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

Charles Crabtree[a,1], **John B. Holbein**[b, 2], **Mitchell Bosley**[c], **and Semra Sevi**[d]

This manuscript was compiled on March 1, 2026

Personalized interpersonal conversations are among the most effective known tools for reducing prejudice, yet they are difficult to scale because they require skilled human facilitators. This study tests whether artificial intelligence can approximate the effects of these interventions. Using OpenAI's GPT-4o, we developed a messaging-based intervention that engaged U.S. participants in individualized, morally aligned dialogues about transgender rights. In a preregistered experiment, these AI-mediated conversations significantly increased support for transgender rights across multiple attitudinal measures. Robustness checks, including weighting and sensitivity analyses, confirmed the reliability of these effects, and analyses of the conversations support the idea that moral matching between the AI and participants played a key role in reducing prejudice. However, follow-up data collected one week later indicated that the attitudinal gains diminished over time, suggesting that reinforcement may be necessary to sustain change. Together, these findings indicate that generative AI can facilitate value-aligned dialogue capable of shifting social attitudes, while highlighting the challenge of achieving durable impacts.

Artificial Intelligence (AI) | Prejudice Reduction | Transgender Rights | Moral Matching | Attitude Change

Significance Statement

Personalized, face-to-face conversations are among the most effective known interventions for reducing prejudice, yet they are difficult to scale. This study demonstrates that artificial intelligence can approximate some of the persuasive and moral dynamics of these interactions. Using large language models to conduct individualized dialogues and analyze their content, we find that AI-mediated conversations can increase support for transgender rights, although effects attenuated when respondents were recontacted a week later. These findings suggest that generative AI can be harnessed to promote empathy and reduce bias at scale, while also highlighting the need for complementary strategies to sustain long-term attitude change.

Reducing prejudice toward marginalized groups remains a central objective in the social sciences, with significant consequences for equity, democratic cohesion, and public policy (1). One approach that has shown remarkable effectiveness is *deep canvassing*: a method that uses structured, empathetic, one-on-one, in-person conversations designed to shift attitudes on contentious social and political issues. Drawing on insights from narrative persuasion and moral reframing (2–5), deep canvassing encourages participants to reflect on personal experiences and core values, often leading to durable attitude change. These conversations have demonstrated notable success in reducing transphobia and other forms of prejudice (e.g., 6–9). One particularly promising approach involves integrating deep canvassing with moral matching: the practice of tailoring persuasive messages to reflect the moral values of the recipient (2, 3, 10, 11). Despite evidence supporting the effectiveness of deep canvassing, this approach is costly and difficult to scale (12). It relies heavily on trained human facilitators and time-intensive interactions, which makes it impractical for large-scale deployment.

Recent advances in AI, particularly the development of large language models (LLMs) like OpenAI's GPT series offer a promising avenue for addressing at least some of these scalability challenges (12–15). These models can engage in human-like conversations, adapt messages to align with users' beliefs and values, and even influence opinion on morally charged topics (16–20). An emerging research stream has demonstrated that AI systems can reduce conspiracy beliefs through rational discourse (18, 21–23), shift issue positions (20, 24–27), and foster more respectful political dialogue (28).

Despite these encouraging early findings, key questions remain. Researchers still lack a comprehensive understanding of when, where, and for whom AI-driven persuasion is most effective (29–37). To our knowledge, no published studies have yet examined whether or the extent to which generative AI, like in-person interventions, reduce prejudice toward marginalized groups, such as sexual and gender minorities, through value-aligned conversational interventions. Addressing this gap is critical for assessing the real-world potential of AI as a scalable and cost-effective tool to promote personalized social change.

Our study investigates whether AI can harness moral matching to reduce prejudice in a cost-effective manner, focusing specifically on prejudice against transgender individuals. Despite growing visibility and some legal advances, transgender people continue to face widespread discrimination, violence, and social exclusion in many parts of the world (e.g., 38, 39). These experiences have far-reaching consequences for mental health, community safety, and civic participation (40), underscoring the urgent need for cost-effective interventions that foster greater acceptance and inclusion.

We examine three questions in this paper: (1) Can morally aligned AI conversations increase support for transgender rights?, (2) Does the effectiveness of these conversations vary depending on individuals' moral profiles and the moral matching of the conversations?, and (3) Are any observed attitude changes durable over time? To address these questions, we conducted a preregistered randomized controlled trial with a national non-probability sample adjusted to match key demographics of U.S. adults. Treated participants engaged in a brief, one-on-one conversation with OpenAI's GPT-4o, which was programmed to deliver persuasive messages tailored to their responses on the Moral Foundations Questionnaire (2, 41). Control participants engaged in no such conversation.

Author affiliations: [a]Monash University, School of Social Sciences, Melbourne, Australia and Korea University, University College, Seoul, South Korea; [b]University of Virginia, School of Leadership and Public Policy, Charlottesville, Virginia, USA; [c]Chilean National Center for Artificial Intelligence (CENIA), Santiago, Chile; [d]University of Toronto, Department of Political Science, Toronto, Ontario, Canada

We declare no competing interests.

[1]C.C., J.H., M.B., and S.S. contributed equally to this work. Author order was determined by random draw with `sample()` in R and a mutually agreed on seed.

[2]To whom correspondence should be addressed. E-mail: holbein@virginia.edu

Despite the brevity of the interaction (six exchanges), we find that participants exposed to the intervention show significant short-term increases in support for transgender rights. Effect sizes range from 0.09 to 0.29 standard deviations across a range of attitudinal measures. These effects are robust to alternative model specifications and weighting strategies. However, follow-up data collected one week later reveal substantial attenuation, suggesting that while AI can spark attitudinal shifts, sustained change may require reinforcement.

To better understand one potential mechanism behind these persuasive effects and examine the role that conversational content plays, we conducted an exploratory analysis of moral matching within the treated group. Leveraging novel text-as-data methods, we generated post-treatment annotations of our chatbot conversations, which we used to assess how closely AI's moral framing matched each participant's self-reported moral foundations. Using this measure, we show suggestive evidence that moral matching of conversational content is associated with stronger persuasive effects in the immediate aftermath of the conversation.

Our findings contribute to four key literatures. First, we advance research on prejudice reduction by showing that AI can emulate central features of effective in-person conversation, such as personalization, value matching, and empathetic engagement, in a cost-effective digital format. Second, we extend Moral Foundations Theory to the domain of machine-mediated persuasion, demonstrating that messages grounded in individuals' moral values can meaningfully influence attitudes relative to no conversation. Because we did not include a condition featuring non-morally aligned or generic messaging, we cannot isolate the specific effect of moral congruence from the broader impact of engaging in a persuasive dialogue. However, because our AI intervention generates a complete transcript for every conversation, we can systematically assess the degree of moral matching and its association with attitudinal outcomes—offering a unique analytic advantage over in-person deep canvassing studies, which typically lack or at least do not report systematic records of conversational content. Our findings that moral matching is associated with stronger persuasive effects in the immediate aftermath of the conversation raise important questions about the idea that interpersonal contact works independently of message content, and highlight the need for further research to isolate when, how, and for whom message matching shapes persuasive outcomes. Third, we contribute to research on digital interventions by identifying both the potential and limitations of AI systems in influencing opinions on socially contested issues. Finally, we contribute to the persuasion literature by testing how emerging technologies like generative AI intersect with core theoretical models—particularly regarding message-receiver matching and the durability of attitudinal change.

Together, our findings suggest that AI systems could serve as a cost-effective tool for delivering personalized, persuasive messages informed by moral psychology. At the same time, they underscore some important limitations, particularly regarding the durability of attitudinal change, and suggest that future work should explore hybrid approaches that combine the scalability of technology with the potential staying power of human-led reinforcement.

1. Background and Theoretical Framework

How can we reduce prejudice? One particularly promising approach to reducing prejudice is deep canvassing, in which trained individuals engage in open-ended, nonjudgmental conversations that invite reflection, encourage perspective-taking, and build interpersonal connection. Despite the effectiveness of this approach, however, there are difficulties scaling it. Deep canvassing, for example, requires intensive training, logistical coordination, and sustained time commitments, factors that limit its reach. As such, scholars and practitioners alike have begun exploring whether new technologies, particularly generative AI, might offer a cost-effective, personalized alternative.

This study draws on several literatures to examine whether large language models (LLMs) can emulate key mechanisms of effective interpersonal interventions: (1) political persuasion generally, (2) moral psychology, and (3) AI-driven persuasion. Together, these literatures provide a foundation for understanding how AI might be used to promote more inclusive attitudes in a more scalable and targeted manner.

A. Political Persuasion. At the heart of prejudice-reduction efforts more broadly, is persuasion: the intentional effort to influence another's beliefs, attitudes, or behaviors through communication (42). The study of persuasion has a long lineage, dating back to Aristotle's Rhetoric, and has evolved into a sprawling interdisciplinary field encompassing psychology, political science, communication, and philosophy. Contemporary theories of persuasion emphasize the interaction between source, message, receiver, and medium (43–45), with dual-process models like the Elaboration Likelihood Model (ELM) and the Heuristic-Systematic Model (HSM) explaining when and how individuals engage in thoughtful or cue-based processing of persuasive messages (46, 47). Yet despite these theoretical advances, the field remains fragmented and full of sometimes contradictory findings. While some studies suggest that political persuasion is difficult, others document meaningful, lasting change.

The recently proposed Generalizing Persuasion Framework (35, 36) offers a way to reconcile these tensions by encouraging scholars to systematically examine variation in actors, treatments, outcomes, and contexts. This framework clarifies when and why persuasion succeeds or fails, especially in complex political environments where strategic actors compete for public support. In this light, prejudice-reduction interventions can be seen as a specific case of political persuasion—one in which moral values, identity salience, and emotional resonance potentially play an important role.

