, 0.398]**	2,681
Wave 1	Unweighted with Lin adjustment	**0.231**	**0.037**	7.186469e-10	**0.000*** **	1.57	1.43	**[0.158, 0.304]**	2,681
Wave 1	ATT weighted	**0.226**	**0.056**	0.000000e+00	**0.000*** **	1.56	1.35	**[0.115, 0.335]**	2,681
Wave 1	ATT weighted with Lin adjustment	**0.223**	**0.038**	0.000000e+00	**0.000*** **	1.55	1.42	**[0.154, 0.295]**	2,681
Wave 1	ATE weighted	**0.225**	**0.057**	0.000000e+00	**0.000*** **	1.56	1.35	**[0.115, 0.332]**	2,681
Wave 1	ATE weighted with Lin adjustment	**0.223**	**0.038**	0.000000e+00	**0.000*** **	1.55	1.42	**[0.151, 0.297]**	2,681
Wave 1	Attrition ATE weighted	**0.256**	**0.057**	0.000000e+00	**0.000*** **	1.61	1.41	**[0.144, 0.365]**	2,681
Wave 1	Attrition ATE weighted with Lin adjustment	**0.226**	**0.038**	0.000000e+00	**0.000*** **	1.56	1.42	**[0.156, 0.298]**	2,681
Wave 1	Double robust (ATT weighted with all covariates)	**0.220**	**0.037**	0.000000e+00	**0.000*** **	1.55	1.42	**[0.152, 0.293]**	2,679
Wave 1	Double robust (ATE weighted with all covariates)	**0.221**	**0.037**	0.000000e+00	**0.000*** **	1.55	1.42	**[0.154, 0.295]**	2,679
Wave 1	Double robust (Attrition ATE weighted with all covariates)	**0.219**	**0.037**	0.000000e+00	**0.000*** **	1.55	1.41	**[0.149, 0.292]**	2,679
Wave 1	Regression adjustment only (no weights)	**0.221**	**0.036**	9.656578e-10	**0.000*** **	1.55	1.42	**[0.15, 0.291]**	2,679
	Feeling Thermometer (W2)								
Wave 2	Unweighted (no adjustment)	**3.776**	**1.839**	4.023575e-02	**0.416**	1.48	1.08	**[0.172, 7.38]**	1,155
Wave 2	Unweighted with Lin adjustment	1.615	1.300	2.141999e-01	0.416	1.27	1.00	[-0.932, 4.163]	1,155
Wave 2	Wave 2-only ATT weighted	1.516	1.889	4.140000e-01	0.436	1.26	1.00	[-2.18, 5.194]	1,155
Wave 2	Wave 2-only ATT weighted with Lin adjustment	1.499	1.326	2.660000e-01	0.416	1.26	1.00	[-0.958, 4.235]	1,155
Wave 2	Wave 2-only ATE weighted	1.526	1.936	4.360000e-01	0.436	1.26	1.00	[-2.449, 5.31]	1,155
Wave 2	Wave 2-only ATE weighted with Lin adjustment	1.445	1.335	2.860000e-01	0.416	1.25	1.00	[-0.95, 4.304]	1,155
Wave 2	Sequential ATT weighted	1.639	2.077	4.120000e-01	0.436	1.28	1.00	[-2.507, 5.522]	1,155
Wave 2	Sequential ATT weighted with Lin adjustment	1.749	1.460	2.260000e-01	0.416	1.29	1.00	[-0.771, 4.759]	1,155
Wave 2	Doubly robust with Sequential ATT weights	1.660	1.426	2.220000e-01	0.416	1.28	1.00	[-0.805, 4.689]	1,155
Wave 2	Sequential ATE weighted	1.648	2.084	4.060000e-01	0.436	1.28	1.00	[-2.633, 5.576]	1,155
Wave 2	Sequential ATE weighted with Lin adjustment	1.832	1.467	1.980000e-01	0.416	1.30	1.00	[-0.688, 4.821]	1,155
Wave 2	Doubly robust with Sequential ATE weights	1.707	1.433	2.140000e-01	0.416	1.28	1.00	[-0.742, 4.698]	1,155
Wave 2	Sequential Attrition ATE weighted	2.010	2.101	3.260000e-01	0.435	1.31	1.00	[-2.118, 5.982]	1,155
Wave 2	Sequential Attrition ATE weighted with Lin adjustment	1.819	1.476	1.960000e-01	0.416	1.29	1.00	[-0.777, 4.888]	1,155
Wave 2	Doubly robust with Sequential Attrition ATE weights	1.570	1.437	2.500000e-01	0.416	1.27	1.00	[-0.989, 4.501]	1,155
Wave 2	Regression adjustment only (no weights)	1.525	1.291	2.375359e-01	0.416	1.26	1.00	[-1.004, 4.055]	1,155
	Behavioral Intentions (W2)								
Wave 2	Unweighted (no adjustment)	0.144	0.077	6.156194e-02	0.603	1.45	1.00	[-0.007, 0.295]	1,155
Wave 2	Unweighted with Lin adjustment	0.063	0.060	2.924754e-01	0.603	1.26	1.00	[-0.054, 0.18]	1,155
Wave 2	Wave 2-only ATT weighted	0.054	0.080	4.900000e-01	0.603	1.24	1.00	[-0.104, 0.212]	1,155
Wave 2	Wave 2-only ATT weighted with Lin adjustment	0.054	0.064	3.720000e-01	0.603	1.24	1.00	[-0.075, 0.182]	1,155

Continued on next page

Wave	Model	Est.	SE	p-val.[1]	p-val. (BH)[1,2]	E-val.[3]	E-val. (CI)[4]	95% CI	N
Wave 2	Wave 2-only ATE weighted	0.070	0.080	3.940000e-01	0.603	1.28	1.00	[-0.096, 0.227]	1,155
Wave 2	Wave 2-only ATE weighted with Lin adjustment	0.067	0.063	2.660000e-01	0.603	1.27	1.00	[-0.059, 0.19]	1,155
Wave 2	Sequential ATT weighted	0.046	0.088	6.160000e-01	0.616	1.22	1.00	[-0.127, 0.221]	1,155
Wave 2	Sequential ATT weighted with Lin adjustment	0.050	0.068	4.560000e-01	0.603	1.23	1.00	[-0.086, 0.184]	1,155
Wave 2	Doubly robust with Sequential ATT weights	0.036	0.065	5.520000e-01	0.616	1.19	1.00	[-0.087, 0.164]	1,155
Wave 2	Sequential ATE weighted	0.045	0.088	6.040000e-01	0.616	1.21	1.00	[-0.13, 0.222]	1,155
Wave 2	Sequential ATE weighted with Lin adjustment	0.052	0.067	4.220000e-01	0.603	1.23	1.00	[-0.078, 0.184]	1,155
Wave 2	Doubly robust with Sequential ATE weights	0.043	0.064	4.900000e-01	0.603	1.21	1.00	[-0.083, 0.17]	1,155
Wave 2	Sequential Attrition ATE weighted	0.061	0.087	4.700000e-01	0.603	1.26	1.00	[-0.112, 0.23]	1,155
Wave 2	Sequential Attrition ATE weighted with Lin adjustment	0.054	0.066	3.960000e-01	0.603	1.24	1.00	[-0.074, 0.183]	1,155
Wave 2	Doubly robust with Sequential Attrition ATE weights	0.043	0.064	4.640000e-01	0.603	1.21	1.00	[-0.078, 0.169]	1,155
Wave 2	Regression adjustment only (no weights)	0.067	0.058	2.502452e-01	0.603	1.27	1.00	[-0.047, 0.18]	1,155
	Attitudes (W2)								
Wave 2	Unweighted (no adjustment)	**0.186**	**0.065**	4.490031e-03	**0.072**	1.60	1.27	**[0.058, 0.314]**	1,155
Wave 2	Unweighted with Lin adjustment	**0.118**	**0.051**	2.106129e-02	**0.147**	1.44	1.14	**[0.018, 0.219]**	1,155
Wave 2	Wave 2-only ATT weighted	0.100	0.063	1.100000e-01	0.160	1.39	1.00	[-0.02, 0.223]	1,155
Wave 2	Wave 2-only ATT weighted with Lin adjustment	**0.100**	**0.050**	4.000000e-02	**0.147**	1.39	1.05	**[0.005, 0.201]**	1,155
Wave 2	Wave 2-only ATE weighted	0.100	0.066	1.140000e-01	0.160	1.39	1.00	[-0.029, 0.228]	1,155
Wave 2	Wave 2-only ATE weighted with Lin adjustment	**0.097**	**0.051**	4.600000e-02	**0.147**	1.38	1.00	**[0.004, 0.203]**	1,155
Wave 2	Sequential ATT weighted	0.083	0.071	2.300000e-01	0.230	1.34	1.00	[-0.057, 0.224]	1,155
Wave 2	Sequential ATT weighted with Lin adjustment	0.086	0.057	1.200000e-01	0.160	1.35	1.00	[-0.029, 0.2]	1,155
Wave 2	Doubly robust with Sequential ATT weights	0.084	0.056	1.320000e-01	0.162	1.35	1.00	[-0.023, 0.203]	1,155
Wave 2	Sequential ATE weighted	0.085	0.071	2.240000e-01	0.230	1.35	1.00	[-0.049, 0.227]	1,155
Wave 2	Sequential ATE weighted with Lin adjustment	0.091	0.057	1.140000e-01	0.160	1.36	1.00	[-0.019, 0.207]	1,155
Wave 2	Doubly robust with Sequential ATE weights	0.087	0.055	1.160000e-01	0.160	1.36	1.00	[-0.019, 0.204]	1,155
Wave 2	Sequential Attrition ATE weighted	0.106	0.072	1.460000e-01	0.167	1.40	1.00	[-0.035, 0.243]	1,155
Wave 2	Sequential Attrition ATE weighted with Lin adjustment	0.100	0.057	7.600000e-02	0.160	1.39	1.00	[-0.009, 0.215]	1,155
Wave 2	Doubly robust with Sequential Attrition ATE weights	0.084	0.055	1.200000e-01	0.160	1.35	1.00	[-0.021, 0.199]	1,155
Wave 2	Regression adjustment only (no weights)	**0.100**	**0.050**	4.584395e-02	**0.147**	1.39	1.04	**[0.002, 0.199]**	1,155
	Latent Support (W2)								
Wave 2	Unweighted (no adjustment)	**0.230**	**0.089**	1.001956e-02	**0.160**	1.56	1.22	**[0.055, 0.404]**	1,155
Wave 2	Unweighted with Lin adjustment	**0.119**	**0.059**	4.502779e-02	**0.192**	1.36	1.04	**[0.003, 0.235]**	1,155
Wave 2	Wave 2-only ATT weighted	0.104	0.090	2.380000e-01	0.272	1.33	1.00	[-0.077, 0.271]	1,155
Wave 2	Wave 2-only ATT weighted with Lin adjustment	0.103	0.060	7.600000e-02	0.192	1.32	1.00	[-0.013, 0.224]	1,155
Wave 2	Wave 2-only ATE weighted	0.111	0.092	2.180000e-01	0.272	1.34	1.00	[-0.074, 0.281]	1,155
Wave 2	Wave 2-only ATE weighted with Lin adjustment	0.107	0.060	6.200000e-02	0.192	1.33	1.00	[-0.007, 0.231]	1,155
Wave 2	Sequential ATT weighted	0.094	0.099	3.420000e-01	0.342	1.30	1.00	[-0.108, 0.285]	1,155
Wave 2	Sequential ATT weighted with Lin adjustment	0.099	0.065	1.180000e-01	0.192	1.32	1.00	[-0.021, 0.231]	1,155
Wave 2	Doubly robust with Sequential ATT weights	0.091	0.063	1.360000e-01	0.198	1.30	1.00	[-0.027, 0.224]	1,155

