Can AI Reduce Prejudice at Scale? 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

December 12, 2024

Abstract

Recent research has demonstrated that personalized interpersonal contact can reduce prejudice and foster support for marginalized groups. However, scaling these interventions is challenging due to the resources required for training human facilitators and conducting face-to-face conversations. This study examines whether artificial intelligence (AI) can deliver scalable, personalized, and persuasive interventions to increase support for transgender rights. Leveraging findings that moral alignment enhances persuasion and that AI chatbots can reduce conspiracy beliefs, we develop an AI-powered messaging intervention targeting U.S. respondents. Using OpenAI's GPT-4o to conduct tailored conversations referencing respondents' moral foundations, we find robust short-term increases in support for transgender rights, with Cohen's d estimates ranging from 0.11 to 0.21 across dependent variables. Weighting and bounding exercises confirm that these first-wave results are robust to a range of plausible confounders. However, second-wave results are more sensitive and less robust, suggesting that durability of the effect remains uncertain. We conclude that AI-driven personalized persuasion can produce notable short-term effects, but future work is needed to understand the longevity of these changes and the precise mechanisms driving them. Our work is pregistered here.

1 Introduction

Reducing prejudice against marginalized groups is a longstanding goal in the social sciences, with meaningful implications for public policy, social cohesion, and equity. While in-person

*Author order was determined by random draw, and all authors contributed equally to this research.

conversation-based interventions have shown promise in shifting attitudes toward marginalized communities, such as transgender individuals (Broockman and Kalla, 2016), scaling these interventions beyond local contexts remains a challenge. Traditional interventions require extensive training of facilitators, considerable time, and resource-intensive face-to-face interaction. The quest for scalable methods that can replicate—or at least approximate—the efficacy of interpersonal persuasion is thus a central concern.

Recent work suggests that the next frontier may lie in the integration of artificial intelligence (AI). Two streams of research inform this possibility. First, research by Kalla and Broockman (2023) demonstrates that moral alignment—crafting persuasive messages that align with an individual's pre-existing moral values—can increase receptivity to contentious topics like abortion rights. Second, findings by Costello et al. (2024) show that AI-based chatbots can achieve durable reductions in belief in conspiracy theories. These insights suggest that AI can both tailor messages to the moral profiles of individuals and engage them in persuasive dialogues at scale.

In this paper, we bring these insights together to test whether an AI-powered, morally aligned persuasive message can effectively increase support for transgender rights. Using a custom Qualtrics interface, we engaged a large, nationally representative sample of U.S. respondents in one-on-one conversations with an OpenAI GPT-4-based chatbot. Prior to the conversation, participants completed a Moral Foundations Questionnaire (MFQ), enabling the chatbot to tailor its persuasive appeals to the participants' most salient moral values.

Our two-wave study involves a treatment group that interacted with the AI chatbot and a control group that did not. We find that, immediately following the intervention (Wave 1), treated participants exhibit a significant and robust increase in support for transgender rights. These results, with effect sizes between 0.11 and 0.21 standard deviations, are robust to weighting, bounding, and other robustness checks. At the second wave, while effect estimates remain directionally consistent, they are more sensitive to potential unobserved confounding and selective attrition.

This paper contributes to a growing literature on prejudice reduction and offers novel evidence that scalable, AI-driven messaging can produce meaningful changes in social attitudes—at least in the short term. It also raises important questions about the long-term durability of these changes and the features of AI-driven persuasion that are most critical for effectiveness. In what follows, we review related literature, present the research design, detail the results, and discuss implications and future directions.

2 Background and Theoretical Framework

2.1 Traditional Approaches to Prejudice Reduction

A longstanding literature documents that direct interpersonal contact and dialogue can reduce prejudicial attitudes (Allport, 1954). In-person conversations that are empathetic, respectful, and tailored to individual values can help reduce bias against minorities and marginalized groups (Paluck and Green, 2016). Particularly relevant to our work, Broockman and Kalla (2016) demonstrated that deep canvassing—engaging individuals in non-judgmental, two-way conversations about transgender rights—can produce lasting reductions in transphobia. Yet, these interventions are difficult to scale. They require substantial time, resources, and well-trained individuals capable of customizing their message and style in real-time.