While the normative implications of persuasion remain contested, especially when distinguishing education from manipulation, many scholars argue that persuasion aimed at promoting inclusion, tolerance, and evidence-based policy outcomes represents a legitimate and even necessary democratic good (e.g., 48–50). Thus, studying persuasive communication, especially in the service of reducing prejudice and promoting marginalized groups' rights, is not only empirically important but normatively justified (51).

B. Conversation-based strategies for prejudice reduction. A robust body of field experiments has demonstrated the power of *deep canvassing*, a structured, one-on-one conversation approach that emphasizes personal storytelling and active listening, in reducing prejudice. An earlier study (6) found that brief, doorstep conversations led by transgender and non-transgender canvassers significantly reduced transphobic attitudes, with effects lasting at least three months. Later studies confirmed that non-judgmental exchanges by canvassers on politically charged topics like immigration could reliably shift opinions (52).

C. Moral Foundations. One key mechanism behind these effects appears to be *moral matching*: tailoring messages to resonate with the listener's core values. Grounded in Moral Foundations Theory (MFT), this approach posits that people process social and political messages through five foundational moral domains, *Care*, *Fairness*, *Loyalty*, *Authority*, and *Purity*, which vary in importance across the ideological spectrum (2, 41). These foundations are weighted differently across the ideological spectrum. For example, progressives tend to emphasize Care and Fairness, while conservatives often prioritize Loyalty, Authority, and Purity.

Experimental work suggests that persuasive efforts can be tailored to reflect the recipient's moral lens. In one study, abortion rights messages framed in morally congruent terms increased support for pro-choice policies (11, 53). These findings highlight the value of personalization and moral resonance in successful persuasion efforts.

However, replicating this kind of moral tailoring in traditional canvassing campaigns is labor-intensive and difficult to execute consistently at scale. This challenge motivates our interest in whether artificial intelligence, specifically large language models, can emulate and automate moral matching techniques across diverse populations.

D. Capabilities of Large Language Models. Recent advances in LLMs, such as OpenAI's GPT-4o, have significantly expanded the potential for machines to hold rich, context-sensitive conversations. These models can simulate key elements of human dialogue, including semantic

nuance, emotional tone, narrative construction, and adaptive communication styles (18, 28). Crucially, they can personalize content in real time, enabling dynamic interactions based on users' stated beliefs, values, and conversational cues.

Initial evidence suggests that LLMs can successfully influence public attitudes when messages are carefully framed. AI-driven interventions have been shown to reduce belief in conspiracy theories, counter misinformation, and increase support for climate change action (16, 54). These findings raise the possibility that AI systems could be used not just to inform (27) but also to persuade, particularly if they are designed to target the same psychological mechanisms that make human-led interventions effective.

However, existing research has largely overlooked how LLMs might be used to reduce prejudice toward marginalized groups when guided by principles of moral matching. The intersection of moral psychology and AI-powered persuasion remains underexplored, particularly in the context of social inclusion and rights-based issues like transgender equality.

2. Research Questions

Research on moral reframing and political persuasion raises an important question: To what extent can the elements associated with successful, personalized prejudice reduction, such as value matching, tailored messaging, and emotionally resonant communication, be approximated in a scalable, cost-effective way? And might artificial intelligence (AI) offer one potential pathway for delivering such personalized persuasion?

To explore these questions, we develop a framework for AI-mediated persuasion informed by three elements emphasized in prior interpersonal interventions:

- Moral matching message content to recipients' dominant moral values;
- Personalization—adapting communication to reflect individual beliefs and concerns, and
- Perspective-taking—eliciting empathy by fostering emotional resonance and cognitive openness.

We evaluate this framework in a randomized experiment in which participants either engaged in a brief, morally aligned conversation with an AI chatbot or received no conversational intervention. This design allows us to assess the effects of value-informed AI persuasion relative to a pure control condition.

Specifically, we address the following research questions:

- Can morally aligned, AI-generated conversations increase support for transgender rights?
- Does the persuasive effect vary across individuals based on their moral profiles? For instance, are effects stronger among those who emphasize Care and Fairness?
- Are any observed attitudinal changes sustained over time, or do they attenuate after the conversation ends?

3. Experimental Design

To test these questions, we conducted a preregistered randomized controlled trial using a national non-probability sample adjusted to match the national demographics of U.S. adults (55).* This study was reviewed by the University of Virginia IRB (6915). This study was executed via CINT (formerly Lucid).† Our experiment was run on a large sample of 3,089 U.S. adults.‡ We collected data in November 2024.§ Demographic quotas on age, gender, race/ethnicity, and region were used to ensure that our sample represented the national population across key attributes. After removing 212 participants who dropped out before treatment assignment, we randomized 2,877 respondents into treatment (n = 1,430) and control (n = 1,447) conditions.¶

Figure 1 shows the flow of our experimental design. Before treatment assignment and exposure, participants completed demographic questions, the Moral Foundations Questionnaire (MFQ), and baseline measures of transgender rights support, including a feeling thermometer and policy attitude items. These responses served dual purposes: they provided covariates for the analysis and, for participants in the treatment group, the MFQ scores were used to personalize the AI interaction.

Figure 1 also provides a visual schematic of the overall treatment strategy, outlining the flow from assessment to intervention to outcome measurement. As depicted in Step 1 of Figure 1, the MFQ responses allowed us to identify the dominant moral foundation of each participant. Participants in our survey engaged in tailored, one-on-one conversations with GPT-4o, which used their responses to the Moral Foundations Questionnaire to frame persuasive messages in morally congruent terms.

Participants in the treatment condition engaged in a GPT-4o-powered chatbot conversation tailored to their MFQ profiles. Step 2 of Figure 1 illustrates this core manipulation, showing how the AI was designed to engage participants in personalized, morally aligned dialogues, with examples of persuasive frames tailored to different primary moral foundations (Care, Fairness, Loyalty, Authority, Purity). The chatbot was instructed to dynamically align persuasive messages with respondents' dominant moral values, emphasizing harm reduction, fairness, respect for authority, or human dignity as appropriate. Each conversation consisted of approximately six exchanges.

The control group received no treatment and immediately proceeded to the post-treatment questionnaire. We intentionally opted for a pure control rather than a placebo conversation because our goal was to estimate the more policy-relevant outcome: the effect of a morally aligned conversation relative to no treatment. As a result, the treatment condition, consisting of six back-and-forth message exchanges, required more time and cognitive engagement than the control condition. This likely contributed to higher attrition in the treatment group during Wave 1 (12.2%) than the control group (1.5%). (For further attrition diagnostics and tests of covariate balance, see Tables S2-S6 in the Supporting Information.) We account for these imbalances through a comprehensive battery of weighting and other robustness checks meant to adjust for attrition, all of which confirm the consistency of our results. These adjustments include a combination of inverse probability weighting, covariate adjustment, and doubly robust estimation methods (detailed in Appendix B.4 of the Supporting Information). Attrition between waves was substantial but similar across conditions: 56.0% in control and 57.9% in treatment group. The final sample sizes at Wave 2 were 627 for control group and 528 for the treatment group. We use the standard set of robustness checks to address potential bias from attrition.

Step 3 of Figure 1 lists the key outcomes collected immediately after the treatment (Wave 1) and again one week later (Wave 2), allowing us to assess both immediate effects and their persistence. These outcome measures mirror those in prior published work about the effects of person-to-person persuasion already cited. We measured support for transgender rights using four complementary outcomes:

1. **Feeling Thermometer:** A 0–100 scale measuring affect toward transgender people.

2. **Attitudinal Support:** Agreement with pro-transgender rights statements, averaged into a composite index.

*Our preregistration can be found here: https://aspredicted.org/rzwn-7pd9.pdf

†CINT (formerly Lucid) labels itself as "the world's largest global research marketplace for getting ... surveys answered." Research has generally shown that market research firms such as these can recruit samples that replicate canonical experimental results (e.g., 56–58), especially with standard attention checks, which we include.

‡We selected this sample size based on *a priori* power calculations in order to be powered at 0.80 to detect a 0.10 SD effect (two-tailed $\alpha = 0.05$), as well as to be well positioned for detecting heterogeneous effects.

§This timing was chosen intentionally, as transgender rights were particularly salient for many participants in the period leading up to the election.

¶We discuss later what dropping these respondents means for our analyses and results.