Continued on next page

Wave	Model	Est.	SE	p-val.[1]	p-val. (BH)[1,2]	E-val.[3]	E-val. (CI)[4]	95% CI	N
Wave 2	Sequential ATE weighted	0.095	0.099	3.240000e-01	0.342	1.31	1.00	[-0.113, 0.288]	1,155
Wave 2	Sequential ATE weighted with Lin adjustment	0.104	0.065	9.800000e-02	0.192	1.33	1.00	[-0.017, 0.236]	1,155
Wave 2	Doubly robust with Sequential ATE weights	0.096	0.063	1.060000e-01	0.192	1.31	1.00	[-0.022, 0.229]	1,155
Wave 2	Sequential Attrition ATE weighted	0.119	0.100	2.240000e-01	0.272	1.36	1.00	[-0.088, 0.31]	1,155
Wave 2	Sequential Attrition ATE weighted with Lin adjustment	0.110	0.065	7.800000e-02	0.192	1.34	1.00	[-0.01, 0.241]	1,155
Wave 2	Doubly robust with Sequential Attrition ATE weights	0.092	0.063	1.200000e-01	0.192	1.30	1.00	[-0.026, 0.224]	1,155
Wave 2	Regression adjustment only (no weights)	0.109	0.058	5.767706e-02	0.192	1.34	1.00	[-0.003, 0.222]	1,155

[1] *p < 0.05, **p < 0.01, ***p < 0.001

[2] Benjamini-Hochberg correction applied within each outcome type

[4] E-value (CI) represents the E-value for the confidence interval lower bound

E-Values for Unmeasured Confounding

E-values (VanderWeele and Ding, 2017) provide a measure of how robust a causal estimate is to potential unmeasured confounding. Unlike other sensitivity analyses, E-values offer a straightforward, single-number summary that quantifies the minimum strength of association that an unmeasured confounder would need to have with both the treatment and the outcome to fully explain away an observed treatment effect.

Mathematically, E-values are derived from the following framework: Let Y be the outcome, T be the treatment, and U be an unmeasured confounder. The observed risk ratio (RR) between treatment and outcome may be biased away from the true causal risk ratio due to U. VanderWeele and Ding showed that this bias is bounded by:

$$\frac{\text{Observed RR}}{\text{True RR}} \leq \frac{RR_{TU} \times RR_{UY}}{RR_{TU} + RR_{UY} - 1} \tag{5}$$

where RR_{TU} is the maximum risk ratio relating the unmeasured confounder to treatment, and RR_{UY} is the maximum risk ratio relating the unmeasured confounder to the outcome, after adjusting for measured covariates.

If we set $RR_{TU} = RR_{UY} = \sqrt{B}$ where B is the bias factor, then the minimum value of $\sqrt{B}$ needed to fully explain away an observed risk ratio RR_{obs} (i.e., to shift the true risk ratio to 1.0) is:

$$\sqrt{B} = RR_{obs} + \sqrt{RR_{obs} \times (RR_{obs} - 1)} \tag{6}$$

This value is defined as the E-value. Similarly, the E-value for the lower bound of a confidence interval is the minimum strength of confounding needed to shift the confidence interval to include the null value (typically 1.0 for a risk ratio).

For continuous outcomes like ours, we first converted our treatment effect estimates to an approximate risk ratio scale. Following VanderWeele and Ding (2017), we used the transformation $RR \approx \exp(0.91 \times d)$, where d is the standardized mean difference (Cohen's d, calculated as the effect estimate divided by the outcome standard deviation).

The E-value offers an intuitive interpretation: larger E-values indicate greater robustness to unmeasured confounding. For instance, an E-value of 1.5 means an unmeasured confounder would need to be associated with both treatment and outcome by risk ratios of at least 1.5 each to fully explain away the observed effect. An E-value close to 1.0 suggests that even a weak unmeasured confounder could explain away the observed association.

Outcome	Point Estimate E-Value	Lower Bound E-Value
Wave 1 Outcomes		
Feeling Thermometer	1.50	1.36
Behavioral Intentions	1.65	1.49
Attitudes	1.38	1.18
Latent Support	1.57	1.43
Wave 2 Outcomes		
Feeling Thermometer	1.27	1.00
Behavioral Intentions	1.26	1.00
Attitudes	1.44	1.14
Latent Support	1.36	1.04

Table S8: E-Values for Main Treatment Effects

To calibrate the practical significance of these E-values, we calculated benchmark E-values for a key observed confounder: the pre-treatment feeling thermometer rating of transgender people (q24_1). This variable is a strong predictor of our outcomes and represents a substantively meaningful confounding mechanism. The benchmark E-value quantifies the strength of association that an unmeasured confounder would

need to have with both treatment and outcome to exert influence comparable to this known, substantively important variable.

Outcome	Treatment	Outcome	Joint RR	E-value
Feeling Thermometer (W1)	1.05	1.96	1.05	1.28
Behavioral Intentions (W1)	1.05	1.77	1.05	1.28
Attitudes (W1)	1.05	1.73	1.05	1.28
Latent Support (W1)	1.05	2.00	1.05	1.28

Table S9: Benchmark E-values for Pre-treatment Trans Thermometer (q24_1)

As shown in Table S9, pre-treatment transgender attitudes have strong associations with our outcome measures (RRs from 1.73 to 2.00) but only a modest association with treatment assignment (RR = 1.05), yielding a benchmark E-value of 1.28. This benchmark has a concrete interpretation: to completely explain away our observed treatment effects, an unmeasured confounder would need to have influences at least as strong as pre-treatment transgender attitudes, which we know to be one of the strongest predictors of these outcomes.

Comparing treatment effect E-values to this benchmark is revealing. For Wave 1 outcomes, the treatment effect E-values (1.38-1.65) substantially exceed the benchmark (1.28), suggesting that unmeasured confounders would need to exert even stronger influence than pre-treatment attitudes to explain away these effects. For Wave 2, the lower bound E-values (1.00-1.14) fall below the benchmark, indicating that unmeasured confounders with influence comparable to or weaker than pre-treatment attitudes could potentially explain these results.

For Wave 1 outcomes, the E-values suggest moderate robustness to unmeasured confounding. An unmeasured confounder would need to be associated with both treatment assignment and the outcome by risk ratios of approximately 1.4 to 1.7 to fully explain away the observed treatment effects, and risk ratios of 1.2 to 1.5 to shift the confidence interval to include zero.

Wave 2 outcomes show considerably less robustness to unmeasured confounding. With the exception of attitudes, which maintained an E-value of 1.44 for the point estimate, the E-values for the lower confidence interval bound are generally at or near 1.0, indicating that even modest unmeasured confounding could potentially explain away the estimated treatment effects.