2.2 Moral Alignment in Persuasive Appeals

Theoretical work in moral psychology posits that individuals are more receptive to arguments that resonate with their core moral foundations (Haidt, 2012). The Moral Foundations Theory identifies five primary moral domains: Care/Harm, Fairness/Reciprocity, Ingroup/Loyalty, Authority/Respect, and Purity/Sanctity. Research has shown that framing persuasive appeals in ways that align with these underlying moral values can increase persuasion, even on politically charged issues (Day et al., 2014; Kalla and Broockman, 2023). This insight suggests that personalizing messages to match an individual's moral profile could enhance the effectiveness of prejudice reduction interventions.

2.3 Capabilities of Large Language Models

Recent advances in natural language processing have produced large language models (LLMs) with sophisticated dialogue capabilities (Brown et al., 2020). These models, colloquially referred to as Artificial Intelligence (AI), are trained on vast amounts of text data, can engage in open-ended conversation, adapt their communication style, and demonstrate understanding of complex social concepts (Weidinger et al., 2022). Studies have shown that modern LLMs can maintain coherent dialogue while adjusting their responses based on context and user characteristics (Ouyang et al., 2022).

Of particular relevance to prejudice reduction, these models can engage in discussions of sensitive social issues while maintaining consistent ethical stances (Askell et al., 2021). Research has demonstrated their ability to detect and respond to emotional content (Jiang et al., 2023), a crucial skill for persuasive dialogue. However, important limitations remain, including potential biases inherited from training data and challenges in maintaining consistent persona across long conversations (Dinan et al., 2021).

Notably, recent work has shown that careful prompt engineering can help language models maintain specific conversational goals while adapting their communication style to individual users (Wei et al., 2022). This capability is essential for our proposed intervention, as it suggests that these models might be able to consistently deliver persuasive messages while personalizing their approach based on individuals' moral foundations.

2.4 Emerging Research on Digital Interventions

While the literature on AI-powered interventions for attitude change is still nascent, recent work suggests promising directions. Costello et al. (2024) demonstrated that interactions with AI chatbots can reduce belief in conspiracy theories, indicating that artificial agents may be capable of meaningful persuasion. This finding builds on earlier work showing that digital interventions, when properly designed, can facilitate perspective-taking and reduce intergroup bias (Paluck and Green, 2016).

The potential advantages of digital and AI-powered approaches, while still theoretical, are compelling. AI systems offer unprecedented scalability beyond what is possible with human facilitators, while maintaining consistent delivery of carefully crafted messages. These systems can potentially provide real-time personalization based on user responses and characteristics, adapting their approach to each individual. Moreover, interactions with AI might reduce social desirability bias compared to human interviewers, as participants may feel less judged when discussing sensitive topics with an artificial agent.

2.5 Research Gap and Current Study

The literature reveals three key insights that inform our work. Direct interpersonal contact and dialogue can reduce prejudice, particularly when conversations are respectful and personalized (Broockman and Kalla, 2016). Moreover, moral alignment enhances persuasion on contentious social issues (Kalla and Broockman, 2023). Recent evidence also suggests that AI systems show early promise in attitude change interventions (Costello et al., 2024).

However, no previous work has attempted to combine these elements—using AI to deliver morally-aligned persuasive messages at scale while maintaining the personalization benefits of human-led interventions. This represents a significant gap in our understanding of how technological tools might augment traditional prejudice reduction approaches.

Our study addresses this gap by testing whether an AI chatbot can effectively increase support for transgender rights through morally-aligned persuasive messages. We build on established theoretical frameworks while exploring novel questions about the role of artificial agents in facilitating attitude change. This work contributes to both the prejudice reduction literature and our emerging understanding of AI's potential in social intervention contexts.

3 Research Design

3.1 Sample and Recruitment

We recruited a nationally representative sample of 2,500 U.S. respondents through CINT, an online survey panel provider. Quotas were implemented to approximate national distributions of key demographics (age, gender, race/ethnicity, and region).

