PNAS RESEARCH ARTICLE | POLITICAL SCIENCES COMPUTER SCIENCES OPEN ACCESS

Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale

Lisa P. Argyle[a,1,2], Christopher A. Bail[b,1,2], Ethan C. Busby[a,1], Joshua R. Gubler[a,1], Thomas Howe[c,1], Christopher Rytting[c,1], Taylor Sorensen[d,1], and David Wingate[c,1]

Edited by Kathleen Jamieson, University of Pennsylvania, Philadelphia, PA; received July 9, 2023; accepted August 18, 2023

Political discourse is the soul of democracy, but misunderstanding and conflict can fester in divisive conversations. The widespread shift to online discourse exacerbates many of these problems and corrodes the capacity of diverse societies to cooperate in solving social problems. Scholars and civil society groups promote interventions that make conversations less divisive or more productive, but scaling these efforts to online discourse is challenging. We conduct a large-scale experiment that demonstrates how online conversations about divisive topics can be improved with AI tools. Specifically, we employ a large language model to make real-time, evidence-based recommendations intended to improve participants' perception of feeling understood. These interventions improve reported conversation quality, promote democratic reciprocity, and improve the tone, without systematically changing the content of the conversation or moving people's policy attitudes.

democratic deliberation | computational social science | generative AI | political science

Significance

We develop an AI chat assistant that makes real-time, evidence-based suggestions for messages in divisive online political conversations. In a randomized controlled trial, we show that when one participant in a conversation had access to this assistant, it increased their partner's reported quality of conversation and both participants' willingness to grant political opponents space to express and advocate their views in the public sphere. Participants had the ability to accept, modify, or ignore the AI chat assistant's recommendations. Notably, participants' policy positions were unchanged by the intervention. Though many are rightly concerned about the role of AI sowing social division, our findings suggest it can do the opposite—improve political conversations without manipulating participants' views.

Social scientists have long observed that "conversation is the soul of democracy" (1, 2). Interpersonal discussions across social divides can help diverse groups of people peacefully identify solutions to shared problems, avoid violent conflict, and come to understand one another better (2–8). Historically, these conversations have occurred face-to-face (8), but online conversations now play a central role in public dialogue. More than 100 billion messages are sent every day on Facebook and Instagram alone (9), and approximately 7 billion conversations occur daily on Facebook Messenger (10). Such conversations can have far-reaching impact. Some of the largest social movements in human history have emerged out of sprawling conversations on social media, and discussions between high-profile social media users can shape the stock market, politics, and many other aspects of human experience (11–14). The internet thus has the capacity to empower an ever-increasing number of people to communicate and deliberate together.

However, there is growing concern that much of online conversation does the opposite (15–18). Nearly half of social media users report observing mean or cruel behavior, and many indicate that online divisiveness and incivility complicate a variety of relationships in their lives—with family, friends, and work colleagues (19). As such, many members of the public either avoid online discussions about politics or unwittingly find themselves arguing online in a corrosive, unconstructive manner (20–23). Such rhetoric has been linked to partisan violence (24, 25), disengagement from politics and public life (23, 26), and reduced capacity to find compromise (27).

These online political conversations are a far cry from the types of conversations scholars identify as the foundation of deliberative democracy (4, 28–30). Democratic deliberation demands conversations built on what deliberative scholars call "democratic reciprocity": a willingness to grant political opponents the same right to express and advocate their views in the public sphere that we hope they will grant us (4). This does not imply conversations that end with agreement, or condoning ideas that are problematic, but instead suggests conversational willingness to listen to and engage in good faith with those who have political opinions that differ from our own.

Thankfully, scholarship on how to facilitate these types of conversations continues to grow across disciplines (7, 31–38). These studies identify a range of strategies to increase the likelihood that members of rival groups thoughtfully listen to and engage with others' perspectives. Though conversations built on such strategies rarely result in immediate resolution of political problems and disagreements, many scholars see them as a necessary condition for increasing mutual understanding, compromise, and coalition-building. That is, "hearing the other side" (39) can be democratically beneficial even if disagreement remains, providing a variety of broader benefits related to social cohesion and democracy (7, 8, 22, 32).