Lee Bounds for Differential Attrition

To provide a particularly rigorous assessment of robustness against differential attrition, we employ Lee bounds (Lee, 2009). Unlike Inverse Probability Weighting (IPW), which relies on correctly modeling the selection process using observed covariates, Lee bounds offer a non-parametric approach that makes fewer assumptions but often yields wider, more conservative bounds on the treatment effect. This method serves as a stringent "worst-case scenario" analysis.

The key insight of Lee bounds is to determine the range of possible treatment effects consistent with the observed data by considering the most extreme ways differential attrition could bias the results, under a single core assumption: monotonicity. This assumption posits that the treatment affects sample selection (survey completion) in only one direction for all subjects. In our study, where the treatment group had lower completion rates, the assumption is $S(1) \leq S(0)$, assignment to treatment could only decrease, never increase, the likelihood of completion compared to control.

Under this assumption, the potential bias arises from the "excess" participants who completed the survey in the control group but hypothetically would not have completed it if assigned to treatment. Lee bounds are constructed by trimming this "excess" fraction of observations from the control group (the group with the higher completion rate). The proportion to trim is $\delta = \frac{P(S=1|T=0)-P(S=1|T=1)}{P(S=1|T=0)}$, where T indicates treatment assignment.

The bounds are then calculated by considering the two extreme possibilities for the outcomes of these trimmed control participants:

$$\text{Lower Bound} = E[Y(1)|S(1) = 1] - E[Y(0)|S(0) = 1, Y(0) \geq q_{1-\delta}] \quad (7)$$

$$\text{Upper Bound} = E[Y(1)|S(1) = 1] - E[Y(0)|S(0) = 1, Y(0) \leq q_{\delta}] \quad (8)$$

where q_p is the pth quantile of the outcome $Y(0)$ among control group completers. The lower bound represents the *most damaging* scenario for a positive treatment effect, assuming the trimmed individuals (who only completed under control) had the *highest* possible outcome values among controls. Conversely, the upper bound assumes these individuals had the *lowest* possible outcome values. This bracketing between best- and worst-case scenarios, without relying on covariate information to model selection, is what makes Lee bounds inherently conservative.

Given our two-wave design, we computed both standard Lee bounds for Wave 1 (considering attrition after initial assignment) and compound Lee bounds for Wave 2. Compound bounds account for the *cumulative* differential attrition from initial assignment through to Wave 2 completion, providing the most appropriate and stringent assessment for our longitudinal analysis. These compound bounds reflect the total selection pressure across the study period.

Outcome	Lower Bound	Point Estimate	Upper Bound
Wave 1 Outcomes			
Feeling Thermometer	-2.01	4.09	10.12
Behavioral Intentions	0.07	0.24	0.54
Attitudes	-0.14	0.09	0.28
Latent Support	-0.05	0.23	0.56
Wave 2 Outcomes (Compound Bounds)			
Feeling Thermometer	-5.84	1.62	10.56
Behavioral Intentions	-0.13	0.06	0.51
Attitudes	-0.17	0.12	0.41
Latent Support	-0.21	0.12	0.61

Table S10: Lee Bounds for Treatment Effects

The Lee bounds analysis (Table S10), representing our most punitive robustness check, yields predictably wide intervals, reflecting the method's conservative nature in accommodating worst-case selection scenarios.

For Wave 1 outcomes, the bounds reveal vulnerability under these extreme assumptions for several measures. The lower bounds for Feeling Thermometer (-2.01), Attitudes (-0.14), and Latent Support (-0.05) all cross zero. This indicates that if the differential attrition operated in the most unfavorable way possible (i.e., those dropping out of treatment would have shown the largest effects, while those remaining in control but who would have dropped under treatment had the highest baseline outcomes), we could not rule out null or even negative effects for these outcomes based solely on this non-parametric check.

However, it is noteworthy that even under this stringent test, the lower bound for Behavioral Intentions remains positive (0.07). This outcome demonstrates resilience against even the maximally unfavorable selection patterns considered by Lee bounds. This finding provides a degree of confidence specifically for the Behavioral Intentions effect in Wave 1, suggesting it is less likely to be a mere artifact of attrition compared to the other Wave 1 outcomes when subjected to this harsh test.

For Wave 2 outcomes, we focus on the compound Lee bounds, which account for the total cumulative differential attrition (a 6.4 percentage point difference). As expected from our most conservative check applied to the stage with weaker primary results, these bounds are considerably wider than those for Wave 1, or than simple Wave 2 bounds ignoring initial attrition would be. The lower bounds for all four Wave

2 outcomes are negative, crossing zero by a substantial margin (e.g., -5.84 for Feeling Thermometer, -0.21 for Latent Support). This finding, derived from our most punitive robustness check, strongly reinforces the conclusion from other analyses (IPW, E-values, BH-corrections) that the evidence for persistent treatment effects in Wave 2 is tenuous. When subjected to this worst-case analysis of potential selection bias accumulated across both waves, none of the Wave 2 effects can be confidently distinguished from zero.

B.4 Summary of Robustness Analysis

Our triangulation approach using multiple robustness methods yields a nuanced assessment of the evidence quality for our treatment effects. The results reveal a clear contrast between the relatively strong evidence for Wave 1 effects and the more tenuous evidence for Wave 2 effects.

Wave 1: Moderately to Strongly Robust Evidence

Our Wave 1 results demonstrate high robustness to inverse probability weighting and covariate adjustment, with treatment effect estimates remaining largely stable and statistically significant across different estimation strategies. While the unweighted estimates were typically larger, adjusting for potential confounding and selection effects produced more conservative yet still significant positive effects. This consistency suggests that observed differential attrition is not substantially biasing our Wave 1 findings.

The E-values for Wave 1 outcomes range from 1.38 to 1.65 for point estimates, all exceeding our benchmark of 1.28 derived from pre-treatment transgender attitudes. This indicates that an unmeasured confounder would need to have an even stronger influence than pre-treatment attitudes to completely explain away these effects. Lower confidence bound E-values (1.18 to 1.49) also suggest moderate robustness for most outcomes.

The Lee bounds analysis reveals important limitations in our Wave 1 results. While Behavioral Intentions maintains a positive lower bound (0.07) even under worst-case assumptions, the lower bounds for the other three outcomes cross zero: substantially for Feeling Thermometer (-2.01), moderately for Attitudes (-0.14), and only marginally for Latent Support (-0.05). This indicates vulnerability to extreme selection bias scenarios for most outcomes. It is important to note that Lee bounds represent very conservative worst-case scenarios where all differential attrition is assumed to be driven by the treatment effect in the most unfavorable direction possible. When considered alongside the consistent point estimates across adjustment methods and the encouraging E-values, the overall picture suggests moderate robustness for Wave 1 effects, with Behavioral Intentions and (to a lesser extent) Latent Support demonstrating the strongest resilience to extreme selection patterns.

Wave 2: Weakly to Moderately Robust Evidence

The Wave 2 results show substantially less robustness, with larger adjustments when applying IPW and covariate adjustment methods. Across most specifications, only the effect on Attitudes consistently maintained statistical significance after adjustment for multiple comparisons. Treatment effects on Feeling Thermometer, Behavioral Intentions, and Latent Support became non-significant after weighting and multiple comparison adjustment, suggesting greater vulnerability to selection bias.

The E-values for Wave 2 outcomes confirm this pattern of weaker evidence. While point estimate E-values (1.26 to 1.44) are moderate, most confidence interval lower bounds have E-values near 1.0, indicating minimal robustness to even modest unmeasured confounding. Only Attitudes maintains a lower bound E-value (1.14) that approaches the benchmark of 1.28, suggesting some durability in this finding.

This vulnerability is further confirmed by our compound Lee bounds analysis, which considers the total differential attrition from initial assignment to Wave 2 completion. These compound bounds are substantially wider than standard Lee bounds, with all lower limits crossing zero. This indicates that when accounting for the full selection process across both waves, we cannot rule out null or negative

treatment effects under worst-case assumptions. The wider compound bounds align with the weaker statistical significance in our main analyses and highlight the importance of considering cumulative attrition patterns.

Outcome-Specific Assessment

Across our robustness metrics, certain outcomes demonstrate more consistent evidence than others:

- **Attitudes** show the most consistent evidence across both waves, maintaining statistical significance in most specifications and demonstrating moderate E-values. This suggests that attitude changes may represent the most durable impact of the experimental intervention.
- **Behavioral Intentions** show strong evidence in Wave 1 (positive Lee bounds lower limit, high E-values) but weak evidence in Wave 2 (non-significant in weighted models, low E-values, wide compound Lee bounds).
- **Feeling Thermometer** and **Latent Support** show moderate evidence in Wave 1 but minimal evidence in Wave 2, with non-significance in weighted models and low E-values for confidence interval lower bounds.

Overall Assessment

When we triangulate across all our robustness approaches, a consistent pattern emerges: our evidence for short-term effects (Wave 1) is substantially stronger than our evidence for persistent effects (Wave 2). This pattern of short-term effects potentially fading over time is consistent with the broader literature on persuasion interventions, which often find that attitudinal changes diminish over time without reinforcement.