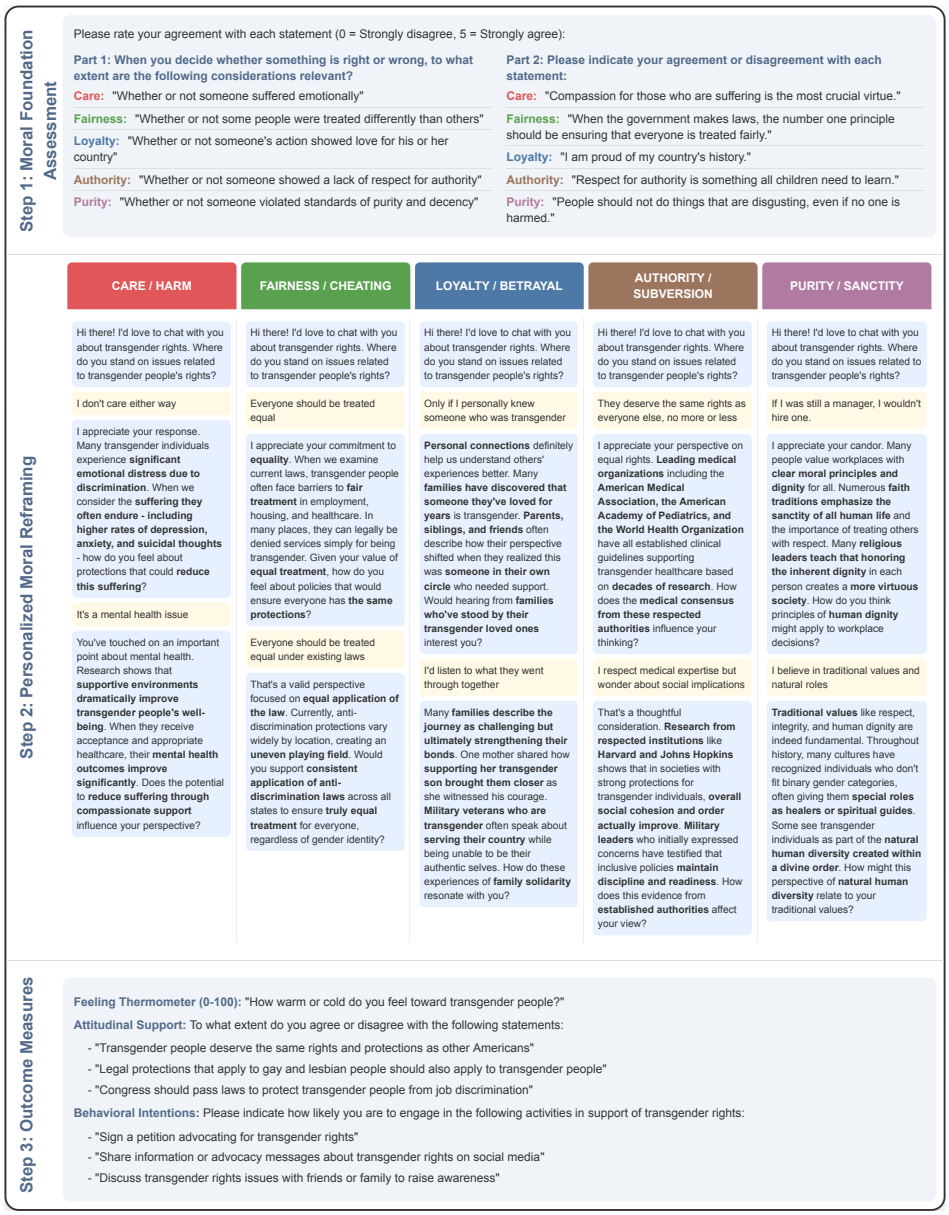

Fig. 1. Summary of Treatment Strategy

3. **Behavioral Intentions:** Willingness to take supportive actions (e.g., signing petitions), averaged into a composite index.

4. **Latent Support Index:** A principal component analysis-based index combining the above three measures.

This design allowed us to assess the causal effect of a morally aligned AI intervention on support for transgender rights and evaluate the durability of those effects over time.

A. Chatbot Design and Implementation. The chatbot was embedded in Qualtrics via custom JavaScript. After the MFQ, participant responses were processed into a moral profile used to guide the chatbot's persuasive framing. The GPT-4o model was prompted to deliver morally aligned arguments in favor of transgender rights, and asked to adapt to participant values and conversational responses in real time. The precise instructions given to the AI (the system prompt) are detailed in Appendix D in the Supporting Information. This prompt is critical as it operationalized our theoretical framework, instructing the AI to act as an expert persuader with the specific goal of increasing support for transgender rights. The prompt outlined a conversational arc inspired by deep canvassing, involving steps such as establishing contact, surfacing the participant's values, validating those values, building understanding through shared values or personal stories (simulated), engaging in personalized moral reframing based on the pre-assessed MFQ profile (e.g., linking fairness values to non-discrimination), and finally, reinforcing key points and suggesting action. Furthermore, the prompt provided guiding principles for the AI's tone (empathy, respect), strategy for handling disagreement (avoid argument, seek common ground), and constraints (e.g., brevity, 9th-grade reading level) to ensure a consistent and ethically grounded interaction. The conversation flow and tone were modeled on effective deep canvassing techniques, emphasizing empathy, common ground, and respectful engagement.

This implementation enabled delivery of personalized persuasion while simulating key features of effective interpersonal interventions. Figure 2 shows an example respondent conversation with the chatbot.

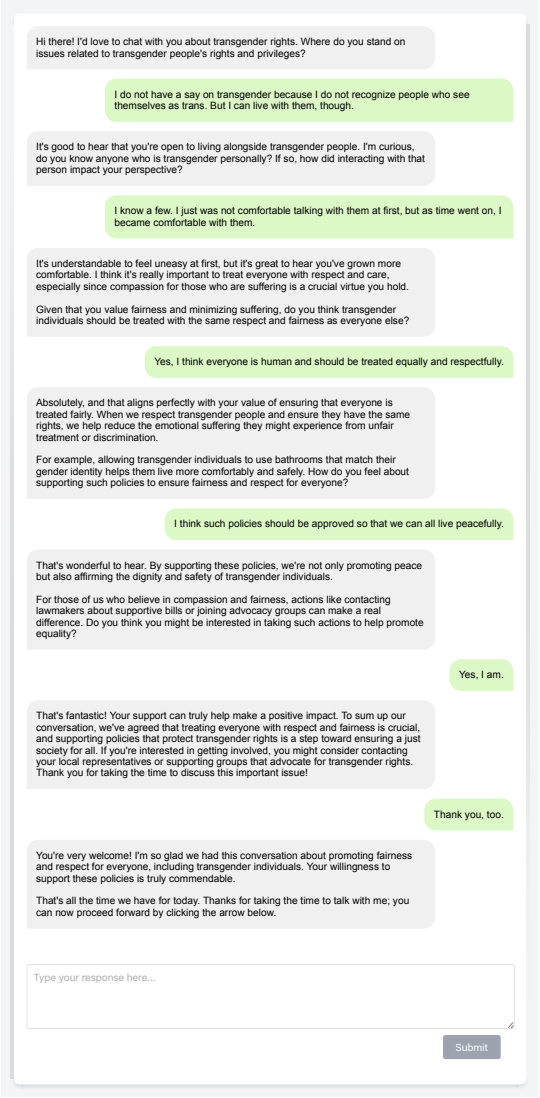

Fig. 2. Example of Chatbot Conversation with a Respondent

B. Estimation Strategy. As discussed above, we tailored our estimation strategy to address two potential threats to inference: differential attrition in the first wave, and non-random selection into the second wave. To estimate treatment effects, we began with unweighted difference-in-means, then adjusted for baseline feelings towards transgender individuals following (59).

To correct for attrition (see also Tables S2, S4, and S5) and ensure covariate balance (see also Tables S1, S3, and S6), we applied inverse probability weighting using Covariate Balancing Propensity Scores (CBPS; (60)) based on our pre-treatment covariates, including demographics, transgender feeling, and AI use. We constructed both treatment assignment weights and attrition-specific weights to account for different attrition scenarios.[‖] For Wave 2 outcomes, we implemented sequential IPW (61), multiplying initial assignment and retention weights.

We then estimated doubly robust models that combine covariate adjustment and weighting (62). To assess the robustness of our results to unobserved confounding, we computed E-values (63). Furthermore, to provide a particularly stringent, non-parametric check against differential attrition under worst-case assumptions, we calculated Lee bounds (64). In addition, we applied Benjamini-Hochberg corrections for multiple comparisons (65). Confidence intervals for all weighted models were derived via block bootstrapping with 1,000 replicates (66), and standard errors for all other models are robust to heteroskedasticity (67). To explore heterogeneity, we estimated Conditional Average Treatment Effects (CATEs) using interaction terms between treatment and moderators such as ideology and moral foundations (68). For a full description of methodological choices, see Appendix B.

4. Results

Figure 3 presents treatment effects across both waves, outcome measures, and model specifications. Tables 1 and 2 provide detailed statistics for Waves 1 and 2, respectively (see also Table S7 in the Supporting Information), while Tables S11 and S12 show heterogeneous treatment effects across moderator variables.

[‖] Specifically, we tested multiple model specifications for both the first and second waves. These included: (1) weights that adjusted the treatment group to more closely resemble the overall population to estimate the Average Treatment Effect (ATE); (2) weights that adjusted the control group to more closely resemble the treatment group to estimate the Average Treatment Effect on the Treated (ATT); and (3) weights that reduced the influence of treatment-group respondents who were similar to those who attrited.

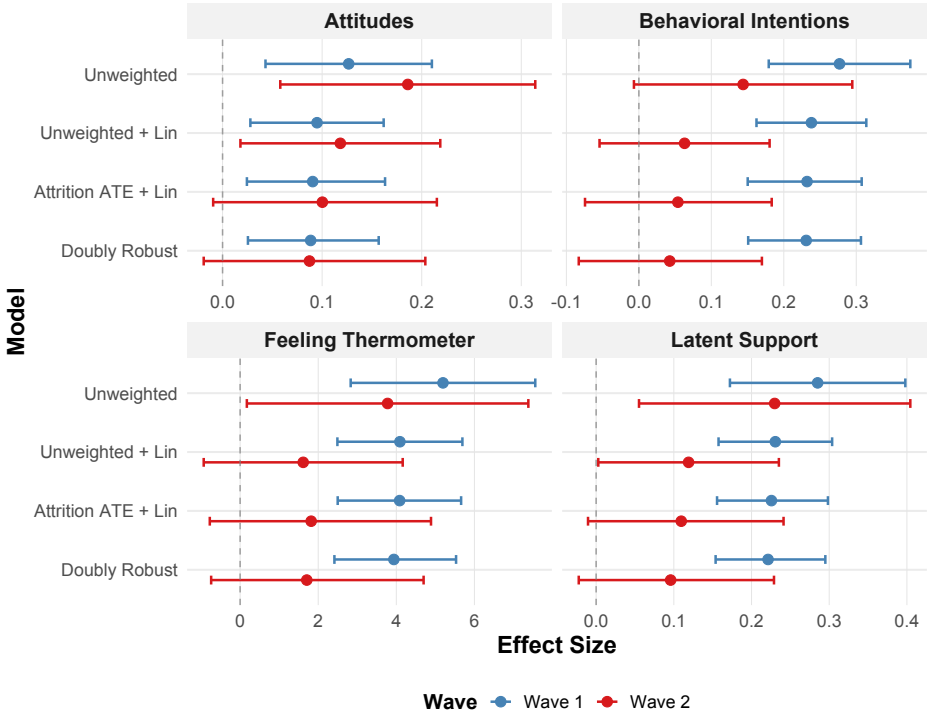

Fig. 3. Effects of Persuasive AI Treatment on Support for Transgender Issues, by Dependent Variable and Model Specification