Author contributions: L.P.A., C.A.B., E.C.B., J.R.G., C.R., T.S., and D.W. designed research; L.P.A., C.A.B., E.C.B., J.R.G., T.H., C.R., T.S., and D.W. performed research; L.P.A., E.C.B., J.R.G., C.R., and D.W. analyzed data; T.H. app development; and L.P.A., C.A.B., E.C.B., J.R.G., C.R., and D.W. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]L.P.A., C.B., E.C.B., J.R.G., T.H., C.R., T.S., and D.W. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: lpargyle@byu.edu or christopher.bail@duke.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2311627120/-/DCSupplemental.

Published October 3, 2023.

In what follows, we present the results of a field experiment that employed cutting-edge AI tools—in this case, the large language model (LM) GPT-3—to scale up evidence-based conversation-improving interventions. We invited proponents and opponents of gun regulation in the United States into online conversations, randomly assigning a pretrained chat assistant powered by GPT-3 to some of the participants. We show that intervention by the chat assistant, which recommended real-time, context-aware, and evidence-based ways to rephrase messages, improved democratic reciprocity in conversation, particularly for the partner of the person assigned the AI assistant.

AI Tools in Social Science. Political actors and social scientists increasingly use AI tools to influence and study the social world (40–43). LMs like ChatGPT and others highlight the ability of AI to generate human-sounding text and perform tasks previously thought impossible (44). Given their potential to identify and replicate complex patterns in text, LMs provide a promising way to explore social outcomes (45). One important advance of these models is their capacity for "few-shot" learning, or their ability to learn to perform a task from just a few exemplars without requiring parameter updates (46).

While many observers are rightfully concerned about the negative effects of biases present in LMs and other AI tools (47–51), the same model features that generate these biases also enable LMs to produce text that is nuanced and multifaceted in its representation of a range of people, tones, ideas, and attitudes (45). Prior AI-in-the-loop applications have demonstrated that AI can help people be more empathetic in peer mental health support conversations (52), and that AI-induced reflection and restatement can improve the quality of conversations (53–55). We build on that work, as well as on frameworks developed by scholars like Fishkin et al. (56), to demonstrate that dynamic, real-time, and context-aware LM-generated recommendations can improve the quality of political conversations by helping people communicate their willingness to respect, acknowledge, and be open to the views of their political opponents.

Strategies to Improve the Perception of Being Understood and Democratic Reciprocity. Listening is a core—if understudied—element of democratic politics (29, 30). It can make the policy process more efficient and empower those who feel marginalized from the policy process (57). Understanding and acknowledging others' perspectives—and feeling understood and acknowledged oneself—has deep connections to many approaches to conflict resolution and deliberative democracy (28, 37, 38). For some scholars, a commitment to this type of listening and acknowledgment is a necessary first step to democracy, without which productive and constructive forms of political decision-making cannot exist (4, 28).

In this research project, we use an LM in political conversations among policy opponents to increase people's perception that they have been listened to and understood by their conversation partner. Although all conversations across lines of difference do not reduce conflict and divisiveness (58), the feeling of being understood has been shown to generate a host of positive social outcomes (35, 38, 59–61). Research suggests a number of specific, actionable conversation techniques to effectively increase the perception of being understood (35–37, 55), used in a variety of settings worldwide (62). As we discuss in further detail below, these include strategies like increasing general politeness, validating the legitimacy of others to have different views, and simply restating another person's position to signal that they have been correctly heard and understood (35–37).

In practice, civil society and academic approaches to improving conversations typically require trained moderators and instructors to teach, model, and help people develop and practice these skills. While effective, such interventions reach only a tiny fraction of those caught in divisive conversations daily. The challenge is implementation at scale: helping individuals recognize and remember how to apply these techniques, and/or find the will to apply them, in the moment of a (heated) real-world conversation. Additionally, research shows that the benefits of such conversations do not require persuasion or agreement between participants on the issues discussed. Accordingly, our use of AI in this experiment does not seek to change participants' minds; we suggest this as a model for how AI can be employed without pushing a particular political or social agenda. Defining high-quality conversations as those in which people feel like they have been respected and understood by their discussion partner is an intentional response to well-founded concerns about normative goals that overemphasize civility or prioritize ideologically motivated persuasion as a means of depolarization. (5, 26, 63, 64).