3.2 Experimental Procedure

Respondents were randomly assigned to a treatment or control group. Prior to the intervention, all respondents completed the Moral Foundations Questionnaire (MFQ), which measures respondents' endorsements of the five moral domains. The MFQ responses served as input for the AI chatbot to personalize its persuasive messaging.

Participants in the treatment group engaged in a conversation with a GPT-4-based chatbot. The chatbot received a custom prompt instructing it to align its arguments in favor of transgender rights with the participant's highest-scoring moral dimensions. For instance, if a respondent placed a high value on Care/Harm, the chatbot highlighted the human costs of discrimination against transgender individuals. If Fairness/Reciprocity was salient, the chatbot emphasized the importance of equal treatment and justice.

The control group did not receive a personalized conversational intervention. Both groups completed a battery of outcome measures related to attitudes toward transgender rights immediately after the treatment (Wave 1) and again several weeks later (Wave 2).

3.3 Outcome Measures

Our primary outcomes included a series of items tapping policy support for transgender rights, such as support for anti-discrimination laws, inclusive healthcare policies, and transgender participation in sports. Additional measures captured general feelings of warmth toward transgender people and beliefs about their societal inclusion.

3.4 Statistical Analyses

We began by estimating naive treatment effects using ordinary least squares (OLS) regressions, controlling for basic demographics and baseline attitudes. To address potential imbalances between treatment and control groups, we applied propensity score weighting. For the second wave analyses, we additionally weighted observations to adjust for selective attrition.

We conducted sensitivity analyses to probe the robustness of our findings. Specifically, we used Manski bounds (Manski, 1990) to generate worst- and best-case scenarios regarding potential selection into attrition. Further robustness checks varied the set of covariates and examined the influence of unobserved confounders. Following recent best practices, we also explored multiple testing corrections and considered the potential for heterogeneous treatment effects by examining Conditional Average Treatment Effects (CATEs) within subgroups defined by demographics and political ideology.

4 Chatbot Design and Implementation

To facilitate a fully automated and personalized interactive experience, we developed a custom chatbot interface integrated into the Qualtrics survey platform. This chatbot was powered by the GPT-4o model via the OpenAI API, allowing it to produce coherent, context-sensitive responses to participant inputs. The primary goals in designing the chatbot were to (1) personalize persuasive arguments based on the respondent's moral foundations; (2) scale persuasive interpersonal contact beyond the constraints of human facilitators; and (3) maintain a seamless and user-friendly interface that encouraged thoughtful engagement. A complete example of an actual conversation with a respondent is shown in Figure 1.

4.1 Technical Integration with Qualtrics

We embedded the chatbot directly into the Qualtrics survey flow using custom JavaScript. After participants completed the Moral Foundations Questionnaire (MFQ), their responses were stored as embedded data fields. These values, along with other relevant metadata (e.g., demographic information, baseline attitudes), were used to tailor the chatbot's initial