Table 1. Treatment Effects on Transgender Rights Support (Wave 1)

	Estimate	SE	p-value[a]	p-value (BH)[a,b]	E-value[c]	95% CI
Attitudes						
Unweighted	0.127	0.043	0.003**	0.006**	1.46	[0.043,0.21]
Unweighted + Lin	0.095	0.034	0.006**	0.006**	1.38	[0.028,0.162]
ATE + Lin	0.088	0.035	0.006**	0.006**	1.36	[0.021,0.16]
Attrition ATE + Lin	0.091	0.035	0.004**	0.006**	1.37	[0.024,0.163]
Doubly Robust	0.088	0.033	0.000***	0.000***	1.36	[0.025,0.157]
Behavioral Intentions						
Unweighted	0.277	0.050	0.000***	0.000***	1.73	[0.179,0.375]
Unweighted + Lin	0.238	0.039	0.000***	0.000***	1.65	[0.162,0.314]
ATE + Lin	0.232	0.040	0.000***	0.000***	1.63	[0.15,0.308]
Attrition ATE + Lin	0.232	0.039	0.000***	0.000***	1.63	[0.15,0.307]
Doubly Robust	0.231	0.039	0.000***	0.000***	1.63	[0.151,0.306]
Feeling Thermometer						
Unweighted	5.194	1.206	0.000***	0.000***	1.60	[2.831,7.557]
Unweighted + Lin	4.091	0.817	0.000***	0.000***	1.50	[2.491,5.692]
ATE + Lin	3.981	0.796	0.000***	0.000***	1.49	[2.388,5.547]
Attrition ATE + Lin	4.084	0.794	0.000***	0.000***	1.50	[2.498,5.659]
Doubly Robust	3.936	0.795	0.000***	0.000***	1.49	[2.414,5.529]
Latent Support						
Unweighted	0.285	0.058	0.000***	0.000***	1.66	[0.172,0.398]
Unweighted + Lin	0.231	0.037	0.000***	0.000***	1.57	[0.158,0.304]
ATE + Lin	0.223	0.038	0.000***	0.000***	1.55	[0.151,0.297]
Attrition ATE + Lin	0.226	0.038	0.000***	0.000***	1.56	[0.156,0.298]
Doubly Robust	0.221	0.037	0.000***	0.000***	1.55	[0.154,0.295]

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

[b] BH: Benjamini-Hochberg correction applied within each outcome

[c] E-values indicate robustness to unmeasured confounding; higher values = greater robustness

Note: N = 2,681 for all models except Doubly Robust (N = 2,679)

Table 2. Treatment Effects on Transgender Rights Support (Wave 2)

	Estimate	SE	p-value[a]	p-value (BH)[a,b]	E-value[c]	95% CI
Attitudes						
Unweighted	0.186	0.065	0.004**	0.022*	1.60	[0.058,0.314]
Unweighted + Lin	0.118	0.051	0.021*	0.053	1.44	[0.018,0.219]
Seq. ATE + Lin	0.091	0.057	0.114	0.116	1.36	[-0.019,0.207]
Seq. Attrition ATE + Lin	0.100	0.057	0.076	0.116	1.39	[-0.009,0.215]
Seq. Doubly Robust	0.087	0.055	0.116	0.116	1.36	[-0.019,0.204]
Behavioral Intentions						
Unweighted	0.144	0.077	0.062	0.308	1.45	[-0.007,0.295]
Unweighted + Lin	0.063	0.060	0.292	0.490	1.26	[-0.054,0.18]
Seq. ATE + Lin	0.052	0.067	0.422	0.490	1.23	[-0.078,0.184]
Seq. Attrition ATE + Lin	0.054	0.066	0.396	0.490	1.24	[-0.074,0.183]
Seq. Doubly Robust	0.043	0.064	0.490	0.490	1.21	[-0.083,0.17]
Feeling Thermometer						
Unweighted	3.776	1.839	0.040*	0.201	1.48	[0.172,7.38]
Unweighted + Lin	1.615	1.300	0.214	0.214	1.27	[-0.932,4.163]
Seq. ATE + Lin	1.832	1.467	0.198	0.214	1.30	[-0.688,4.821]
Seq. Attrition ATE + Lin	1.819	1.476	0.196	0.214	1.29	[-0.777,4.888]
Seq. Doubly Robust	1.707	1.433	0.214	0.214	1.28	[-0.742,4.698]
Latent Support						
Unweighted	0.230	0.089	0.010*	0.050	1.56	[0.055,0.404]
Unweighted + Lin	0.119	0.059	0.045*	0.106	1.36	[0.003,0.235]
Seq. ATE + Lin	0.104	0.065	0.098	0.106	1.33	[-0.017,0.236]
Seq. Attrition ATE + Lin	0.110	0.065	0.078	0.106	1.34	[-0.01,0.241]
Seq. Doubly Robust	0.096	0.063	0.106	0.106	1.31	[-0.022,0.229]

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

[b] BH: Benjamini-Hochberg correction applied within each outcome

[c] E-values indicate robustness to unmeasured confounding; higher values suggest greater robustness

Note: N = 1,155 for all models

A. Wave 1 Treatment Effects. As shown in Table 1, the treatment effects in Wave 1 are positive, substantial, and statistically significant across all outcome measures and the vast majority of modeling approaches. For the feeling thermometer, the AI intervention produced increases ranging from 3.94 to 5.19 points (all $p < 0.001$), with E-values between 1.49 and 1.60 indicating reasonable robustness to unmeasured confounding.** For behavioral intentions, effects ranged from 0.23 to 0.28 standard deviations (all $p < 0.001$, E-values: 1.63-1.73). For attitudes, effects ranged from 0.09 to 0.13 standard deviations (all $p < 0.01$, E-values: 1.36-1.46), and for the latent support index, from 0.22 to 0.29 standard deviations (all $p < 0.001$, E-values: 1.55-1.66). (See also table S9 in the Supporting Information.)

The stability of these results across specifications using weighting and covariate adjustment is notable. Even our most conservative weighted approaches (see Appendix B, Table S7) yield statistically significant effects with 95% confidence intervals that exclude zero after Benjamini-Hochberg correction. This suggests strong robustness against observable confounding and selection bias. To provide an even more punitive assessment under worst-case assumptions, we also employed a non-parametric Lee bounds analysis (Appendix B, Table S10), which examines the potential impact of differential attrition under the most unfavorable scenarios. Essentially, this method provides very conservative bounds by estimating the range of possible treatment effects if the 'excess' participants who completed the survey in the higher-completion group (here, control) had the most extreme outcomes, thereby testing the result against the most challenging attrition bias.

This test revealed that the lower bound for Behavioral Intentions remains positive (0.07), demonstrating resilience even under these highly conservative assumptions. The lower bounds for Feeling Thermometer (-2.01), Attitudes (-0.14), and Latent Support (-0.05) cross zero, indicating that while these effects are robust to standard adjustments, null effects cannot be entirely ruled out if differential attrition operated in the maximally unfavorable way assumed by this specific method. However, considering the Lee bounds represent an extreme scenario and the consistent positive findings across multiple other adjustment methods (IPW, DR) and sensitivity analyses (E-values), the overall evidence supports moderate robustness for the Wave 1 effects, with Behavioral Intentions demonstrating particularly strong resilience across all checks.

B. Wave 2 Treatment Effects. In contrast, Wave 2 results (Table 2) demonstrate considerably less robustness, with substantial attenuation and high sensitivity to model specification, suggesting minimal evidence for persistent effects. While the directional pattern remains consistent with Wave 1 for some outcomes, the magnitude and statistical significance diminish markedly, particularly after accounting for attrition and multiple comparisons. For the feeling thermometer, only the unweighted, unadjusted model yields a statistically significant effect (3.78 points, $p < 0.05$), but this significance disappears after correcting for multiple comparisons or applying weighting adjustments. All other specifications produce smaller, non-significant estimates with E-values for the confidence interval lower bound at or near 1.0 (Appendix B, Table S8), indicating high vulnerability to even modest unmeasured confounding.

** The E-value is the minimum strength of association that an unmeasured confounder would need to have simultaneously with treatment assignment and with the outcome (on the risk-ratio scale) to explain away the observed effect (63). As a benchmark, the strongest measured pre-treatment covariate in our data—the baseline transgender feeling-thermometer score—corresponds to an E-value of 1.28, whereas our Wave 1 treatment effects range from 1.36 to 1.73. Hence, an unobserved variable would have to be more strongly related to both treatment and outcome than respondents' prior affect toward transgender people in order to fully nullify the findings.

Similar patterns emerge for behavioral intentions and latent support: none of the weighted specifications yield statistically significant results after correcting for multiple comparisons. For attitudes, although some specifications retain nominal significance, the effects attenuate considerably in the more rigorous sequential weighting models that account for cumulative attrition across both waves. These sequential weighting models (Appendix B, Table S7), representing our most robust approach to selection bias, yield non-significant estimates for most Wave 2 outcomes.

This pattern is further confirmed by our compound Lee bounds analysis (Appendix B, Table S10), which accounts for the total differential attrition across both waves under worst-case assumptions. The lower bounds for all four Wave 2 outcomes cross zero by substantial margins (e.g., -5.84 for Feeling Thermometer). This stringent test reinforces the conclusion from other analyses (IPW, E-values, BH-corrections) that the apparent treatment effects in Wave 2 are not robust and cannot be reliably distinguished from zero when accounting for potential selection bias.