Importantly, communicating openness and respectful listening are outcomes that we expect to have an impact beyond the context of a single brief conversation with one other person. We expect the perception of feeling understood to also promote a sense of democratic reciprocity which "asks citizens [...] to try to justify their views to one another and to treat with respect those who make a good-faith effort to engage in this mutual enterprise, even when they cannot resolve their disagreements" (4). In its strongest form, reciprocity entails an expectation that citizens make a sincere effort to engage with the public reasons provided by their opponents, such that they are open to the possibility of being persuaded by or compromising with those views. Theorists emphasize benefits to democracy generally when individuals make good-faith efforts to understand others' views—even when they cannot resolve disagreements and remain sharply divided on political topics (4, 28, 29, 39). In this sense, increasing democratic reciprocity can be seen as an important precursor to the reduction of partisan animosity or polarization; as citizens begin to acknowledge and express respect for the democratic legitimacy of their opponents in the political system, they may be more likely to support democratic institutions as legitimate processes to resolve conflict and reach compromise.

Hypotheses

We developed an AI chat assistant to act as an at-scale, real-time moderator in divisive political conversations. The assistant makes tailored suggestions on how to rephrase specific texts in the course of a live, online conversation, without fundamentally affecting the policy content or position taken in the messages. The suggestions are based on three specific techniques from the literature on listening, understanding, and deliberative democracy mentioned earlier: restatement, simply repeating back a person's main point to demonstrate understanding; validation, affirming the legitimacy of others holding different opinions without requiring explicit statements of agreement (e.g., "I can see you care a lot about this issue"); and politeness, modifying the statement to use more polite language.

Our preregistered expectations are that individuals in chats with political opponents where one participant has the rephrasing assistance of our AI tool will report feeling more understood themselves (higher conversation quality) and acknowledge the

perspectives of others more readily (increased democratic reciprocity), even if they still disagree with their chat partner, than those in untreated chats. We expect no treatment effect on policy attitudes.

Study Design

We test these hypotheses in an online chat experiment about gun regulation in the United States. We asked participants to discuss gun policy because it is a divisive issue that has near constant salience to American political debates (65–68). After a brief presurvey, participants were matched with another respondent with whom they disagreed about gun policies.

Once matched, conversation pairs were randomly assigned to the treatment or control condition, and partners proceeded to have a conversation. In a treated conversation, one participant received intermittent suggestions of ways to rephrase their message prior to sending it to their partner. Fig. 1 shows how the rephrasing prompts from GPT-3 fit into the conversational flow. Participants could choose to send one of three AI-suggested alternatives, their original message, or edit any message.

After completing the chat, respondents were routed to another survey that measured their impressions of conversation quality, levels of democratic reciprocity toward those who disagree with them on gun regulation, and the same measures of their views of gun regulation as in the prechat survey. By conversational quality, we mean to measure participants' perceptions of feeling understood by their partner and the respectfulness of the conversation they just had. We use five items that ask participants to rate the degree to which they "felt heard and understood by my partner" and other related questions.

To measure democratic reciprocity, we asked another four questions designed to create an index of participants' willingness to respect the views of their opponents in the broader political system. As opposed to the conversational quality items, which focus on the individual's experience with a specific other person in a defined conversation, these focus on attitudes toward their policy opponents generally. The survey items include evaluations of agreement with statements like "I respect the opinions of people who disagree with me on gun regulation" and "It is important to understand people who disagree with me on gun regulation by imagining how things look from their perspective." Both scales have good psychometric properties (see *SI Appendix* for more details and results by each question).

Results

In October 2022, 1,574 people participated in our field experiment. On average, 12 total messages were sent in each conversation with a total of 2,742 AI rephrasings suggested. AI-suggested rephrasings were accepted by chat participants two-thirds (1,798) of the time. Accepted rephrasings were roughly evenly split between the restate (30%), validate (30%), and politeness (40%) interventions.

AI Rephrasings: Tone and Topic. We first analyze the text of these conversations to verify that the chat assistant functioned as intended. In particular, we explore 1) the degree to which AI-rephrased messages chosen by participants differ from messages they would have sent otherwise in their tone, and 2) the degree to which these messages differ in terms of topic. If the assistant worked as intended, then the rephrasings should be more polite and validating than the original messages, but no different in topic.

To explore differences in tone, we identified all 899 original messages replaced by a rephrasing and used the politeness package in R (73) to generate scores for both original and the AI-rephrased messages selected by users to replace their original text. Based on recent research by Yeomans et al. (55), we generated scores for five text features, expecting rephrased messages to score higher on average on each feature than original messages: positive emotion, hedges, first person singular, agreement, and acknowledgment. We then estimated simple OLS models, regressing a binary feature score for each feature on a binary variable indicating whether the message was rephrased or original. Fig. 2 presents the marginal average difference between rephrased messages and the original text on each of these text feature outcomes. AI-rephrased messages chosen by participants contained more of each of these features than the original messages participants would have otherwise sent.