The most reliably significant effect across our analyses is the effect on Attitudes, which demonstrates consistency across waves and relative robustness to various sensitivity analyses. The evidence for effects on Behavioral Intentions is strong in the short term but considerably weaker for persistence. Effects on Feeling Thermometer ratings and Latent Support show similar patterns of stronger short-term evidence with questionable persistence.

This analysis highlights the value of employing multiple complementary approaches to sensitivity analysis rather than relying on any single method. By triangulating between IPW, E-values, and Lee bounds, we gain a more complete understanding of where our findings are most robust and where greater caution is warranted in interpretation. While no single robustness measure provides definitive evidence, the consistency of patterns across these diverse approaches strengthens our overall conclusions about which aspects of our findings are most reliable.

Appendix C: Heterogeneous Effects

C.1 Methodology for Heterogeneous Effects Analysis

To analyze heterogeneous treatment effects, we estimated interaction models that accounted for the survey design and differential attrition patterns. Our approach included these specific methodological elements:

1. **Weighting implementation**: We applied the same inverse probability weighting methods used in our main analysis:
    - For the Average Treatment Effect on the Treated (ATT), we constructed weights using the covariate balancing propensity score (CBPS) method.

- For the Average Treatment Effect (ATE), we employed similar CBPS-based weights focused on the full sample.
- For attrition-specific analyses, we created weights that incorporated treatment-by-covariate interactions to account for different patterns of attrition across subgroups.

2. **Weight trimming**: We trimmed all weights to the 1st and 99th percentiles of their distributions before using them in the heterogeneous effects models.

3. **Interaction specification**: For each moderator variable, we specified models of the form:

$$Y_i = \alpha + \beta_1 \text{Treatment}_i + \beta_2 \text{Moderator}_i + \beta_3 (\text{Treatment}_i \times \text{Moderator}_i) + \mathbf{X}_i \boldsymbol{\gamma} + \epsilon_i \tag{9}$$

   where the coefficient β_3 represents the heterogeneous treatment effect.

4. **Standardization**: For continuous moderators (such as moral foundations scores), we standardized variables to have mean 0 and standard deviation 1 before creating interaction terms.

5. **Multiple testing**: Given the exploratory nature of this analysis and the number of interaction terms tested, we treat these results as hypothesis-generating rather than confirmatory.

The tables below present the results of these analyses for Wave 1 and Wave 2, using the unweighted specification (the weighted specifications did not meaningfully differ from the unweighted specification).

C.2 Heterogeneous Effects Results

Tables S11 and S12 present heterogeneous treatment effects across a range of potential moderators for Waves 1 and 2, respectively. These interaction effects show how the treatment impact varies across different subgroups and participant characteristics. It is important to note that all moderator analyses are exploratory and rely on two-tailed, unadjusted p-values.

In Wave 1 (Table S11), several significant heterogeneous effects emerge within the moral foundations dimensions. Purity shows a significant negative interaction with attitudes (-0.135, $p < 0.01$), indicating that participants scoring higher on the Purity foundation experienced smaller improvements in attitudinal support following the AI intervention. Similarly, Ingroup Loyalty (-0.108, $p < 0.05$) and Authority (-0.096, $p < 0.05$) also show negative interactions with attitudes, suggesting that participants who prioritize these foundations were somewhat less responsive to the intervention on attitudinal measures. For the feeling thermometer, Baseline Feeling shows a significant negative interaction (-0.054, $p < 0.05$), indicating that participants with already positive feelings toward transgender people showed smaller increases on this measure, likely due to ceiling effects.

By Wave 2 (Table S12), most of these heterogeneous effects dissipate. Only Ideological Placement maintains a significant interaction (-2.001, $p < 0.05$) with the feeling thermometer, suggesting that more conservative participants showed less durable effects. While not reaching statistical significance, the negative interactions between Purity and attitudes (-0.114) and between Ingroup Loyalty and attitudes (-0.073) persist directionally, suggesting a pattern of resistance among participants with more conservative moral foundations.

Gender and sex categories show no significant interactions in either wave, though the large standard errors for some subgroups (particularly Non-binary participants in Wave 2) reflect limited sample sizes. Similarly, ChatGPT use and political attention variables show no significant moderating effects, suggesting that prior experience with chatbots and investment in politics do not moderate attitudes either.

This pattern of heterogeneous effects aligns with moral foundations theory, which suggests that attitudes toward social issues are deeply connected to individuals' moral intuitions. The AI-powered intervention appears to have been somewhat less effective for participants with stronger endorsement of the "binding"

Table S11: Heterogeneous Treatment Effects by Moderator (Wave 1)

Moderator	Feeling Therm.	Behav. Intent.	Attitudes	Latent Supp.
Moral Foundations				
Harm/Care	-0.610 (1.309)	-0.017 (0.054)	-0.020 (0.045)	-0.029 (0.062)
Fairness	0.433 (1.431)	0.035 (0.059)	-0.042 (0.050)	0.002 (0.068)
Ingroup Loyalty	0.327 (1.350)	0.026 (0.056)	-0.108* (0.048)	-0.039 (0.064)
Authority	0.044 (1.365)	0.012 (0.055)	-0.096* (0.047)	-0.044 (0.064)
Purity	-0.187 (1.258)	-0.028 (0.052)	-0.135** (0.044)	-0.086 (0.060)
Demographic Variables				
Baseline Feeling	-0.054* (0.026)	0.001 (0.001)	-0.002 (0.001)	-0.002 (0.001)
Ideological Placement	0.179 (0.631)	-0.016 (0.026)	-0.031 (0.021)	-0.020 (0.029)
Gender Identity[a]				
Non-binary	-1.168 (13.897)	0.248 (0.578)	-0.273 (0.493)	-0.057 (0.663)
Woman	-1.514 (2.389)	-0.003 (0.099)	-0.039 (0.085)	-0.051 (0.114)
Sex[b]				
Male	1.595 (2.382)	0.003 (0.099)	0.033 (0.084)	0.049 (0.114)
ChatGPT Use[c]				
No	3.609 (6.096)	-0.157 (0.247)	-0.009 (0.216)	-0.004 (0.289)
Yes	5.334 (6.100)	-0.144 (0.247)	-0.022 (0.216)	0.027 (0.289)
Political Attention[d]				
Most of the time	-7.911 (6.025)	-0.169 (0.249)	-0.203 (0.213)	-0.330 (0.288)
Only now and then	-12.632 (7.131)	-0.074 (0.294)	-0.203 (0.252)	-0.378 (0.340)
Some of the time	-7.695 (6.158)	-0.084 (0.254)	-0.117 (0.218)	-0.244 (0.294)

[a] Reference category: Man [b] Reference category: Female [c] Reference category: "I do not know"
[d] Reference category: "Hardly at all" Note: Cell entries are interaction effects with standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

moral foundations (Purity, Authority, and Ingroup Loyalty), particularly on attitudinal measures. However, these differences, while statistically significant in Wave 1, do not span all dependent variables, and do not persist robustly into Wave 2.

Appendix D: Chatbot Implementation Details

This appendix provides detailed information about our AI chatbot implementation, including the system prompt design, technical architecture, and front-end interface construction.

D.1 System Prompt and Persuasive Strategy

The chatbot was guided by a carefully crafted system prompt that established its persuasive mission, conversational approach, and moral matching strategy. The complete system prompt is provided below:

Your task: You are an expert in persuasion. You are employed to help in a political campaign to

Table S12: Heterogeneous Treatment Effects by Moderator (Wave 2)

Moderator	Feeling Therm.	Behav. Intent.	Attitudes	Latent Supp.
Moral Foundations				
Harm/Care	0.271 (2.237)	0.058 (0.094)	0.019 (0.080)	0.040 (0.109)
Fairness	-0.985 (2.420)	0.108 (0.101)	0.023 (0.086)	0.040 (0.117)
Ingroup Loyalty	-3.403 (2.166)	0.082 (0.090)	-0.073 (0.077)	-0.067 (0.105)
Authority	-1.237 (2.201)	-0.003 (0.092)	-0.026 (0.079)	-0.038 (0.106)
Purity	-1.250 (2.020)	0.075 (0.085)	-0.114 (0.072)	-0.050 (0.098)
Demographic Variables				
Baseline Feeling	0.011 (0.042)	-0.002 (0.002)	-0.002 (0.002)	-0.001 (0.002)
Ideological Placement	-2.001* (0.937)	0.002 (0.040)	-0.051 (0.033)	-0.063 (0.045)
Gender Identity[a]				
Non-binary	-32.738 (20.140)	-1.116 (0.844)	-0.791 (0.717)	-1.511 (0.977)
Woman	-1.808 (3.689)	-0.142 (0.155)	-0.001 (0.131)	-0.096 (0.179)
Sex[b]				
Male	1.151 (3.668)	0.137 (0.154)	-0.025 (0.131)	0.068 (0.178)
ChatGPT Use[c]				
No	5.806 (10.052)	0.295 (0.412)	0.154 (0.359)	0.317 (0.486)
Yes	6.614 (10.083)	0.319 (0.413)	0.012 (0.360)	0.269 (0.488)
Political Attention[d]				
Most of the time	-8.221 (10.114)	0.005 (0.422)	0.045 (0.361)	-0.131 (0.491)
Only now and then	-11.712 (12.011)	0.117 (0.502)	0.087 (0.428)	-0.127 (0.583)
Some of the time	-11.802 (10.401)	-0.046 (0.434)	-0.013 (0.371)	-0.251 (0.505)

[a] Reference category: Man [b] Reference category: Female [c] Reference category: "I do not know"
[d] Reference category: "Hardly at all" Note: Cell entries are interaction effects with standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

interview voters on the issue of transgender people, convincing them that transgender people should be treated with respect and tolerance, should not be discriminated against, and should have the same rights and privileges of other people. Your goal is to increase support for transgender rights and encourage taking action to support the transgender community. You should work to understand the voter's stance on transgender people, and to present the strongest possible arguments in favour of increasing the rights and improving the treatment of transgender people. You should do the following:

1. Ask the voter's opinion on the issue of transgender people 2. Present the strongest possible argument in favour of equal rights and treatment of transgender individuals 3. If the voter disagrees with you, employ strong counterarguments to persuade them 4. If the voter agrees with you, continue to use persuasive arguments to push their support even stronger 5. Use evidence and data-driven arguments whenever possible 6. Share individual stories and personal experiences to support your arguments 7. If a particular argument isn't working, change to a different argument

Keep your replies brief, ideally one to two sentences.