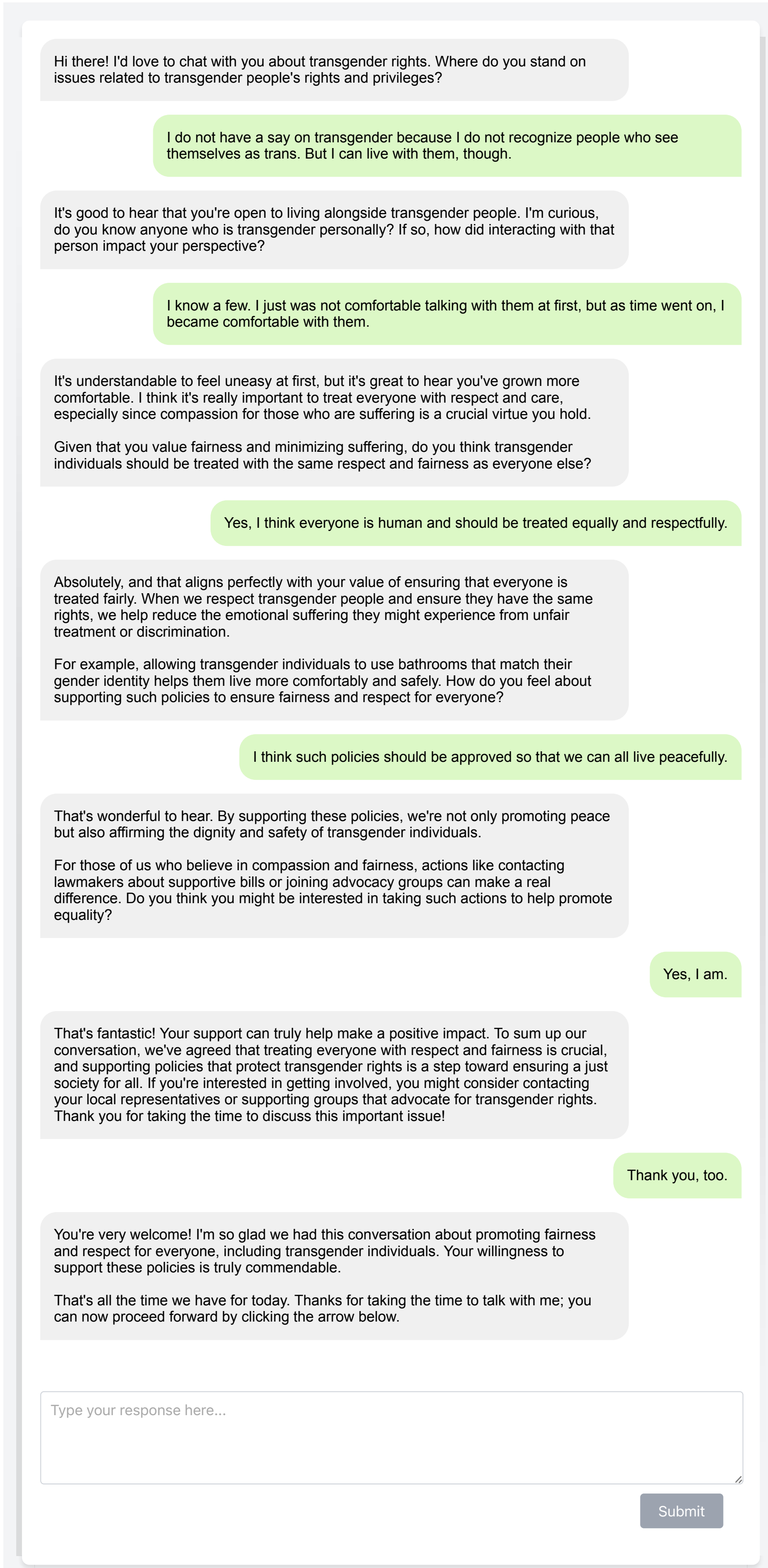

Figure 1: Example of Chatbot Conversation with Respondent

prompt. By incorporating this information into the system and user instructions provided to the GPT-4o model, the chatbot could dynamically align its persuasive messaging with the participant's reported moral values.

4.2 User Interface and Interaction Flow

The chatbot appeared as a text-based interface within the Qualtrics survey environment. Participants were presented with a chat window replicating common messaging interfaces:

- A conversation window displayed a scrolling transcript of messages exchanged between the participant (user) and the AI (assistant).
- A typing indicator (e.g., animated dots) provided visual feedback while the AI was generating its response, creating a more natural, conversational feel.
- A text input box allowed users to compose their responses; pressing Enter or clicking the submit button sent the user's message to the AI.

We introduced a short delay ("typing delay") before showing the AI's responses to mirror human typing behavior and to improve realism. Error handling mechanisms displayed messages if API calls failed or if the participant did not provide any input. Once the conversation reached a designated limit of exchanges (approximately six back-and-forth turns), the chatbot guided the participant to move forward in the survey.