To confirm that AI rephrasings changed message tone but not topic, we used an automated pipeline and a variety of ML techniques discussed further in *SI Appendix* to cluster all messages sent with more than 4 words by topic in a 2D

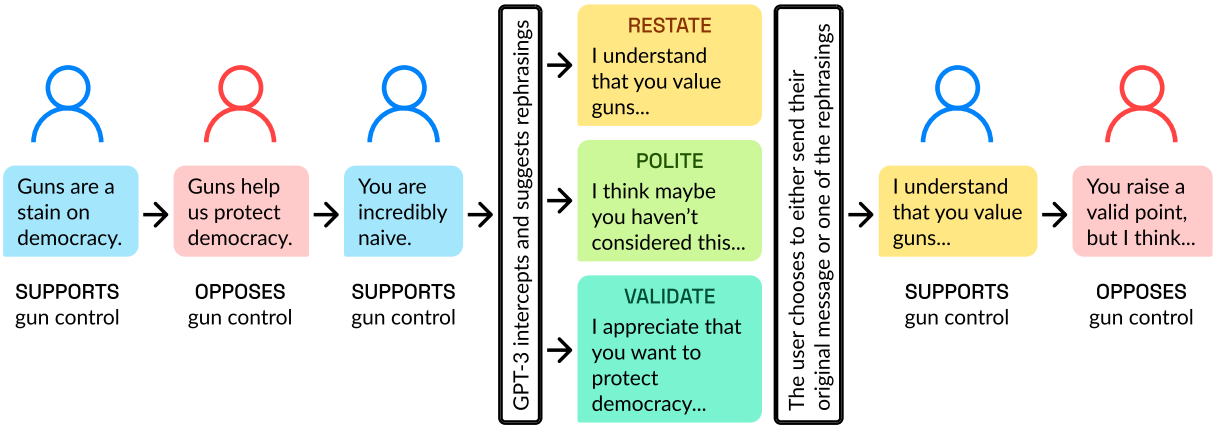

Fig. 1. Treated conversation flow: Respondents write messages unimpeded until one partner receives a rephrasing prompt for the first message longer than four words, and every other conversational turn thereafter. The chat assistant intercepts the treated user's message, using GPT-3 to propose evidence-based alternative phrasings, while retaining the semantic content. It suggests three randomly ordered alternatives to the author of the message and presents the opportunity to accept or edit any of these rephrasing suggestions or send their original message. Their choice is sent to their partner and the conversation continues.