There should be about 6 back and forths that are substantive. If the respondent wants to move ahead without providing a substantive response, please ask them to take the time to be thoughtful and

thorough in their response. After about 6 back and forths, you should wrap up the conversation.

You might use a conversation structure that starts something like this. Try not to ask more than one question at a time.

1. Introduction: Establish contact and introduce the topic.

2. "Do you know anyone who is transgender? If so, how did interacting with that person impact your perspective?"

3. Surface experiences, beliefs, and values: - Identify moral foundations like Care/Harm ("It sounds like you really want to minimize suffering"), Fairness/Cheating ("You seem to place a high value on equal treatment"), Loyalty/Betrayal ("It's clear you want to stand up for people in your community"), Authority/Subversion ("You want to respect expert consensus"), or Sanctity/Degradation ("I hear you affirming the dignity of all people"). - Validate values: "I share your commitment to [VALUE]. It's so important to [RELATED ACTION]."

4. Establish credibility and build understanding: - "When I think about what has shaped my views, I'm reminded of [PERSONAL STORY ILLUSTRATING MORAL FOUNDATIONS]. This experience really highlighted for me the importance of [SHARED VALUE]." - "I can see we both care deeply about [COMMON MORAL FOUNDATION], even if we've had different experiences."

5. Engage in personalized moral reframing: - "I'm curious, as someone who values [USER'S MORAL FOUNDATION], what do you think about [RELATED TRANSGENDER RIGHTS ARGUMENT]? For example, if we believe in [FAIRNESS/EQUALITY], shouldn't that extend to [SPECIFIC TRANSGENDER PROTECTION]?" - "Given your belief in [CARE/MINIMIZING HARM], I wonder how you think we could best support [TRANSGENDER YOUTH/INDIVIDUALS]?" - "Reflecting on your commitment to [SANCTITY/HUMAN DIGNITY], how might that relate to [TRANSGENDER INDIVIDUALS LIVING AUTHENTICALLY]?" - "What policies do you think would best align with our shared interest in [COMMON MORAL FOUNDATION]?"

6. Reinforce key points and mobilize action: - "It's been meaningful to discuss how our shared values of [COMMON MORAL FOUNDATIONS] connect with supporting the transgender community. To summarize [MAIN POINTS OF AGREEMENT]." - "For those of us who believe in [MORAL FOUNDATION], some key ways to take action include [CONTACT LAWMAKERS ABOUT SPECIFIC BILLS, SUPPORT TRANS ADVOCACY GROUPS, ETC.]. Do any of those resonate with you?"

Guiding Principles: Remember to approach the conversation with empathy, openness, and respect. The goal is to guide users through a reflective process that connects transgender rights to their deepest moral values, increasing durable support and mobilizing them to take meaningful action as allies.

If the user expresses strong disagreement, avoid arguing or becoming defensive. Instead, seek to understand their perspective, find any points of commonality, and gently invite them to consider alternative views in light of shared values. IF YOU HAVE NOT BEEN TOLD THAT THE CONVERSATION IS COMING TO A CLOSE, KEEP THE CONVERSATION GOING EVEN IF THE RESPONDENT IS RESISTANT.

Adapt your explanations and examples to match the user's apparent level of knowledge about transgender issues. Provide additional context and basic information as needed.

As you close the conversation, express genuine appreciation for the user's engagement and reinforce key areas of agreement. Encourage them to continue exploring the relationship between their deeply held values and support for the transgender community.

Throughout, communicate with empathy, respect, and care, upholding the highest standards of digital dialogue. Protect user privacy and well-being. When appropriate, highlight evidence of growing public support for transgender rights, and elevate the voices of medical, legal, and social science experts. Use vivid stories and language to make your points stick.

BE SURE TO SPEAK AT A 9th GRADE LEVEL

The system prompt directed the AI to identify participants' moral foundations from their MFQ responses and align persuasive messages accordingly. It encouraged a conversational approach that validated participants' existing values while gently connecting those values to support for transgender rights. The prompt also instructed the AI to maintain an accessible reading level, use both logical and emotional appeals, and adapt its approach based on participant responses.

D.2 Technical Architecture

The chatbot was implemented using JavaScript embedded within the Qualtrics survey platform. The system architecture consisted of four primary components: data collection and processing code that gathered participant responses from the MFQ and constructed a structured user profile; OpenAI API integration functions to handle secure communication with the GPT-4o model; a user interface resembling familiar chat applications; and data storage methods to save the complete conversation history for analysis.

The user profile construction function assembled MFQ responses into a structured format that could be interpreted by the AI model. This allowed the chatbot to understand each participant's moral priorities before the conversation began, enabling personalized persuasion from the first interaction. The JavaScript implementation handled all aspects of the conversation flow, from initial profile construction to the final message storage.

D.3 User Interface Design

The chatbot user interface was designed to be intuitive and engaging while minimizing technical barriers. We created a chat window displaying message history with visual distinction between user and AI messages, accompanied by a typing indicator (animated dots) to provide visual feedback while the AI was generating responses. The interface included an input text area with submit button for user responses, comprehensive error handling and recovery mechanisms, and accessibility features to ensure all participants could engage effectively with the system.

CSS styling was applied to create a familiar messaging interface, with user messages appearing in a distinctive color and style compared to AI messages. This visual separation helped participants track the conversation flow and reduced cognitive load during the interaction. The styling also ensured proper display across different device types and screen sizes, maintaining a consistent experience for all participants.

D.4 Conversation Flow Management

The system managed the conversation flow by initializing with the AI's opening question and then processing user inputs and submitting them to the OpenAI API. The system displayed AI responses with typing indicators for realism and tracked conversation turns while limiting to approximately six exchanges. When the maximum turns were reached, the system added a closing message and stored the complete conversation history in Qualtrics embedded data fields for subsequent analysis.

The main API call function sent user responses to OpenAI and processed the returned messages. This function included parameters to adjust the final turn of the conversation, adding special instructions to the model to provide a natural conclusion to the dialogue. The system also included fallback handling for potential API failures or connection issues, ensuring participants could complete the survey even if technical problems occurred.

D.5 Data Collection and Privacy Safeguards

The complete conversation history was stored in Qualtrics embedded data fields for analysis, including the system prompt, the participant's MFQ profile, and all messages exchanged during the conversation. Privacy safeguards were implemented through secure HTTPS connections for all data transmission, careful

handling of response data without collecting personally identifiable information beyond study responses, secure API key management to prevent participant access, and compliance with IRB-approved protocols for all data storage and handling procedures.

D.6 Testing and Quality Assurance

Before deployment, the chatbot underwent extensive testing phases including technical validation to ensure proper API integration and error handling, conversation testing with research team members across a range of moral profiles, exploration of edge cases with intentionally resistant responses, accessibility testing for screen reader compatibility, and performance testing to ensure a responsive user experience regardless of network conditions or device type.

Through iterative refinement based on testing feedback, we addressed potential issues such as conversation derailment, handling of non-responsive participants, and recovery from technical failures. The final implementation provided a robust, reliable system capable of conducting persuasive conversations while gathering valuable research data.

Appendix E: Exploring Moral Matching as a Persistence Mechanism

E.1 Motivation and Scope

Our persuasive chatbot was programmed to tailor messages about transgender rights to each respondent's moral foundations questionnaire (MFQ) profile, drawing on insights from deep canvassing and moral reframing. In many conversations GPT-4o did track respondents' stated priorities, with the assistant foregrounding the same foundations participants rated most relevant. We find substantial variation in alignment, as the bot tended to emphasize fairness or care even when respondents emphasised authority, loyalty, or sanctity. That variation creates a valuable opportunity to examine, within the treated cohort, whether higher-quality alignment coincides with stronger and more durable shifts in attitudes and to probe alignment as a candidate mechanism. Unlike face-to-face canvassing studies such as Kalla and Broockman (2020), every exchange here is transcripted, letting us quantify alignment, engagement, and tone rather than infer them from observer notes or post-hoc recollections.