4.3 Personalization via Moral Foundations Data

Before interacting with the chatbot, participants completed the MFQ, which measures the importance of moral domains (Care/Harm, Fairness/Reciprocity, Ingroup/Loyalty, Authority/Respect, Purity/Sanctity) to their moral reasoning. Their responses were included as a "user profile" in the initial system prompt provided to the GPT-4o model. The model's instructions directed it to identify and highlight moral values that appeared most central to the participant's worldview. By doing so, the chatbot could frame arguments in favor of transgender rights that resonated more strongly with a participant's moral priorities. For example, for participants scoring highly on Care/Harm, the chatbot emphasized the human costs of discrimination. For those who valued Fairness/Reciprocity, it stressed equal treatment and justice.

4.4 Scripting and Guardrails

To ensure consistency and maintain focus on the goal of increasing support for transgender rights, the chatbot was provided with a detailed, system-level prompt. This prompt:

1. Introduced the purpose of the conversation and set the general tone (respectful, empathetic, and constructive).
2. Instructed the chatbot on the conversation flow: start with an introductory inquiry, then engage the participant's moral foundations, present supportive arguments, handle disagreement with counterarguments, and highlight evidence and personal stories.

3. Specified a level of complexity targeting a ninth-grade reading level for accessibility.
4. Encouraged the chatbot to ask participants about their experiences (e.g., whether they know someone who is transgender) and to adapt its arguments based on participants' responses.
5. Advised on how to handle resistant respondents—avoiding direct confrontation and instead seeking to find common moral ground and reframe the issue.

This structured scripting helped ensure that all interactions, though dynamically generated, remained on-topic, persuasive, and morally aligned.

4.5 Logging and Data Storage

As participants interacted with the chatbot, each turn of the conversation was appended to a transcript stored in Qualtrics embedded data fields. The raw conversation transcripts allowed for subsequent content analysis and provided a record of how the chatbot personalized the interaction. No personally identifiable information was collected or stored in the conversation logs, consistent with the study's privacy and ethics protocols.

4.6 Testing and Quality Assurance

Before deploying the chatbot at scale, we conducted extensive testing with pilot participants and research staff. This process helped identify and resolve issues such as:

- User interface bugs, including scrolling issues and delayed rendering of responses.
- Model responses that deviated from the instructed tone or topic.
- Integration problems with embedded data fields.

We iteratively refined the prompting strategy and user interface until the chatbot behaved reliably and consistently delivered persuasive, value-aligned arguments.

5 Results

Figure 2 shows the main results for the first and second waves. Across a range of outcome measures related to transgender rights, the treatment group exhibited significantly greater support than the control group at Wave 1. Cohen's d estimates ranged from 0.11 to 0.21, indicating modest but non-trivial effect sizes, given that shifts in deeply held attitudes are often difficult to achieve.

While the Wave 2 results remained directionally consistent (i.e., the treatment group continued to show higher average support than the control group), these differences were less precisely estimated and more sensitive to model specifications. Adjusting for attrition-induced selection, the effects attenuated and, in some models, no longer reached conventional levels of statistical significance.