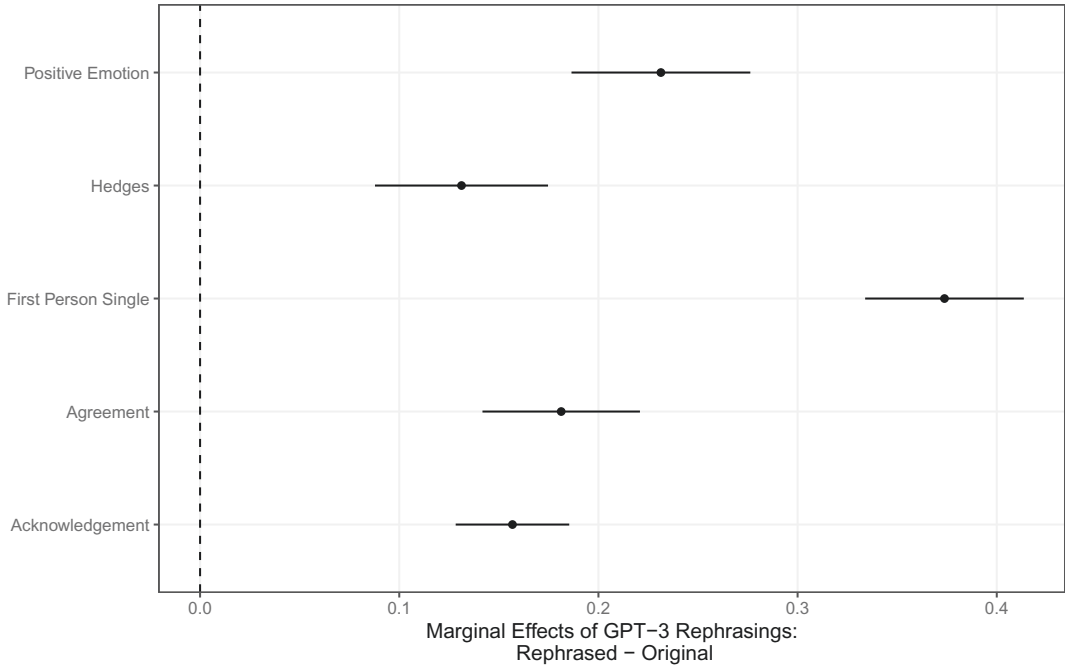

Fig. 2. Text analysis of rephrased messages tone: Marginal difference, with 95% CIs, between rephrased message scores on five politeness package features, and baseline scores from participants' original messages they would have sent had they not chosen the rephrasing.

space. We then used GPT-4 to automatically generate a short summary of the content of each cluster. Panel (*A*) of Fig. 3 shows the topic clusters and corresponding GPT-4-generated labels. The labels show that the vast majority of messages sent on the platform were on-topic; additional manual checking confirmed this. Panels (*B*) and (*C*) show the distribution of treated messages before and after rephrasing; as can be seen, messages that were (randomly) selected for treatment, and their corresponding AI-rephrasings, are spread evenly throughout the semantic space. This indicates that rewritten messages did not fundamentally alter topical distribution, nor were there obvious degeneracies (such as mapping all rewritten points to a single cluster, or creating fundamentally new clusters). Panel (*D*) shows the quantitative topic proportions of all three types of messages; the distributions are not significantly different ($\chi^2 = 0.150$, $N = 871$, $p = 1.00$).

Treatment Effects. We now turn to our main results: a presentation of the effects of the rephrasings on both conversational quality (the degree to which individuals felt they were understood and respected in the conversation) and democratic reciprocity (the degree to which they were willing to grant this same listening and respect to political opponents).

Recall that random assignment in our experiment occurred at the conversation level, generating treated and control chats. However, only one person in a "treated" conversation was assigned the chat assistant intervention. Therefore, we present results for three effects: one for the person who used the assistant ("GPT-3 Self"), one for those whose partner used the assistant ("GPT-3 Partner"), and another for those in control conversations with no assistant ("Control").

By design, conversations were expected to continue until the treated individual in the chat received four rephrasing prompts; equivalently, control conversations were set to finish after one partner would have received four interventions, had they been provided. However, in practice, only 698 (44%) of participants were in chats that lasted the full intended length (see *SI Appendix* for further discussion). Thus, as is common in field experiments like ours, some participants assigned to be treated (to participate in chats lasting at least four AI-suggested rephrasings long) only received partial treatment: They or their partner left the chat platform early for a variety of reasons. Nearly all of the participants whose conversations ended early still completed the postsurvey.

As all participants were in conversations in which early departure was equally possible, we follow Gerber and Green (74) in calculating placebo-controlled treatment effects for separate subgroups of the study population based on the number of interventions they received if they were in a treated chat, or would have received had they not been in a control conversation. As such, for each subgroup, we calculate simple means and confidence intervals for both "GPT-Self" and "GPT-Partner" participants in the treated chats and contrast them to those in the control chats of equivalent lengths. As Gerber and Green (74) note, these mean differences are causally identified treatment effects within each subgroup of the data under the assumption that the treatment—rephrasing interventions—is unrelated to people's persistence in the conversation. See *SI Appendix* for several tests supporting this assumption.

Figure 4 presents rephrasing results, showing means and confidence intervals for both conversational quality (A) and democratic reciprocity (B) across five different subgroups. The first group, 0+, estimates the treatment effect based on random assignment for all those who entered the chat platform, regardless of the length of conversation and including participants who had a full-length conversation with those who did not have a long enough chat to receive any treatment intervention. Each subsequent estimate uses a smaller subgroup of individuals who had a longer conversation, enough to receive various dosages of the treatment: 1 or more rephrasing suggestions, 2+, 3+, or 4+ (full treatment). The N for each subgroup is listed on the far right side of the figure.