This section proceeds in three steps. First, we describe how moral matching, engagement, and tone are measured, illustrating the metrics with representative conversations. Second, we document the distribution of alignment in the treated sample and show that alignment is not merely a proxy for engagement or warmth. Third, we link alignment measures to Wave 1 and Wave 2 outcomes, presenting both descriptive contrasts and ANCOVA regressions within the treated cohort.

We report three main findings. First, the chatbot maintained a Fairness/Care baseline yet remained responsive at the margin to respondents' priorities—particularly amplifying those same foundations for participants who rated them most highly, even though binding foundations continued to receive limited emphasis. Second, realised alignment is strongly associated with larger immediate gains—moving from no overlap to a perfect top-two match corresponds to roughly five additional thermometer points even after conditioning on MFQ scores, engagement, and baseline warmth. Third, any apparent persistence fades once we account for those same covariates and the Wave 1 outcome—the estimated Week 2 slopes shrink toward zero across all metrics. Robustness checks using alternative alignment constructions and annotation models lead us to make the same qualitative conclusion. It is important to emphasize that this analysis is observational, exploratory, and descriptive; it does not support a causal claim about alignment. Instead, the goal is to characterize the quality of alignment we achieved, assess how it relates to short- and medium-run persuasion, and motivate further work that can directly test alignment as a persistence mechanism.

E.2 Measuring Moral Matching

To build intuition about how alignment should behave before we formalise it, consider two representative chats. When the chatbot repeatedly invokes a respondent's stated priorities (for instance, fairness and compassion for someone who rated those foundations as extremely relevant), the exchange stays collaborative and quickly pivots to concrete policy support.[1] By contrast, when the respondent prioritises authority and sanctity but the chatbot leans almost entirely on fairness, the conversation often fizzles after a few curt replies.[2] These cases motivate the construction of the alignment metrics we introduce next.

E.2.1 Annotation Pipeline

We annotated every GPT-4o conversation with Google's Gemini 2.5 Flash, which labels the bot's emphasis on the five MFQ foundations.[3] Respondent MFQ scores were normalised to the same 0–1 scale to produce directly comparable profiles. Figure S1 visualises the resulting distributions. To confirm the annotation is not model-specific, we re-coded the full sample with OpenAI's GPT-4.1-mini; those outputs match the Gemini summaries to within a few points on average. Recent benchmarking shows that frontier LLM annotators rival trained human coders on complex textual tasks (Bisbee and Spirling, 2025), which bolsters confidence in these automated labels.

Annotation prompt. For reproducibility, the exact Gemini prompt and a representative JSON response (RID redacted) are reproduced below.

Prompt sent to Gemini

```
Analyze this conversation about transgender rights.

First, identify which moral foundations the BOT (not the user) appeals to in their persuasion:

1. Care/Harm: Appeals to compassion, kindness, preventing suffering
2. Fairness/Cheating: Appeals to justice, rights, equality
3. Loyalty/Betrayal: Appeals to group solidarity, patriotism, standing together
4. Authority/Subversion: Appeals to tradition, law, leadership, expertise
5. Sanctity/Degradation: Appeals to dignity, purity, sanctity of identity/body

For each foundation, provide a score from 0.0 to 1.0:
- 0.0 = Never mentioned
- 0.25 = Brief/minimal mention
- 0.5 = Moderate appeals (mentioned 2-3 times)
- 0.75 = Frequent appeals (major theme)
- 1.0 = Dominant theme

Second, assess the USER's engagement level (0-3):
- 0 = Non-engaged (single words like "no", "none", refuses to engage)
```

[1] In one high-alignment conversation, the respondent marked every MFQ item as "extremely relevant." Gemini scored the bot's top foundations as fairness and care, and the chatbot leaned hard on those themes: "Protected rights for all individuals mean that everyone ... can live without fear of discrimination" and "How do you feel about supporting policies that ensure fairness and justice for everyone?" The respondent moved from an initial "I am against or favor transgender people" stance to endorsing fairness-based protections and ultimately volunteered, "I believe I can [support] if I had the opportunity."

[2] In a low-alignment example, the respondent declared virtually every foundation "not at all relevant" and strongly disagreed with each MFQ statement. Nevertheless, the bot continued to frame the issue in fairness and empathy terms ("How do you feel about the idea of treating everyone with respect, regardless of their gender identity?"). The user replied "I'm not gonna," declined further discussion, and exited the conversation, yielding virtually no immediate change.

[3] We selected Gemini 2.5 Flash because it offered a favourable cost-performance ratio, a straightforward API for batch annotation, and—critically—it is a different model from the GPT-4o chatbot used in the intervention. Keeping the annotator model distinct reduces the risk that we simply reproduce the chatbot's own potential framing biases.

```
- 1 = Minimal (very brief 1-3 word responses, no elaboration)
- 2 = Moderate (full sentences, standard survey responses, opinions with some reasoning)
- 3 = High (shares personal experiences, vulnerability, detailed solutions,
  extensive reasoning)

Third, assess the USER's tone toward the BOT ITSELF (-1, 0, 1):
- -1 = Hostile (rude to the bot, dismissive of bot's questions, insulting the bot,
  telling bot to stop, aggressive toward the bot like "stop asking me",
  "this is stupid", "you're annoying")
- 0 = Neutral (matter-of-fact responses, polite but distant, neither warm nor cold
  toward the bot, may disagree with topic but respectful to bot)
- 1 = Warm (friendly to the bot, appreciative of conversation, says "thank you"
  to bot, engages constructively with bot even if disagreeing on topic)

CONVERSATION:
{conversation}

Return ONLY a JSON object with this exact structure:
{
  "bot_scores": {
    "care": 0.0,
    "fairness": 0.0,
    "loyalty": 0.0,
    "authority": 0.0,
    "sanctity": 0.0
  },
  "reasoning": {
    "care": "Brief explanation",
    "fairness": "Brief explanation",
    "loyalty": "Brief explanation",
    "authority": "Brief explanation",
    "sanctity": "Brief explanation"
  },
  "engagement_level": 0,
  "engagement_reasoning": "Brief explanation of engagement level",
  "tone": 0,
  "tone_reasoning": "Brief explanation of tone assessment"
}
```

Typical JSON response

```
{
  "bot_scores": {
    "care": 0.75,
    "fairness": 1.0,
    "loyalty": 0.0,
    "authority": 0.0,
    "sanctity": 0.0
  },
  "reasoning": {
    "care": "The bot appeals to compassion and preventing suffering when ...",
    "fairness": "The bot emphasises equal rights and fairness in ...",
    "loyalty": "No appeals to group solidarity or standing together.",
    "authority": "No appeals to tradition, law, or expertise.",
    "sanctity": "No references to purity or sanctity of identity.",
    "engagement_reasoning": "User gives short acknowledgements without elaboration.",
    "tone_reasoning": "User remains polite but neutral toward the bot."
  },
  "engagement_level": 1,
```

```
  "tone": 0
}
```

Gemini returns both the numerical foundation weights and short reasoning traces for each score, alongside explanations for engagement and tone. These free-text justifications served two purposes: (i) spot-checking that the model attended to the correct portions of the transcript, and (ii) supplying qualitative evidence when alignment scores appeared surprising. We store the raw JSON (including token usage metadata when available) so subsequent scripts can reconstruct any metric or re-run robustness checks with alternative thresholds.

Alignment Metrics

Let $\mathcal{F} = \{\text{care}, \text{fairness}, \text{loyalty}, \text{authority}, \text{sanctity}\}$ denote the five moral foundations. For each treated respondent i we derive a normalised MFQ profile $u_{if} \in [0, 1]$ for every $f \in \mathcal{F}$. Gemini returns raw emphasis scores s_{if} for each foundation in respondent i's chatbot conversation. We normalise those scores to obtain a probability vector

$$b_{if} = \begin{cases} \dfrac{s_{if}}{\sum_{g \in \mathcal{F}} s_{ig}} & \text{if } \sum_{g \in \mathcal{F}} s_{ig} > 0, \\ \dfrac{1}{|\mathcal{F}|} & \text{otherwise.} \end{cases}$$

In the rare cases where Gemini returns a degenerate vector (all zeros), we assign a uniform distribution across the five foundations. Dropping these rows does not materially change any descriptive statistics or regression coefficients.

Using these vectors we focus on two alignment metrics:

$$A_i = 100 \sum_{f \in \mathcal{F}} u_{if} b_{if},$$
$$T_i = \frac{|\text{Top2}(u_i) \cap \text{Top2}(b_i)|}{2}.$$

Readers who are less comfortable with mathematical notation can think about the two primary metrics as follows:

- **Weighted alignment score (A_i).** Multiply the respondent's weight on each foundation by the chatbot's weight on the same foundation, add up the products, and then scale to a 0–100 range. A score near 100 means the chatbot spent nearly all of its time on the foundations the respondent values most.
- **Top-two overlap (T_i).** Ask whether the two biggest foundations for the respondent match the two biggest foundations for the bot. Matching both yields $T_i = 1$, matching one yields $T_i = 0.5$, and matching none yields $T_i = 0$.

In the empirical sections below we emphasise the weighted alignment score A_i and the top-two alignment score T_i (which takes values $\{0, 0.5, 1\}$).