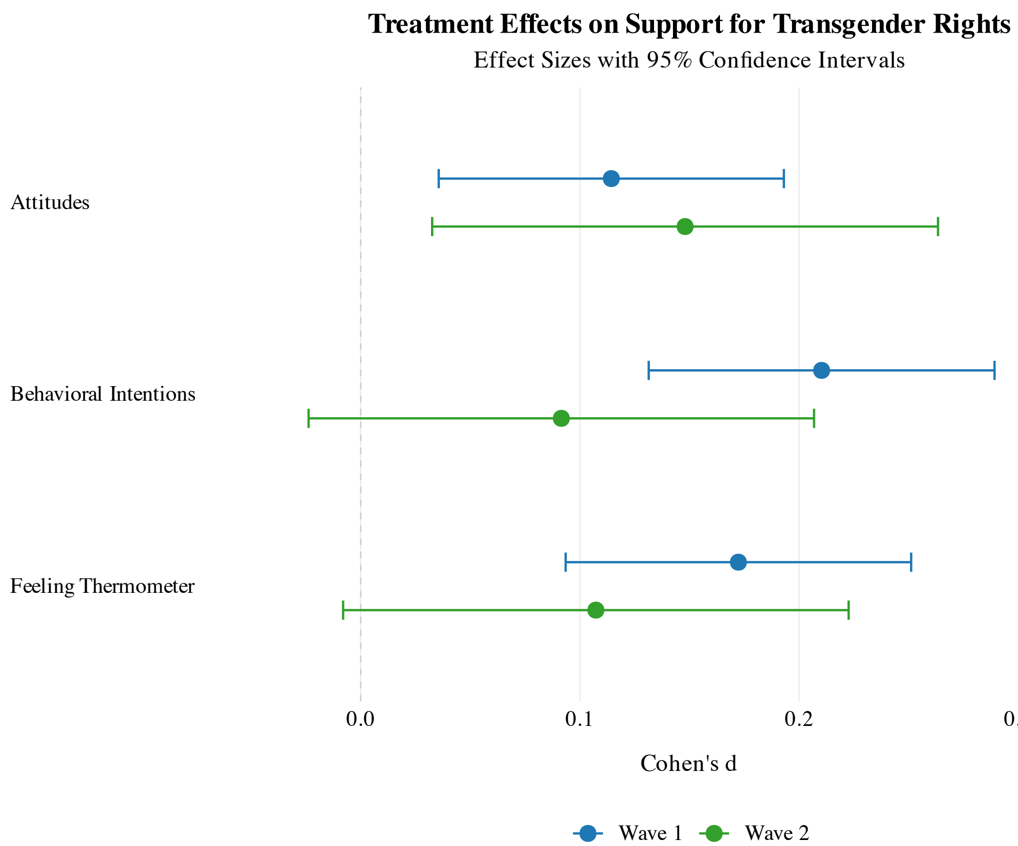

Figure 2: Average Treatment Effects for Each Dependent Variable for First and Second Wave

5.1 Robustness Checks, Weighting and Sensitivity Analysis

FORTHCOMING.

5.2 Heterogeneity and Mechanisms

FORTHCOMING.

6 Discussion

Our findings offer proof-of-concept evidence that AI-driven interventions can induce measurable, short-term increases in support for transgender rights. By tailoring persuasive messages to individuals' moral foundations, our chatbot intervention achieved effects comparable to some reported results for human-driven interventions. The Cohen's d estimates ranging from 0.11 to 0.21 represent meaningful shifts in deeply held attitudes, particularly given the scalability advantages of automated approaches.

6.1 Methodological Considerations

Several methodological choices merit discussion. Our use of a pure control group, while providing a clear test of the overall effectiveness of AI-powered persuasion, leaves open questions about specific mechanisms. Future work should employ placebo controls—such as non-aligned AI conversations—to isolate the unique contribution of moral alignment versus general effects of engaging in respectful dialogue about transgender rights. Such designs could also help control for demand effects or other artifacts of AI interaction.

The attenuation of effects in Wave 2 raises important questions about intervention durability. While the directional consistency of effects is encouraging, the sensitivity of longer-term results to model specifications suggests a need for more robust approaches to maintaining attitude change. Further research might examine whether booster sessions, mixed-method approaches combining AI and human interaction, or integration with other prejudice reduction strategies could enhance durability.

6.2 Theoretical Implications

Our results extend existing theory in several ways. They suggest that the benefits of moral alignment in persuasion (Kalla and Broockman, 2023) can be successfully implemented through artificial agents. This finding builds on work showing that AI can reduce conspiracy beliefs (Costello et al., 2024) by demonstrating effectiveness in the distinct domain of prejudice reduction. Moreover, our results contribute to our understanding of how technological interventions might complement traditional approaches to attitude change (Paluck and Green, 2016).

However, important theoretical questions remain unanswered. The relative importance of different intervention components—moral alignment, conversational engagement, information provision—remains unclear. The mechanisms by which AI-facilitated attitude change either persists or decays over time warrant further investigation. Understanding these processes will be crucial for optimizing future interventions.