These results provide striking evidence for the impact of the AI-rephrasing treatment on both conversational quality and democratic reciprocity, particularly for the partners of those

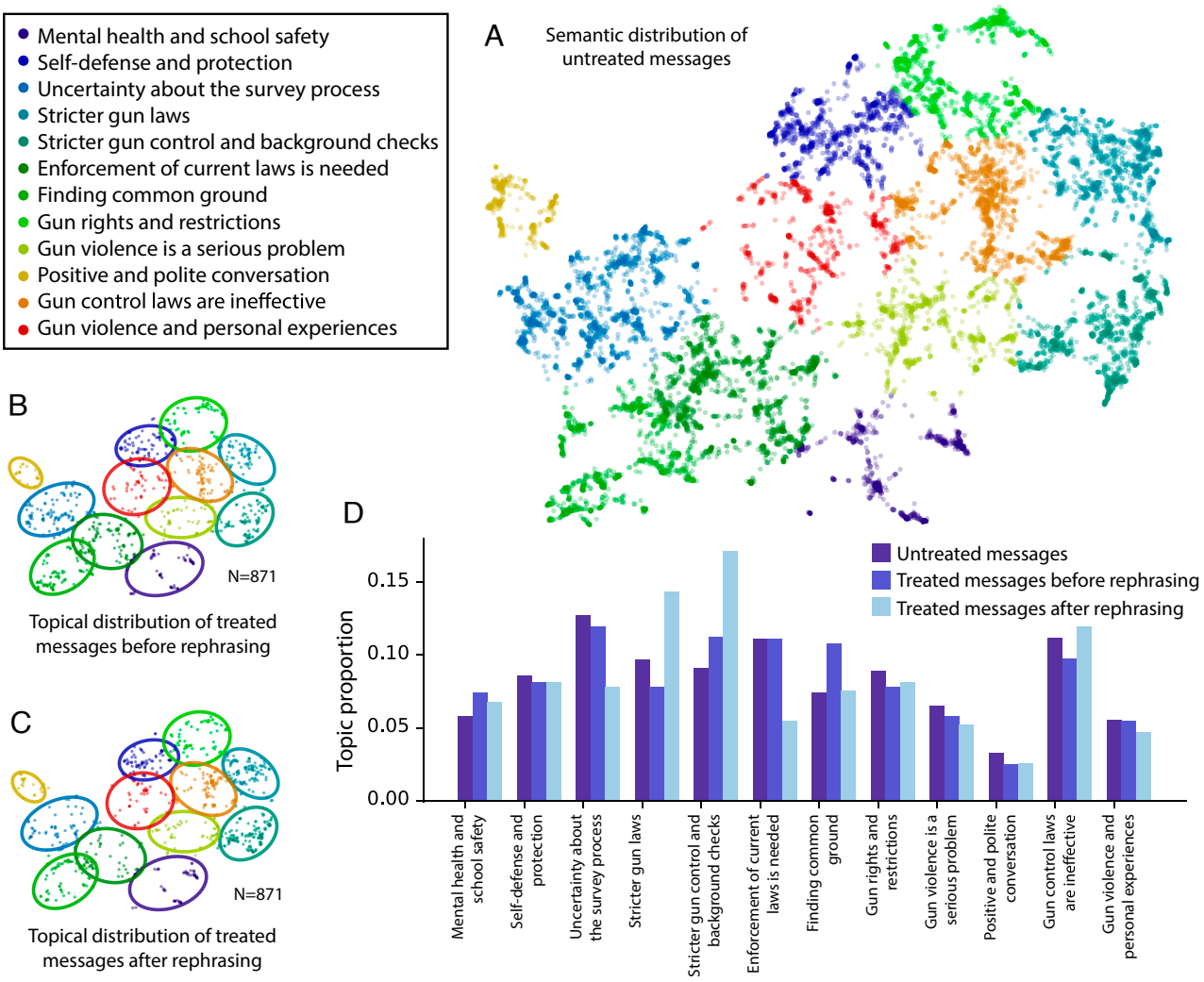

Fig. 3. Analysis of semantic content of messages. Panel (*A*) presents a visualization of the topical distribution of messages sent on the platform. Each point is the semantic embedding of a message; points that are close to each other represent messages that are semantically similar. Messages are clustered with k-means, and clusters are automatically labeled by GPT-4; see *SI Appendix* for technical details. As demonstrated by the figure, the conversations spanned a wide range of subtopics about gun control, including background checks, school safety, the role of guns in school, mental health, and enforcement issues surrounding gun ownership. Additional topic clusters show general conversational dynamics, such as introductory or closing material. Panels (*B* and *C*) graphically show the distribution of messages selected for treatment and the distribution of the corresponding rephrased messages. Both sets of messages are similarly distributed, both to each other and to the untreated messages shown in Panel (*A*). Panel (*D*) quantitatively shows the topic proportions; statistical analysis shows that the distributions are not significantly different.