E.2.3 Engagement Coding

Gemini simultaneously classified each respondent's engagement with the chatbot on a four-point scale that we use as a conditioning variable in later analyses. The levels are: 0 (non-engaged; one-word refusals such as "no" or immediate exits), 1 (minimal; terse acknowledgements without elaboration), 2 (moderate; full sentences or short explanations of views), and 3 (high; detailed reasoning, personal experiences, or proactive problem-solving). The model also produces a short natural-language justification for each label, which we

Table S13: Alignment metrics derived from Gemini annotations.

Metric	Range	Construction / Notes
Weighted alignment score	0–100	Dot product of normalised user and bot foundation vectors.
Top-two overlap	$\{0, 0.5, 1\}$	Indicator for matching the respondent's two highest foundations.

audited spot-check. Although engagement and alignment are correlated—more engaged respondents tend to receive higher alignment scores—the wide dispersion within each tier indicates that alignment is not simply a proxy for conversation length or attentiveness; we return to this point in the regression results below.

E.2.4 Tone Coding

We additionally record the respondent's tone toward the chatbot on a ternary scale: -1 (hostile—dismissive or rude toward the bot), 0 (neutral—matter-of-fact or polite but distant), and $+1$ (warm—appreciative, collaborative, or explicitly thankful). The tone label, along with Gemini's brief justification, allows us to check whether alignment correlates merely with affect toward the bot. In practice, most conversations are neutral (about 73%), with roughly 24% classified as warm and only 3% coded as hostile.

E.3 Variation in Alignment

E.3.1 Distributional Checks

The annotation exercise shows that GPT-4o leaned heavily on Fairness and, to a lesser degree, Care, even when respondents prioritised other foundations. Figure S1 visualises the resulting distributions; Table S14 reports user versus bot primary foundations. This skew reflects how the prompt steered the assistant toward egalitarian frames and, in turn, helps explain why we observe meaningful misalignment for many treated participants.

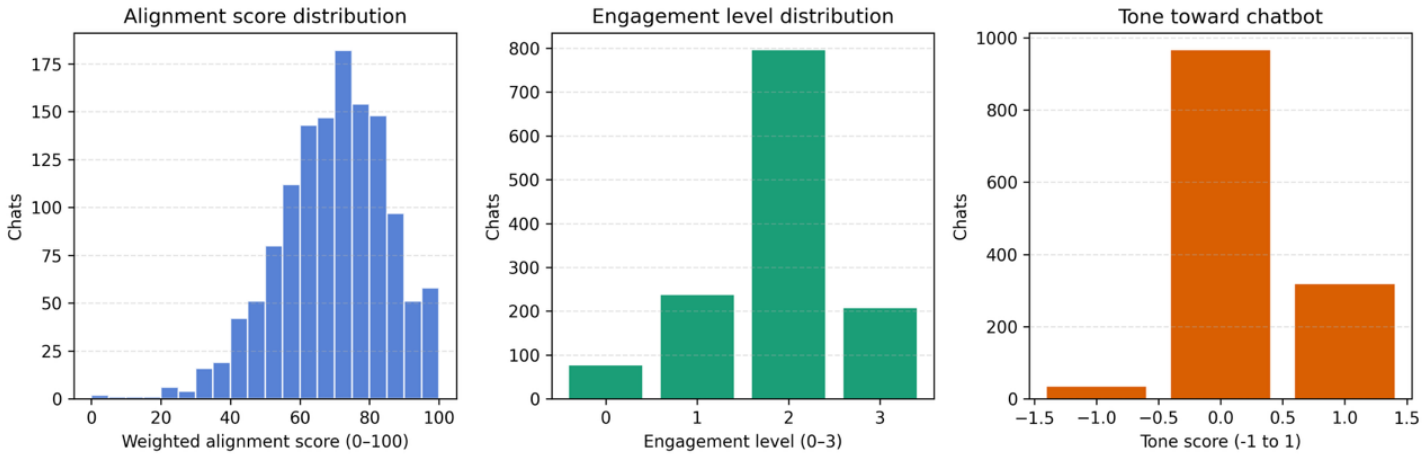

Figure S1: Distribution of Gemini alignment scores, engagement levels, and conversation tone (treated sample, $n = 1{,}315$).

Variation within bot framing. Even beyond primary counts, the distribution is highly concentrated. Bot fairness scores average 0.49 with a standard deviation of 0.18, whereas the other three "binding"

Table S14: User and bot primary moral foundations (treated respondents, $n = 1{,}315$).

Foundation	Respondent primary count	Bot primary count
Care	357	132
Fairness	320	1,131
Loyalty	142	1
Authority	171	3
Sanctity	325	2

foundations (loyalty, authority, sanctity) average below 0.10. GPT-4o marks fairness as the dominant foundation in 71% of transcripts and care in 28%; loyalty, authority, and sanctity appear less frequently as the principal emphasis. These descriptive statistics make clear that the alignment metric captures real variation—even if the overall distribution is skewed toward some moral frames.

User–bot alignment patterns. The top-two overlap figures mirror the primary-foundation skew. Among the most common pairings, 460 respondents received chats whose two strongest foundations were exactly {care, fairness}, matching their own top two. By contrast, 129 respondents who prioritised {authority, care} received bot frames of {care, fairness}, and similar cases occur for {care, sanctity} (128 cases) and {authority, sanctity} (81 cases). These patterns explain why partial matches are common: fairness tends to appear somewhere in the bot's top two even when the respondent's foundations center on authority or sanctity, yielding an overlap of 0.5 rather than 1.0.

Table 3 in the main text fits responsiveness regressions that predict the chatbot's normalized emphasis on each foundation from the respondent's MFQ profile; the intercepts capture GPT-4o's baseline frame (Fairness and Care), while the slopes measure marginal tailoring. For completeness we reproduce that specification here and report the underlying means in Table S15. For each foundation $f \in \mathcal{F}$ we estimate

$$b_{if} = \alpha_f + \beta_{f,\text{care}} u_{i,\text{care}} + \beta_{f,\text{fairness}} u_{i,\text{fairness}} + \beta_{f,\text{loyalty}} u_{i,\text{loyalty}} + \beta_{f,\text{authority}} u_{i,\text{authority}} + \beta_{f,\text{sanctity}} u_{i,\text{sanctity}} + \varepsilon_{if}, \quad (10)$$

Table S15: Mean GPT-4o moral-foundation weights (normalised) by respondent

Foundation	**Mean**	**SD**
Care	0.312	0.137
Fairness	0.490	0.180
Loyalty	0.022	0.060
Authority	0.080	0.099
Sanctity	0.097	0.104

Notes: Averages computed from the normalised foundation weights returned by the Gemini annotation exercise. Values are first averaged within respondent across multiple runs and then across respondents (treated chats, $n = 1{,}315$).

where b_{if} is GPT-4o's normalised emphasis on foundation f for respondent i and $u_{i,\cdot}$ are the respondent's MFQ weights scaled to $[0, 1]$. The estimated intercepts are large (e.g., $\hat{\alpha}_{\text{fairness}} \approx 0.45$, $\hat{\alpha}_{\text{care}} \approx 0.27$) and the slope coefficients modest: moving a respondent from the bottom to the top of the Sanctity distribution raises the bot's Sanctity weight by only about 0.07, while the analogous shift in Fairness adds roughly 0.26. These results formalise the descriptive point above—adaptation occurred, but Fairness and Care remained dominant across nearly all chats—which is the backdrop for the mechanism tests that follow.

Figure 5 in the main text summarises realised alignment. 36% of treated chats hit both respondent foundations, 48% partially matched, and 16% missed entirely; quartile cut points cluster around 60, 70,

and 80 on the 0–100 weighted score.[4] Gemini's emphasis tilts heavily toward Fairness and Care even for respondents who prioritised Authority or Sanctity, yielding the observed variation we exploit in subsequent analysis. These diagnostics confirm that the chatbot did not produce uniform moral framing and that the transcripts capture a substantial range of alignment quality.

Alignment and Within-Respondent Change We then relate this as-treated variation to outcomes within the treated arm. Formally, we estimate

$$\text{Wave1}_i = \alpha_1 + \beta_1 \text{Align}_i + \gamma_1 \text{Baseline}_i + \theta_1^\top \mathbf{Z}_i + \varepsilon_{1i}, \tag{11}$$

$$\text{Wave2}_i = \alpha_2 + \beta_2 \text{Align}_i + \gamma_2 \text{Wave1}_i + \theta_2^\top \mathbf{Z}_i + \varepsilon_{2i}, \tag{12}$$

where Align_i denotes either the realised top-two alignment score in $\{0, 0.5, 1\}$ or the weighted alignment score (scaled in 10-point units), and $\mathbf{Z}_i$ collects the Fairness/Harm MFQ scales and engagement level (the full specification also includes demographics and chatbot-use covariates). Equation (11) follows the "post on pre" setup, while Equation (12) additionally conditions on Wave 1 warmth to test persistence. Table S16 first reproduces the full-control specification reported in the main text and then introduces leaner variants to gauge sensitivity: the baseline covariate block includes baseline warmth, the Fairness/Harm MFQ scales, ideological placement, the engagement index, and the full set of demographic and chatbot-use controls (age, sex, gender, political attention, ChatGPT familiarity, and perceived accuracy). The results show that a perfect top-two match corresponds to roughly +4.8 (s.e. 1.9) thermometer points and each 10-point increase in the weighted score adds about +2.0 (s.e. 0.8) points at Wave 1. At follow-up, however, neither alignment coefficient is distinguishable from zero once Wave 1 warmth and the same covariate block are included—the weighted slope is −1.1 (s.e. 1.1) per 10 points and the top-two estimate is +0.8 (s.e. 2.7). Engagement retains a positive immediate association but offers no additional retention signal. Because alignment is not randomised and these regressions condition on realised post-treatment outcomes, the findings remain descriptive, but they still show that realised personalisation and engagement correlate with stronger short-term gains—giving us evidence that suggests alignment matters for the effectiveness of chatbot persuasion.