This contrast—moderately robust effects across specifications in Wave 1 but weak and highly sensitive results in Wave 2—suggests two important conclusions. First, the immediate impact of the AI-powered intervention is substantively meaningful and statistically reliable against a range of standard assumptions about selection and confounding. Second, there is substantially less evidence for the durability of these effects over one week.

C. Heterogeneous Treatment Effects. Tables S11 and S12 in Appendix C 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 our effects.

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

D. Personalization and Moral Matching Patterns. The case for moral foundation matching as a persuasion mechanism rests on the extent to which the chatbot's moral framing actually tracked respondents' priorities. In this section, we describe how the chatbot framed conversations relative to respondents' moral foundations and we ask whether the realized degree of matching is associated with within-respondent changes in warmth to trans individuals among treated participants. (See also Appendix E and Table S13 specifically.) We show two main results: matching varied across conversations, although the chatbot mostly framed its messages in fairness and care even when respondents prioritized other foundations; and matching is linked to larger immediate gains for treated respondents after accounting for MFQ, engagement, ideology, and demographics.

We begin by characterizing what personalization looked like in practice. A distinctive feature of our setting (compared to in-person conversations) is that every exchange is transcribed, allowing us to assess how the chatbot framed conversations and how closely that framing tracked respondents' moral foundations. Because personalization is bundled with treatment, our analysis here is descriptive rather than causal; our aim is to document the realized pattern of moral framing that any mechanism account must build on. We annotated each exchange with Google's Gemini 2.5 Flash, a model independent from the GPT-4o model that generated the conversations[††]—to score the chatbot's emphasis on each Moral Foundations axis and to label respondent engagement and tone (four-point intensity scale plus sentiment categories). Respondent MFQ profiles were normalized to the same 0–1 scale, enabling a weighted matching score and a "top-two" indicator that records whether the chatbot conversation's two dominant foundations matched the respondent's own moral priorities.[‡‡] We repeated the annotation with OpenAI's GPT-4.1-mini and obtained nearly identical distributions, consistent with benchmarks showing that frontier LLM annotators approach human-expert reliability (69). Full implementation details—including prompts, preprocessing, and diagnostics—appear in Appendix E.

Figure 4 makes clear that the chatbot centered its messaging on the "individualizing" foundations. (See also Figure S1 in the Supporting Information.) Gemini's annotations code roughly nine in ten treated chats as assigning the highest weight to the Fairness foundation, with a more modest but consistent emphasis of the Care dimension, whereas respondents' own MFQ weights are more dispersed. To quantify how strongly the chatbot responded to variation in respondent MFQ profiles, we regress each normalized chatbot foundation weight on respondents' MFQ scores. Table 3 shows that while in absolute terms the chatbot tended to employ Fairness and Care frames, adaptation did occur at the margins: respondents who scored higher on Loyalty, Authority, or Sanctity received correspondingly higher chatbot emphasis on those domains, signaling that GPT-4o registered users' binding priorities even as Fairness and Care remained its dominant default.[§§]

Figure 5 summarizes how this observed variation translates into our matching metrics. The weighted matching score is centered around 69 points on the 0–100 scale, with quartiles at roughly 60, 70, and 80 and a left tail capturing the samples where the chatbot missed the respondent entirely.[¶¶] The overlap categories tell a similar story: GPT-4o employed both of a respondent's top-two foundations in 36% of chats, one of the two in 48%, and neither in 16%. Together with Table 3, these descriptive results show that while GPT-4o applied moral foundations language broadly, it did so primarily through the Fairness and Care dimensions.

We then use this observed variation to relate matching to within-respondent changes among the treated arm. Table 4 reports ANCOVA models that mirror the covariate adjustment used in our main treatment-effect specifications: baseline warmth, respondents' MFQ Fairness and Harm scores, ideological placement, the engagement index, and the complete set of demographic and chatbot-use controls (age, sex,

[††] We selected Gemini 2.5 Flash because it offers strong classification accuracy at low cost and, crucially, provides an audit that is independent of the OpenAI's LLM offerings, including the GPT-4o conversation agent that we used.

[‡‡] The weighted score is the dot product of the respondent and bot vectors (scaled 0–100); the top-two indicator equals 1 when the two top foundations across the chatbot and respondent match, 0.5 when they overlap on exactly one element, and 0 otherwise.

[§§] The reasons why these foundations are more prevalent deserve further investigation. One possible explanation is that the task that we are giving the bot itself lends itself more to a Care and Fairness framing and/or LLM guiderails emphasize values or restraints in line with these foundations.

[¶¶] Appendix Table S14 reports full distributional statistics for both matching metrics and Table S15 reports mean moral-foundation weights (normalised) by respondent.

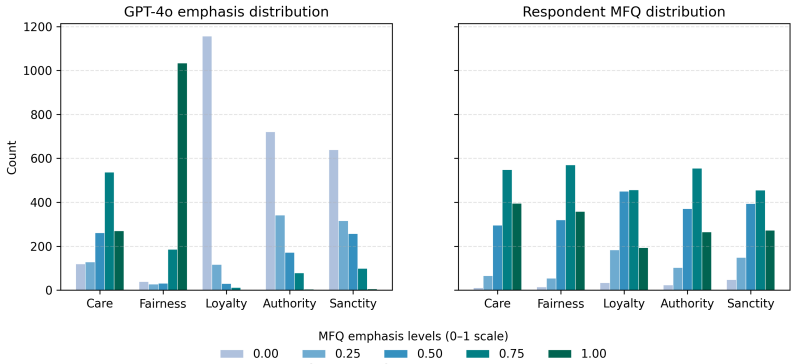

Fig. 4. Chatbot and respondent moral emphasis. The left panel displays GPT-4o's Gemini-coded emphasis levels across foundations; the right panel shows respondents' MFQ weights binned on the same 0–1 scale (treated sample, $n = 1{,}315$). See also Tables S14 and S15.

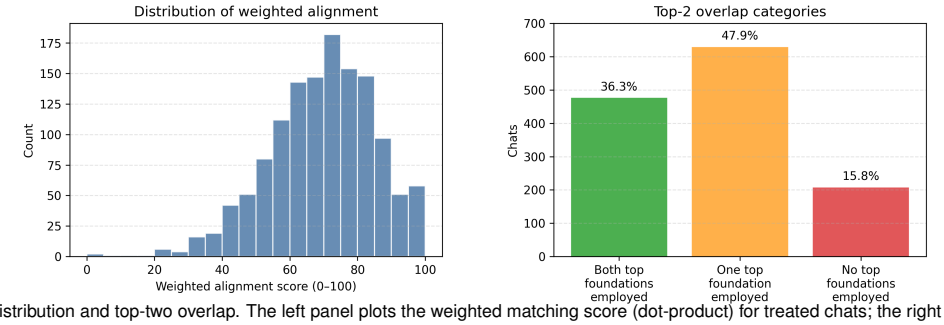

Fig. 5. Matching score distribution and top-two overlap. The left panel plots the weighted matching score (dot-product) for treated chats; the right panel reports the share of conversations where GPT-4o employed both, one, or none of the respondent's top-two MFQ foundations.

gender, political attention, ChatGPT familiarity, and perceived accuracy). (See Table S16 for ANCOVA regressions linking alignment measures to warmth outcomes among treated respondents only.) With these covariates in place, both matching metrics remain strongly associated with the immediate Wave 1 thermometer: moving from zero to a perfect top-two match corresponds to roughly +4.8 (s.e. 1.9) points and each 10-point increase in the weighted score adds about +2.0 (s.e. 0.8) points. These coefficients are estimated alongside respondents' MFQ profiles, baseline warmth, ideology, engagement, and the demographic/chatbot-use block, indicating that matching captures additional explanatory power beyond those predictors. At follow-up, however, neither matching coefficient is distinguishable from zero once we condition on Wave 1 warmth and the same covariate block: 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 does not predict additional week-later gains. Because matching is observed rather than randomized and the regressions control for realized post-treatment outcomes, we interpret the coefficients descriptively; even so, they show that realized personalization and engagement correlate with stronger short-term gains for treated respondents, offering suggestive evidence that moral matching matters for convincing people to support transgender rights.

5. Discussion

Our study demonstrates that an AI-driven intervention using OpenAI's GPT-4o, instructed to tailor conversations to respondents' moral beliefs, significantly increased short-term support for transgender rights relative to a pure control. These findings show that generative AI can deliver value-aligned persuasive messages with measurable attitudinal effects—offering a scalable, low-cost alternative to resource-intensive, human-led interventions such as deep canvassing (6, 12). The results from the follow-up data that we collected one week later revealed notable attenuation, raising questions about durability. To better understand the extent to which personalized moral re-framing mattered, we also conducted an exploratory matching analysis with data collected from respondents who completed the treatment arm, which revealed that GPT-4o's persuasive efforts strongly emphasized the Fairness and Care MFQ dimensions. We leveraged this variation in messaging matching to show suggestive evidence that closer moral matching is associated with larger immediate gains, above and beyond respondents' MFQ profiles, engagement, ideology, and demographics.

Our findings should be interpreted alongside recent studies that find limited evidence of persuasive advantage for AI microtargeting or personalization relative to generic messaging (e.g., 26, 70). Our design does not directly compare AI-mediated persuasion to human-delivered or non-customized messages, and we therefore refrain from making claims about this comparison. Instead, we demonstrate that AI can reproduce some of the mechanisms theorized to underlie effective interpersonal persuasion—personalization, moral matching, and empathetic engagement—and we evaluate whether those features can produce measurable attitudinal change relative to no engagement at all. In this sense, our findings speak to AI's capabilities and limitations, rather than its superiority, underscoring the need for future research that systematically benchmarks AI persuasion against human and generic communication.

A. Theoretical Implications. Our main theoretical contribution is to demonstrate that a brief, moral-foundations–framed exchange with an AI agent can shift attitudes in the short run relative to a pure control. This shows that value-laden message content (relative to a pure control) matters for moving views on a contested issue, and that the persuader need not be human for such content to be effective.