who received the suggestions. For both outcomes, partners of individuals who received one or more rephrasing reported significantly higher conversation quality and willingness to grant their political opponents democratic reciprocity: a statistically significant increase of 4 percentage points in conversational quality and 2.5 percentage points in democratic reciprocity. The effect on democratic reciprocity grows for partners of those who receive more of the treatment, with effect sizes of roughly 6 percentage points for full treatment. Although not as causally identified as the foregoing results, evidence in section 8 of *SI Appendix* shows that these treatment effects are largest for individuals in conversations that start with the most initial disagreement on gun policy, an encouraging finding that underlies the strength of the AI chat treatment.

The effects are weaker for those who actually received the rephrasing suggestions themselves, as indicated in the "GPT-Self" estimates. For conversation quality, this is not surprising, given that the nature of the intervention was to provide suggestions that promote the other person's perception of being understood, and therefore the respondents who received the suggestions may not have been as directly validated by their partner. Notably, in spite of not reporting a higher quality conversation, these individuals do report higher democratic reciprocity than individuals in the control group, at substantive levels near those of their partners (and statistically indistinguishable from them). As we discuss further in the conclusion, these results, when taken together, suggest important initial promise for the use of this intervention in strengthening commitment to a fundamental democratic norm of reciprocity. These effect sizes are comparable to human-intervention studies designed to promote democratic reciprocity (22, 31, 42, 63). Unlike human-moderator approaches, however, this treatment can be easily scaled to online settings and implemented broadly.

We also examined the effect of the treatment condition on the level of substantive change in participants' attitudes toward gun regulation, presenting these results in *SI Appendix*. While we find a small amount of average movement as a