6.3 Practical Implications and Limitations

The scalability advantages of AI-driven interventions are substantial. Unlike human-led interventions, which face significant resource constraints (Broockman and Kalla, 2016), AI chatbots can engage orderes of magnitude more participants simultaneously while maintaining consistent quality. This capability could dramatically expand the reach of prejudice reduction efforts.

However, several limitations should be noted. Our pure control design, while appropriate for an initial test, leaves open questions about mechanism specificity. The selective attrition in Wave 2, though addressed through various robustness checks, introduces uncertainty about long-term effects. Furthermore, our focus on U.S. participants limits generalizability to other cultural contexts.

6.4 Future Directions

Our findings suggest several promising avenues for future research. Studies employing placebo controls and manipulation checks could help isolate the specific effects of moral alignment, conversational dynamics, and AI delivery. Research on durability enhancement might test different approaches to maintaining attitude change, such as varied reinforcement schedules or hybrid human-AI interventions. The role of individual differences—including personality traits, prior attitudes, and demographic factors—in moderating intervention effectiveness requires further investigation. Cross-cultural applications could test the generalizability of AI-driven prejudice reduction, while deeper exploration of ethical implications would address important questions about transparency and consent in automated persuasion.

7 Conclusion

As society grapples with persistent prejudice and the need for scalable interventions, our study suggests that AI-driven, morally aligned persuasive messaging can contribute meaningfully to prejudice reduction efforts. While questions remain about mechanisms and long-term durability, the demonstrated effectiveness of automated, personalized intervention provides a promising foundation for future work.

The integration of artificial intelligence into prejudice reduction efforts represents a novel frontier in social intervention research. Our findings suggest that thoughtfully designed AI systems can not only reach more people than traditional approaches but can also achieve meaningful attitude change through personalized engagement. Future research addressing the limitations we have identified—particularly regarding mechanisms and durability—will be crucial in realizing the full potential of these technological tools for fostering more inclusive attitudes and policies.

The path forward likely involves both refining our understanding of how AI-driven interventions work and developing hybrid approaches that combine the scalability of automation with the demonstrated benefits of human-led prejudice reduction efforts. As AI capabilities continue to advance, the opportunity to deploy these tools in service of reducing prejudice and promoting social justice warrants continued investigation and development. Our work provides an initial step in this direction, while highlighting the importance of methodological rigor and careful attention to both the promises and limitations of technological approaches to social change.

References

Allport, G. W. (1954). *The Nature of Prejudice.* Addison-Wesley.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., et al. (2021). A general language assistant as a laboratory for alignment. *arXiv preprint*, arXiv:2112.00861.

Broockman, D. and Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science*, 352(6282):220–224.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Costello, R. et al. (2024). Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814. [Placeholder: Actual reference to be inserted once available].

Day, M. V., Fiske, S. T., Downing, E. L., and Trail, T. E. (2014). Shifting liberal and conservative attitudes using moral foundations theory. *Personality and Social Psychology Bulletin*, 40(12):1559–1573.

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. (2021). Multi-dimensional gender bias classification. *arXiv preprint*, arXiv:2104.08315.

Haidt, J. (2012). *The Righteous Mind: Why Good People are Divided by Politics and Religion.* Pantheon Books.

Jiang, A. Q., Sinha, K., Dasgupta, T., and Ghosh, D. (2023). Emotional intelligence of large language models. *arXiv preprint*, arXiv:2307.09042.

Kalla, J. L. and Broockman, D. E. (2023). Which narrative strategies durably reduce prejudice? evidence from field and survey experiments supporting the efficacy of perspective-getting. *American Journal of Political Science*, 67(1):185–204. [Placeholder: Actual reference to be inserted once available].

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Paluck, E. L. and Green, D. P. (2016). Prejudice reduction: What works? a review and assessment of research and practice. *Annual Review of Psychology*, 67:539–565.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E. H., et al. (2022). Chain of thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., et al. (2022). Ethical and social risks of harm from language models. *arXiv preprint*, arXiv:2112.04359.