Our findings also contribute directly to Moral Foundations Theory, suggesting that value-congruent moral framing can shift attitudes even when the messenger is an AI agent. Within the treated arm, we show evidence that larger immediate gains are associated with closer message–person fit—i.e., greater matching between the chatbot's framing and respondents' dominant foundations—above and beyond MFQ, baseline warmth, engagement, ideology, and demographics. These results suggest that moral reframing is operative in AI-mediated persuasion and that the magnitude of short-term change increases with matching, even as durable changes remain limited.

That said, while our design focuses on morally aligned messaging, it does not include a non–morally aligned or generic messaging condition. As such, our findings speak to the effects of value-congruent communication but do not directly disentangle moral congruence

Table 3. Chatbot foundation weights regressed on respondent MFQ scores.

Outcome	Predictor	Coef.	SE	Sig.
Bot Care weight	Intercept	0.269	0.017	***
	Respondent Care	0.239	0.037	***
	Respondent Fairness	-0.122	0.038	***
	Respondent Loyalty	-0.069	0.031	**
	Respondent Authority	-0.022	0.037	
	Respondent Sanctity	0.023	0.029	
Bot Fairness weight	Intercept	0.452	0.022	***
	Respondent Care	-0.157	0.049	***
	Respondent Fairness	0.257	0.051	***
	Respondent Loyalty	0.075	0.040	*
	Respondent Authority	-0.041	0.049	
	Respondent Sanctity	-0.078	0.038	**
Bot Loyalty weight	Intercept	0.043	0.007	***
	Respondent Care	-0.044	0.016	***
	Respondent Fairness	-0.008	0.017	
	Respondent Loyalty	0.063	0.013	***
	Respondent Authority	-0.036	0.016	**
	Respondent Sanctity	0.001	0.013	
Bot Authority weight	Intercept	0.108	0.012	***
	Respondent Care	-0.000	0.027	
	Respondent Fairness	-0.074	0.028	***
	Respondent Loyalty	-0.031	0.022	
	Respondent Authority	0.075	0.027	***
	Respondent Sanctity	-0.012	0.021	
Bot Sanctity weight	Intercept	0.127	0.013	***
	Respondent Care	-0.038	0.028	
	Respondent Fairness	-0.054	0.029	*
	Respondent Loyalty	-0.038	0.023	
	Respondent Authority	0.023	0.028	
	Respondent Sanctity	0.066	0.022	***

Notes: Stars indicate significance at $p < 0.10$ (*), $p < 0.05$ (**), $p < 0.01$ (***). Each outcome is the chatbot's average normalized weight on the indicated foundation (treated chats, $n = 1{,}315$). Predictors are respondents' MFQ scores (0–1) with HC2 standard errors.

as the sole underlying mechanism. Future research can build on this work by explicitly testing whether value matching itself, rather than other features of the conversation, drives these effects, extending the broader understanding of how moral framing shapes persuasive communication (2).

Our results also highlight a potential temporal limitation of AI conversations aimed at prejudice reduction rather than factual correction. Although short-term effects were statistically robust across multiple outcomes, they substantially attenuated after one week. This pattern raises questions about the durability of attitude change following brief, AI-mediated interactions focused on moral foundations and suggests that deeper cognitive or emotional engagement, often linked to enduring persuasion, may be harder to achieve through non-human agents. These findings underscore the need for future research on the depth of processing involved in human–AI persuasion and its implications for the stability of attitude change over time.

At the same time, we acknowledge that beliefs and values do not form or persist in isolation. While our study isolates individual-level attitudinal change, the broader social context—norms, networks, and media environments—likely shapes whether and how such change endures. Structural and community-level factors can either reinforce or erode individual persuasion effects, suggesting that AI-mediated interventions should be studied across social ecosystems. We also stress that the intervention we test is not the only way to address pervasive anti-transgender beliefs. Instead, our findings should be seen as identifying one micro-level mechanism that could complement community-based and structural approaches aimed at transforming the broader sociocultural environment in which attitudes are embedded.

Our analysis of heterogeneous treatment effects offers additional insight into how individuals process AI as a persuasive source. Participants who prioritize "binding" moral foundations (Purity, Authority, Ingroup Loyalty) were less responsive to the intervention, consistent with (2)'s account of moral resistance to progressive messaging. This suggests that some of the same psychological mechanisms that constrain human-delivered persuasion may similarly limit the reach of AI-driven messages. At the same time, the broader pattern of results, most interaction effects were near zero, indicates that the overall persuasive impact was not driven by any particular subgroup. Instead, effects appear diffuse across moral foundations, ideological orientations, and demographic lines. This finding aligns with (51)'s argument that persuasive messages tend to produce remarkably homogeneous effects across social and political groups. Our data extend this insight to the digital realm, showing that even value-aligned AI interlocutors can elicit short-term attitude change that is broadly distributed rather than concentrated within specific moral or social segments.

Together, these findings advance our understanding of both prejudice reduction and AI persuasion. While previous studies have shown that LLMs can influence beliefs about conspiracy theories (e.g., 18) or political topics (e.g., 16), our study extends this to a domain traditionally rooted in human empathy and interpersonal conversation. While morally-focused AI can reduce prejudice in the short term (relative to a pure control), the inconsistent durability of these effects raises critical questions about the "dose-effect relationship"—the

Table 4. ANCOVA regressions linking matching measures to warmth outcomes (treated respondents only).

	Weighted matching (per 10 pts)		Top-two matching (0–1)	
	Wave 1 post	Wave 2 post	Wave 1 post	Wave 2 post
Matching coefficient	1.95^{**}	-1.08	4.83^{**}	0.79
	(0.77)	(1.09)	(1.93)	(2.74)
Baseline warmth (q_{24})	0.66^{***}	0.34^{***}	0.66^{***}	0.34^{***}
	(0.03)	(0.06)	(0.03)	(0.06)
Wave 1 warmth	–	0.43^{***}	–	0.43^{***}
		(0.06)		(0.06)
Fairness MFQ	-0.86	-0.46	0.77	-1.37
	(1.26)	(1.83)	(1.08)	(1.58)
Harm MFQ	-0.61	3.59^{**}	0.90	2.85^{*}
	(1.19)	(1.48)	(1.05)	(1.41)
Engagement level	2.75^{***}	-0.60	2.91^{***}	-0.68
	(0.95)	(1.33)	(0.95)	(1.32)
Observations	1,254	528	1,254	528
R^2	0.552	0.657	0.552	0.656

Notes: Ordinary least squares with HC2 robust standard errors. All models use data from only those who completed the treated arm and include the full set of baseline covariates from the main outcome analyses: age, sex, gender, political attention, ideological placement, ChatGPT familiarity, and perceived chatbot accuracy (categorical indicators omitted for brevity). Engagement level is coded 0–3. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

amount, frequency, and intensity of AI interactions needed to produce lasting attitudinal change. Future work should explore how AI persuasion operates within sociocultural contexts, where collective norms and group identities may amplify or constrain the effects we observe at the individual level.

B. Broader Lessons. Our findings offer practical and conceptual lessons for the design of AI-based interventions. One key takeaway is that engagement quality and receptivity may be central to creating scalable solutions. The fact that some participants dropped out of the AI conversation suggests that even brief persuasive interactions can be cognitively or emotionally taxing when they address contested topics. Future interventions should consider friction points in the user experience, such as the potential discomfort of moral confrontation or the novelty of interacting with an AI persuasive agent. While our weighting procedures and robustness checks mitigate this concern in estimating the treatment effects, the attrition pattern itself is informative.

Our use of a pure control condition establishes a baseline effect size for AI-mediated persuasion. However, future research should more precisely isolate the role of moral matching by comparing value-aligned conversations with generic or value-incongruent persuasive messaging. Incorporating placebo controls, such as non-aligned AI conversations or non-personalized persuasive content, would help disentangle the unique contribution of moral matching from the broader effects of engaging in a persuasive exchange.

Despite instructing the model to personalize its messaging to survey respondents, the bot did return to Fairness (and, to a lesser extent, Care) even when respondents prioritized authority, loyalty, or sanctity. Several factors could plausibly explain this pattern: safety and RLHF guardrails that privilege fairness/harm language; anchoring from the ordering and phrasing of examples in the prompt; domain priors that make fairness a safe default when the model is uncertain in this topic area; and finally, the inherent nature of the task that we asked the bot to engage in. Whatever the mix of causes, the concentration is informative: it clarifies how current systems implement personalization by default and helps bound what we should expect without additional controls or prompt engineering.

Methodologically, future work should work to further understand these patterns. Possible avenues include pre-testing chatbot systems on synthetic MFQ profiles and automatically monitoring foundation coverage and fit, as well as including examples of both good and bad patterns of argument personalization and, where feasible, applying light fine-tuning on a balanced, foundation-coded corpus. These steps should make personalization even more faithful while preserving safety constraints. These may provide a practical path for steering models toward controlled, foundation-aware framing in future studies.

Finally, our study challenges the assumption that AI is inherently neutral or disengaging as a communicator. Instead, GPT-4o successfully elicited empathetic, self-reflective responses, even on a highly politicized topic. This suggests that AI can be an effective persuasive actor, with its moral framing, emotional tone, and responsiveness shaping human attitudes in subtle and powerful ways.

C. Practical Applications and Cost-Effectiveness. AI-driven interventions offer significant cost advantages in prejudice reduction efforts. Unlike traditional deep canvassing, which requires extensive training and one-on-one conversations, our AI approach can engage thousands of individuals simultaneously while maintaining consistent quality. To illustrate the cost benefit: deep canvassing typically requires 10-15 minutes of training per volunteer hour and about 10 minutes of conversation per participant (6), resulting in substantial resource investments. In contrast, our AI system required only initial prompt engineering and could process thousands of conversations concurrently, with costs primarily limited to API usage (around $0.15 per conversation).