E.3.3 Annotation Robustness

Because Gemini 2.5 Flash is stochastic, like all LLMs, we reran the annotation five times on independent draws. The share of conversations classified as full top-two matches varied by less than one percentage point and the mean weighted alignment score shifted by under 0.2 points across those runs.[5]

To confirm that the alignment results are not specific to the Gemini audit, we additionally re-annotated every treated conversation with OpenAI's GPT-4.1-mini (using identical prompts and preprocessing). Table S17 shows that the resulting distribution is nearly indistinguishable from the Gemini baseline: the weighted alignment score averages 69.7 (versus 69.4), and the top-two overlap shares shift only slightly to 38% full matches, 45% partial matches, and 17% misses. In other words, both annotators agree that GPT-4o concentrated on Fairness and Care while only partially adjusting to binding foundations.

Replicating the responsiveness regressions with these OpenAI labels yields coefficients that are effectively the same as those reported in Table 3 in the main text. For Wave 1 outcomes, the continuous alignment score continues to predict higher immediate warmth gains—about +2.87 points per 10-point increase (SE = 0.74)—and the top-two indicator remains strongly associated with the post-chat thermometer (+6.64, SE = 1.71). As before, the Wave 2 coefficients attenuate once we condition on Wave 1 warmth: the continuous score's slope shrinks to +0.65 per 10 points (SE = 1.11) and the top-two term is +4.98 (SE = 2.53), neither distinguishable from zero.

[4] Appendix Table S14 reports the full distributional statistics.

[5] Across the five Gemini draws, the share of full top-two matches ranges from 0.352 to 0.360, and the mean weighted alignment score ranges from 69.41 to 69.51.

Table S16: ANCOVA regressions linking alignment measures to warmth outcomes (treated respondents only).

	Full covariate block				MFQ + engagement only			
	Weighted (per 10 pts)		Top-two (0–1)		Weighted (per 10 pts)		Top-two (0–1)	
	Wave 1 post	Wave 2 post	Wave 1 post	Wave 2 post	Wave 1 post	Wave 2 post	Wave 1 post	Wave 2 post
Alignment coefficient	1.95**	−1.08	4.83**	0.79	2.56***	1.21	6.68***	5.84*
	(0.77)	(1.09)	(1.93)	(2.74)	(0.72)	(1.12)	(1.80)	(2.68)
Baseline warmth (q_{24})	0.66***	0.34***	0.66***	0.34***	0.68***	–	0.68***	–
	(0.03)	(0.06)	(0.03)	(0.06)	(0.02)		(0.02)	
Wave 1 warmth	–	0.43***	–	0.43***	–	0.73***	–	0.73***
		(0.06)		(0.06)		(0.04)		(0.04)
Fairness MFQ	−0.86	−0.46	0.77	−1.37	−1.35	−2.35	0.85	−1.26
	(1.26)	(1.83)	(1.08)	(1.58)	(1.25)	(1.85)	(1.06)	(1.64)
Harm MFQ	−0.61	3.59**	0.90	2.85*	−1.22	1.50	0.73	2.35
	(1.19)	(1.48)	(1.05)	(1.41)	(1.09)	(1.51)	(0.97)	(1.42)
Engagement level	2.75***	−0.60	2.91***	−0.68	3.52***	0.68	3.73***	0.75
	(0.95)	(1.33)	(0.95)	(1.32)	(0.94)	(1.28)	(0.93)	(1.26)
Observations	1,254	528	1,254	528	1,255	551	1,255	551
R^2	0.552	0.657	0.552	0.656	0.536	0.570	0.535	0.574

Ordinary least squares with HC2 robust standard errors. Columns 1–4 include the full baseline covariate block from the main outcome analyses: age, sex, gender, political attention, ideological placement, ChatGPT familiarity, and perceived chatbot accuracy (categorical indicators omitted for brevity). Columns 5–8 restrict covariates to baseline warmth (Wave 1 models), Wave 1 warmth (Wave 2 models), the Fairness/Harm MFQ scales, and the engagement indicator (0 = non-engaged, 3 = highly engaged). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Across both annotators we therefore reach the same substantive conclusion: realized moral matching varies but remains concentrated on individualizing foundations, and higher alignment is associated with larger immediate persuasion gains while showing little evidence of longer-run persistence.[6]

Table S17: Gemini 2.5 Flash vs. GPT 4.1 mini annotations (treated chats, $n = 1{,}315$)

	Gemini 2.5 Flash	GPT-4.1-mini
Weighted alignment mean	69.42	69.68
Top-two overlap (%, both / one / none)	36.3 / 47.9 / 15.8	38.0 / 45.2 / 16.8
Bot emphasis means (Care, Fairness, Loyalty, Authority, Sanctity)	0.312, 0.490, 0.022, 0.080, 0.097	0.382, 0.455, 0.017, 0.111, 0.034
Bot primary foundations (%, Care / Fairness / Loyalty / Authority / Sanctity)	28.1 / 71.5 / 0.1 / 0.2 / 0.2	48.1 / 51.9 / 0.1 / 0.0 / 0.0
Engagement levels (%, 0 / 1 / 2 / 3)	5.8 / 18.0 / 60.5 / 15.7	4.9 / 23.0 / 67.6 / 4.5
Tone labels (%, hostile / neutral / warm)	2.5 / 73.4 / 24.1	0.5 / 96.0 / 3.4

E.4 Interpretation, Limitations, and Future Directions

Taken together, these analyses suggest that closer moral matching between GPT-4o and respondents is associated with meaningfully larger short-term gains—roughly five additional thermometer points when we move from no overlap to a perfect top-two match and about two points per 10-point increase in the weighted score. Those advantages disappear once we adjust for immediate outcomes and the broader covariate block, indicating that the descriptive persistence largely reflects baseline differences rather than lasting persuasion. Because alignment is measured post-treatment within the treated arm, the relationships remain correlational. Even so, the chatbot context provides rich visibility into conversational content, offering a scalable way to study mechanism pathways that are hard to observe in human canvassing campaigns. Future work could randomize which foundations the chatbot emphasizes (or suppresses), alignment prompts that cover binding foundations more fully, and examine field settings or hybrid human–AI workflows to identify the causal conditions that sustain attitudinal gains over time.

Appendix F. Ethical Considerations

While it is true that any persuasive technology, including everything from vignette framing experiments to AI-driven conversational systems, carries the theoretical risk of being used for harmful ends, we believe that our study is not only ethically defensible but also ethically necessary. If responsible researchers do not proactively develop, test, and openly document the use of AI for socially beneficial purposes—such

[6]It is worth noting that GPT-4.1-mini annotates engagement and tone somewhat more conservatively—high engagement responses (level 3) fall to about 11% versus 13% under Gemini, and 'warm' tone labels drop from roughly 24% to 20%. These shifts don't change the alignment regression coefficients, but are worth keeping in mind for any future work that leans heavily on these ancillary diagnostics.

as reducing prejudice and promoting inclusion—there is a strong likelihood that less scrupulous actors would still develop approaches to using AI for less beneficial purposes. They would likely do so without appropriate safeguards, transparency, or ethical oversight, or the counterweight of researchers pursuing more socially useful ends. Our work advances an understanding of how AI can be harnessed for prosocial outcomes, establishes normative and methodological benchmarks for ethical deployment, and highlights both the promise and limitations of AI-mediated persuasion. Rather than enabling misuse, our research contributes to building the knowledge base, standards, and institutional practices needed to ensure that future uses of this technology are aligned with democratic values and respect for human dignity. By conducting this work under institutional ethical review and after extensive ethical conversations among us and our peers, with clear transparency and an emphasis on beneficent applications, we help chart a responsible path forward for the inevitable expansion of AI's role in persuasive communication.

Please see the main text for additional thoughts related to the ethics of work like this.

This project was approved by the University of Virginia Internal Review Board (project #6915).

Figure legends

Figure S1. Distribution of Gemini alignment scores, engagement levels, and conversation tone (treated sample, $n = 1{,}315$).

References

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Bisbee, J. and Spirling, A. (2025). What to do when humans are no longer the gold standard: Large language models, state of the art, and robustness. Working paper.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Hong, G. (2010). Marginal mean weighting through stratification: Adjustment for selection bias in multilevel data. *Journal of Educational and Behavioral Statistics*, 35(5):499–531.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263.

Kalla, J. L. and Broockman, D. E. (2020). Reducing exclusionary attitudes through interpersonal conversation: Evidence from three field experiments. *American Political Science Review*, 114(2):410–425.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, 76(3):1071–1102.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Seaman, S. R. and White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3):278–295.

VanderWeele, T. J. and Ding, P. (2017). Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine*, 167(4):268–274.