This cost-efficiency could dramatically expand the reach of advocacy organizations working with limited resources. Political campaigns supporting transgender-inclusive candidates could deploy similar chatbots to engage undecided voters at scale, reaching audiences that would be difficult to reach through traditional canvassing. Educational institutions such as universities and schools could implement AI-driven modules to complement existing diversity training, providing personalized engagement on issues of gender identity and inclusion. Advocacy organizations like the National Center for Transgender Equality could integrate similar chatbots into their websites, offering visitors personalized, values-aligned conversations about transgender rights. Healthcare organizations could use this approach to address stigma and misinformation about transgender healthcare, potentially improving health outcomes through attitude change.

However, the diminished effects seen in Wave 2 suggests that AI alone may be enough to produce durable attitude change. Organizations should consider supplementary strategies, such as periodic "booster" interactions or hybrid approaches combining AI with human follow-up, to sustain initial gains.

D. Future Work. Our findings point to several promising directions for future research. Methodologically, studies that vary the frequency, duration, and intensity of AI interactions could help clarify the optimal "dose" for durable attitude change. These might include experimental designs comparing single extended conversations with multiple shorter interactions over time.

From a theoretical perspective, further research on which moral foundations are most susceptible to AI persuasion and whether tailoring strategies can be optimized for individuals with specific moral profiles would refine the model we propose. Studies on how the attribution of persuasive intent differs between human and AI sources could explain the durability patterns we observed. Additionally, research on the depth of cognitive processing triggered by AI versus human persuasion could shed light on whether AI's ephemeral effects stem from shallower processing, potentially informing future interventions designed to deepen engagement.

Practically, testing models that combine AI's scalability with the durability of human connection would be valuable. For example, AI could be used for initial outreach and personalized framing, followed by human facilitators who build on the established rapport in follow-up conversations. Another approach might integrate AI into human canvassing, providing real-time suggestions for moral framing based on ongoing conversation content. Testing whether our approach can be applied to other prejudice domains would help assess the broader applicability of our moral matching approach.

Additionally, future research should examine the perceived credibility and trustworthiness of AI agents across political and demographic subgroups. Understanding why certain groups find AI persuasive can help refine communication strategies and guide the ethical use of AI in sensitive areas.

Finally, we hope that our work will encourage followup studies that continue to try to evaluate whether personalization matters for persuasion. Two recent experiments (26, 70) conclude that AI microtargeting offers little advantage over generic GPT-generated messages. We take that null benchmark seriously, but our analysis points in a different—albeit suggestive—direction. In our multi-turn chat setting, conversations that achieve high moral matching generate substantially larger immediate gains than those that miss respondents' foundations. Because matching is not randomized, we cannot treat this as definitive causal evidence. Still, the pattern is consistent across annotators and outcomes, indicating that personalization may matter when it is deeply woven into a responsive dialogue rather than applied as a light touch. We hope this evidence motivates further experiments that directly randomize matching intensity or compare aligned vs. generic bot conversations head-to-head.

E. Technological Developments and Future Possibilities. Recent advancements in large language models offer promising avenues to address the limitations we observed, particularly in durability. Three technological developments merit specific attention. New research (71) has demonstrated that increased inference-time computation can substantially enhance model performance on complex reasoning tasks. Applied to persuasion, this approach could enable more sophisticated adaptation to individual values and resistance patterns, potentially deepening engagement and strengthening attitude change. Specifically, models using techniques like tree-of-thought reasoning might produce more cognitively engaging arguments that result in deeper processing and more durable change (72).

Newer language models with multimodal capabilities (processing both text and visual information) could enhance persuasive impact by integrating emotionally resonant imagery with text-based moral arguments. This approach might address the emotional engagement gap that potentially contributes to the limited durability we observed.

Models with expanded context windows capable of processing thousands of previous interaction tokens enable richer "memory" of earlier conversations. This capability could support longitudinal engagement where the model recalls personal details and previous discussions, potentially building stronger rapport and facilitating more durable persuasion through perceived relationship continuity. The work by (73) on using AI for long-form, qualitative interviews with detailed personal context suggests promising avenues for deeper, more personalized persuasive interactions that might enhance durability through more thorough understanding of individuals' values and experiences.

F. Ethical Considerations. While any persuasive technology carries theoretical risks of misuse, our study addresses an ethically necessary question in the current landscape of AI development. The techniques we demonstrate could theoretically be applied to less beneficial ends—including potentially increasing prejudice toward other groups or advancing goals inconsistent with democratic values. The known liberal biases introduced by matching processes in language models complicate these concerns, as current AI systems may be more effective at promoting certain political perspectives than others. Whether similar techniques would prove equally effective for reducing prejudice toward conservative outgroups or groups typically disliked by right-wing audiences (such as political activists or ideological minorities) remains an open empirical question that merits systematic investigation.

These complex normative and empirical implications underscore why responsible researchers must proactively develop, test, and openly document AI's persuasive capabilities rather than leaving such powerful technologies to be developed without academic oversight. Our work advances understanding of how AI can be harnessed for prosocial outcomes while simultaneously revealing important limitations and potential risks. The systematic validation analysis we present, for instance, demonstrates that AI systems do not always perform as intended—a finding crucial for both beneficial applications and safeguards against misuse.

Rather than enabling harmful applications, our research contributes to building the knowledge base, methodological standards, and institutional practices needed to ensure that future uses of persuasive AI align with democratic values and respect for human dignity. The concentration of persuasive power in AI systems controlled by relatively few actors raises fundamental questions about information manipulation and democratic discourse that our field must grapple with directly. By conducting this work under institutional ethical review and with clear transparency about both promise and limitations, we help establish responsible frameworks for the inevitable expansion of AI's role in persuasive communication. This approach ensures that persuasive AI technologies are developed within ethical frameworks that prioritize public accountability and human welfare rather than private profit or political manipulation.

G. Conclusion. In this paper, we have shown evidence that conversations with AI couched in the language of moral foundations theory can significantly increase support for transgender rights—even in brief, text-based conversations with politically diverse respondents. We have also shown preliminary correlational evidence that when these conversations are more persuasive when they are aligned with one's personal moral framework. Taken together, these findings extend key theories of moral persuasion and deep canvassing to artificial interlocutors, and reveal new possibilities for cost-effective social interventions.

Integrating AI into prejudice reduction efforts marks a new and promising direction for social science and intervention design. Our study shows that carefully crafted AI systems can drive meaningful shifts in attitudes by engaging users with messages tailored to their moral values—at a cost that traditional human-led methods cannot match (12). However, the rapid decline in effect over one week highlights a crucial insight: persuasion is not a one-time event, but an ongoing process. For AI to play a lasting role in reducing prejudice, it must support sustained, adaptive engagement—not just deliver a single interaction.

Addressing the limitations we observed, especially around the mechanisms and longevity of change, will be essential to unlocking AI's full potential in fostering inclusive attitudes. Future research should refine both theoretical understanding and practical design, potentially combining the reach of AI with the depth of human connection to improve long-term outcomes.

Our results chart a promising but still preliminary path forward. While AI-mediated conversations can produce short-term shifts in attitudes, our findings underscore the challenges of sustaining these effects over time. Our exploratory analyses suggest that the moral content of persuasive messages may shape their immediate impact, but we find no evidence that matching predicts longer-term attitude change once baseline attitudes and engagement are taken into account. Still, the AI setting offers a unique analytic advantage: full transcripts of each conversation allow researchers to systematically analyze message-receiver matching—something rarely possible in human-led interventions. As future studies build on this foundation, hybrid approaches that combine the reach of AI with the relational depth of human interaction may hold the greatest promise. The task ahead is not just to scale persuasion, but to design it in ways that are targeted, resonant, and enduring—advancing the broader goal of inclusive democratic change.

Data availability

All data, unique materials, documentation, and code used in the analyses underlying this article will be posted at the Harvard Dataverse upon publication.

ACKNOWLEDGMENTS. We thank the following individuals for their feedback on this project: Alexander Coppock, Jesse Crosson, Kyle Dobson, David Broockman, Sunshine Hillygus, Peter Loewen, Walter Mebane, Robert Mickey, Christopher Mills, Jacob Montgomery, Brendan Nyhan, Yuki Shiraito, Andrew Simon, Arthur Spirling, George Tsebelis and the participants of workshops Koc University, Korea Advanced Institute of Science & Technology, Korea University, McGill University, Stanford University, Sungkyunkwan University, Sungshin Women's University, University of Virginia, Yonsei University, and the 2025 American Political Science Association Annual Meeting, and the 2025 Midwest Political Science Association Annual Meeting. Funding: the University of Toronto (S.S.), Social Sciences and Humanities Research Council Insight Development Grant (519977) (S.S.), the University of Virginia (J.B.H.) and Dartmouth College (C.C.). This project was previously posted as a preprint https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5229084here.

1. EL Paluck, SA Green, DP Green, The contact hypothesis re-evaluated. *Behav. Public Policy* **3**, 129–158 (2019).
2. J Haidt, *The Righteous Mind: Why Good People are Divided by Politics and Religion*. (Pantheon Books), (2012).
3. M Feinberg, R Willer, From gulf to bridge: When do moral arguments facilitate political influence? *Pers. Soc. Psychol. Bull.* **41**, 1665–1681 (2015).
4. M Feinberg, R Willer, Moral reframing: A technique for effective and persuasive communication across political divides. *Soc. Pers. Psychol. Compass* **13**, e12501 (2019).
5. M Feinberg, R Willer, The moral roots of environmental attitudes. *Psychol. science* **24**, 56–62 (2013).
6. D Broockman, J Kalla, Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science* **352**, 220–224 (2016).
7. HA Hatch, RH Warner, MR Grundy, KB Heck, Effectiveness of interventions for transgender prejudice reduction: A meta-analysis. *Sex Roles* **91**, 1–12 (2025).
8. K Brock-Petroshius, Race talk to change carceral attitudes: a field experiment on deep canvass organizing. *Soc. Serv. Rev.* **98**, 585–623 (2024).
9. E Santoro, DE Broockman, JL Kalla, R Porat, Listen for a change? a longitudinal field experiment on listening's potential to facilitate persuasion. *OSF Prepr.* (2024).
10. J Kalla, F Rosenbluth, Building support for controversial moral positions through moral reframing. *Proc. Natl. Acad. Sci.* **120**, e2217095120 (2023).
11. JL Kalla, AS Levine, DE Broockman, Personalizing moral reframing in interpersonal conversation: A field experiment. *The J. Polit.* **84**, 1239–1243 (2022).
12. Z Chen, et al., A framework to assess the persuasion risks large language model chatbots pose to democratic societies. *Preprint* (2025).
13. V Capraro, et al., The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS nexus* **3**, pgae191 (2024).
14. YR Velez, P Liu, Confronting core issues: A critical assessment of attitude polarization using tailored experiments. *Am. Polit. Sci. Rev.* pp. 1–18 (2024).
15. L Ibrahim, S Huang, L Ahmad, M Anderljung, Beyond static ai evaluations: advancing human interaction evaluations for llm harms and risks. *arXiv preprint arXiv:2405.10632* (2024).
16. JG Voelkel, S Muldowney, R Willer, , et al., Ai-generated messages can be used to persuade humans on policy issues. *arXiv preprint* (2025).
17. TW Kim, A Duhachek, Artificial intelligence and persuasion: A construal-level account. *Psychol. science* **31**, 363–380 (2020).
18. TH Costello, G Pennycook, DG Rand, Just the facts: How dialogues with ai reduce conspiracy beliefs. *arXiv preprint* (2025).
19. A Rogiers, S Noels, M Buyl, T De Bie, Persuasion with large language models: a survey. *arXiv preprint arXiv:2411.06837* (2024).
20. S Sevi, C Mekik, Ai weakens, but does not strengthen, political attitudes. *PsyArXiv preprint* (2026).
21. M Linegar, B Sinclair, S van der Linden, RM Alvarez, Prebunking elections rumors: Artificial intelligence assisted interventions increase confidence in american elections. *arXiv preprint* (2024).
22. G Spitale, N Biller-Andorno, F Germani, Ai model gpt-3 (dis) informs us better than humans. *Sci. Adv.* **9**, eadh1850 (2023).
23. JA Goldstein, J Chao, S Grossman, A Stamos, M Tomz, How persuasive is ai-generated propaganda? *PNAS nexus* **3**, pgae034 (2024).
24. F Salvi, MH Ribeiro, R Gallotti, R West, On the conversational persuasiveness of large language models: A randomized controlled trial. *arXiv preprint* (2024).
25. CA Bail, Can generative ai improve social science? *Proc. Natl. Acad. Sci.* **121**, e2314021121 (2024).
26. K Hackenburg, H Margetts, Evaluating the persuasive influence of political microtargeting with large language models. *Proc. Natl. Acad. Sci.* **121**, e2403116121 (2024).
27. YR Velez, DP Green, S Sevi, Chatbot voting advice applications inform but seldom sway young unaligned voters. *Proc. Natl. Acad. Sci.* **122**, e2515516122 (2025).
28. LP Argyle, et al., Leveraging ai for democratic discourse: Chat interventions can improve online political conversations at scale. *Proc. Natl. Acad. Sci.* **120**, e2311627120 (2023).
29. JD Teeny, SC Matz, We need to understand "when" not "if" generative ai can enhance personalized persuasion. *Proc. Natl. Acad. Sci.* **121**, e2418005121 (2024).
30. B Bago, JF Bonnefon, Generative ai as a tool for truth. *Science* **385**, 1164–1165 (2024).
31. S Abdurahman, et al., Perils and opportunities in using large language models in psychological research. *PNAS nexus* **3**, pgae245 (2024).
32. M Burtell, T Woodside, Artificial influence: An analysis of ai-driven persuasion. *arXiv preprint arXiv:2303.08721* (2023).
33. M Dehnert, PA Mongeau, Persuasion in the age of artificial intelligence (ai): Theories and complications of ai-based persuasion. *Hum. Commun. Res.* **48**, 386–403 (2022).
34. SC Matz, et al., The potential of generative ai for personalized persuasion at scale. *Sci. Reports* **14**, 4692 (2024).
35. JN Druckman, A framework for the study of persuasion. *Annu. Rev. Polit. Sci.* **25**, 65–88 (2022).
36. JN Druckman, Persuasive political targeting in *The Handbook of Personalized Persuasion*. (Routledge), pp. 236–259 (2025).
37. J Green, et al., Using general messages to persuade on a politicized scientific issue. *Br. J. Polit. Sci.* **53**, 698–706 (2023).
38. RL Stotzer, Violence against transgender people: A review of united states data. *Aggress. Violent Behav.* **14**, 170–179 (2009).
39. AL Wirtz, TC Poteat, M Malik, N Glass, Gender-based violence against transgender people in the united states: A call for research and programming. *Trauma, Violence, & Abus.* **21**, 227–241 (2020).
40. RJ Testa, et al., Effects of violence on transgender people. *Prof. Psychol. Res. Pract.* **43**, 452 (2012).
41. J Graham, J Haidt, BA Nosek, Liberals and conservatives rely on different sets of moral foundations. *J. personality social psychology* **96**, 1029 (2009).
42. RM Perloff, *The dynamics of persuasion: Communication and attitudes in the 21st century*. (Routledge), (1993).
43. M Arundel, The complete works of aristotle. ed. by j. barnes. princeton, nj: Princeton university press, 1984. *State Commonwealth: The Theory State Early Mod. England, 1549–1640* p. 235 (2016).
44. HD Lasswell, The structure and function of communication in society. *The communication ideas* **37**, 136–139 (1960).
45. WJ McGuire, The nature of attitudes and attitude change. *The handbook social psychology/Addison-Wesley* (1969).
46. RE Petty, P Briñol, J Teeny, J Horcajo, The elaboration likelihood model: Changing attitudes toward exercising and beyond. *Persuas. communication sport, exercise, physical activity* pp. 22–37 (2017).
47. S Chaiken, Y Trope, *Dual-process theories in social psychology*. (Guilford Press), (1999).
48. JN Druckman, A Lupia, Preference change in competitive political environments. *Annu. Rev. political science* **19**, 13–31 (2016).
49. D Chong, JN Druckman, Framing public opinion in competitive democracies. *Am. political science review* **101**, 637–655 (2007).
50. A Gutmann, DF Thompson, *Democracy and disagreement*. (Harvard University Press), (2009).
51. A Coppock, *Persuasion in parallel: How information changes minds about politics*. (University of Chicago Press), (2023).
52. JL Kalla, DE Broockman, Reducing exclusionary attitudes through interpersonal conversation: Evidence from three field experiments. *Am. Polit. Sci. Rev.* **114**, 410–425 (2020).
53. DE Broockman, JL Kalla, JS Sekhon, The design of field experiments with survey outcomes: A framework for selecting more efficient, robust, and ethical designs. *Polit. Analysis* **25**, 435–464 (2017).
54. TH Costello, S Yang, JJ Van Bavel, Ai chatbots can reduce belief in misinformation and conspiracy theories. *Sci. Reports* **14**, 2059 (2024).
55. C Crabtree, JB Holbein, M Bosley, S Sevi, Dataset for: Can ai help reduce prejudice? evaluating the effectiveness of ai-powered personalized persuasion on support for transgender rights (https://doi.org/xx.xxxx/xxxx) (2025).
56. MN Stagnaro, et al., Representativeness versus attentiveness: a comparison across nine online survey samples. *PsyArXiv* **22** (2024).
57. A Coppock, OA McClellan, Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Res. & politics* **6** (2019).
58. C Crabtree, Yes, you can do global, cross-cultural behavioral science research using existing survey firms. *Proc. Natl. Acad. Sci.* **122**, e2418102122 (2025).
59. W Lin, Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals Appl. Stat.* **7**, 295–318 (2013).
60. K Imai, M Ratkovic, Covariate balancing propensity score. *J. Royal Stat. Soc. Ser. B (Statistical Methodol.* **76**, 243–263 (2014).
61. G Hong, Marginal mean weighting through stratification: Adjustment for selection bias in multilevel data. *J. Educ. Behav. Stat.* **35**, 499–531 (2010).
62. JM Robins, TS Richardson, Comment on "causal effects in nonexperiments and principal stratification" by judea pearl. *J. Am. Stat. Assoc.* **102**, 683–686 (2007).
63. TJ VanderWeele, P Ding, Sensitivity analysis in observational research: introducing the E-value. *Annals Intern. Medicine* **167**, 268–274 (2017).
64. DS Lee, Training, wages, and sample selection: Estimating sharp bounds on treatment effects (2009).
65. Y Benjamini, Y Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Royal Stat. Soc. Ser. B (Methodological)* **57**, 289–300 (1995).
66. AC Cameron, JB Gelbach, DL Miller, Bootstrap-based improvements for inference with clustered errors. *The Rev. Econ. Stat.* **90**, 414–427 (2008).
67. H White, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**, 817–838 (1980).
68. AS Gerber, DP Green, *Field Experiments: Design, Analysis, and Interpretation*. (W. W. Norton & Company, New York, NY), (2012).
69. J Bisbee, A Spirling, What to do when humans are no longer the gold standard: Large language models, state of the art, and robustness. Working paper (2025).
70. LP Argyle, et al., Testing theories of political persuasion using ai. *Proc. Natl. Acad. Sci.* **122**, e2412815122 (2025).
71. C Snell, J Lee, K Xu, A Kumar, Scaling llm test-time compute optimally can be more effective than scaling model parameters (2024).
72. S Yao, et al., Tree of thoughts: Deliberate problem solving with large language models. *Adv. neural information processing systems* **36**, 11809–11822 (2023).
73. JS Park, et al., Generative agent simulations of 1,000 people (2024).