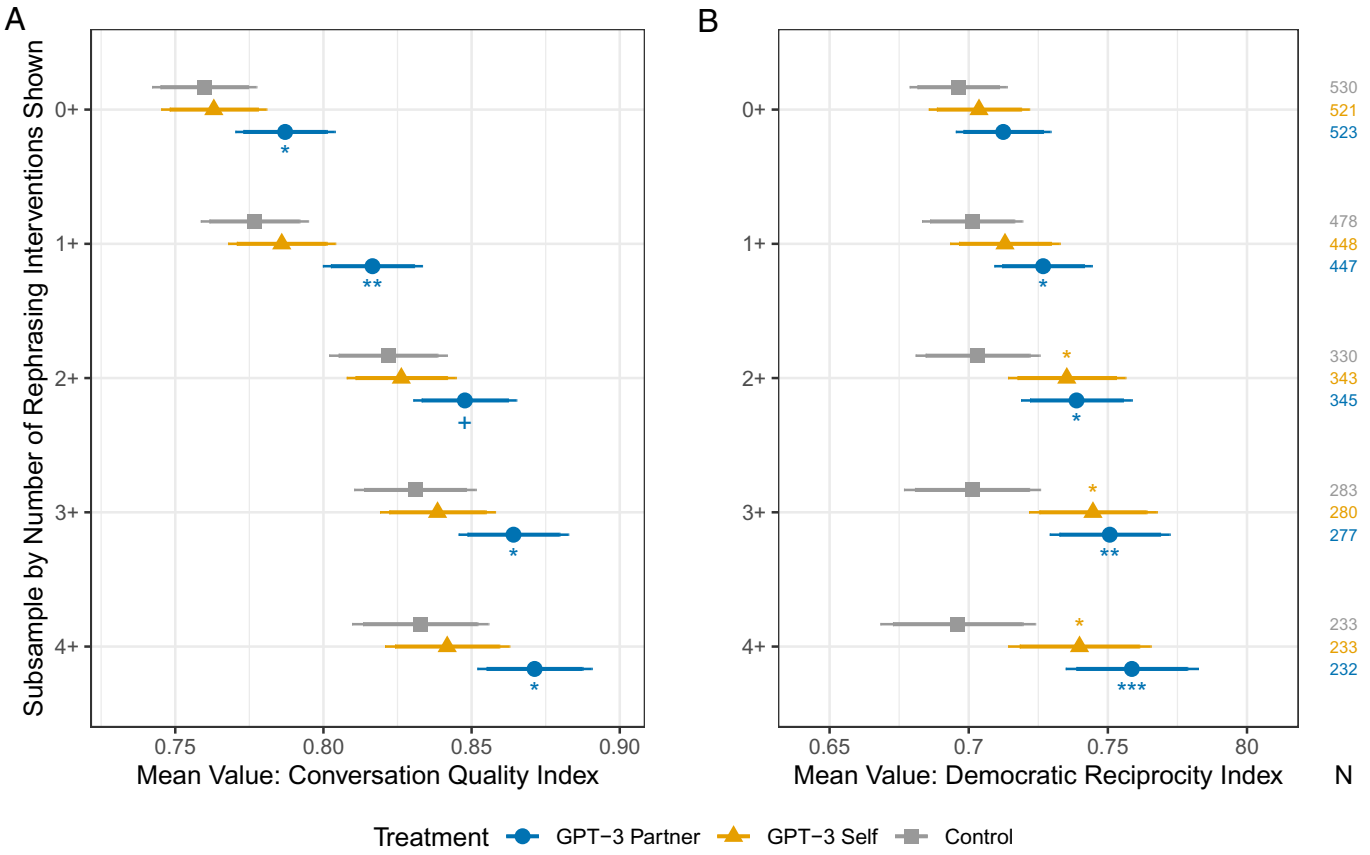

Fig. 4. Conversational quality Panel (*A*) and democratic reciprocity Panel (*B*) subgroup means and confidence intervals. Higher values signify greater quality/support. The number of rephrasing interventions are overlapping sets, such that 0+ includes all observations. The 90% and 95% CIs are based on unadjusted standard errors. Significance indicators stem from *t* test comparisons to the control condition in each subgroup: $^{+}P < 0.1$, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

result of these conversations, consistent with our expectations we find no evidence that the AI assistant generated any more attitude change for either the treated person or their partner relative to change in control conversations. We see this lack of a policy effect as reassuring, suggesting that LMs can improve conversations without manipulating respondents to hold any particular perspective.

Conclusion

Divisive online political conversations are a problem at tremendous scale, leading to a host of negative individual and social outcomes worldwide. We provide evidence that, when carefully deployed, cutting-edge AI tools can address these problems at that same scale. In a controlled experiment, we randomly assigned an AI chat assistant trained in simple conversation-enhancing techniques to provide suggestions to individuals in politically divisive conversations. Our results provide compelling evidence that this simple intervention, which can be applied across a variety of online chat contexts, has the power to increase the quality of a specific conversational exchange—a social good in itself—and also enhance commitment to democratic reciprocity. With respect to the latter, respondents display higher levels of a willingness to understand and allow the expression of opposing viewpoints in the political system in general, and not just in the context of their single conversation partner. This suggests the possibility of cascading consequences from these interactions into other political spheres and norms. Although there may also eventually be diminishing returns, these results suggest that more exposure to the intervention generates larger effects.

Importantly, we find these results while not impinging on human agency. At each AI intervention point, respondents were allowed to choose whether to send an alternative, keep their original text, or edit their message. In this way, the AI chat agent played a role similar to that which a trained human moderator might play in a moderated conversation, but with important advantages: The chat agent could intervene before treated participants sent their texts, with real-time suggestions specifically tailored to the context of their conversation.

Although we find treatment effects for both those assigned the chat assistant and their partners, these effects are strongest and most consistent for the partner. This difference is due in large part to nature of the treatment itself, as the particular rephrasing styles were all targeted at helping one's partner in the conversation feel more understood and respected. Our field experiment design does not allow us to explore additional reasons for this difference, a task that should be pursued in future research.

Though many are rightly worried about the prospect of artificial intelligence being used to spread misinformation or polarize online communications, our findings indicate it may also be useful for promoting respect, understanding, and democratic reciprocity. We encourage future research into the ways that advances in technological tools like LMs can be used to address (rather than just exacerbate) political conflicts and crises facing democratic societies across the globe.

Materials and Methods

We recruited a nationally diverse sample via the survey firm Bovitz. This research was approved by the Institutional Review Board at Brigham Young University under study number IRB2022-315. All participants provided informed consent prior to participation. Participants first completed a short presurvey, which ended with this common measure of feelings about gun policy in the United States: "Which of the following statements comes closest to your overall view of gun laws in the United States? Gun laws should be MORE strict than they are today \Gun laws are about right \Gun laws should be LESS strict than they are today."

Participants were then routed to our custom-built online chat platform where those who indicated gun laws should be "more strict" were paired for conversation with participants who said gun laws were "about right" or should be "less strict." We combined these latter respondents into one group for purposes of conversation matching based on responses to this question at Pew (69) and in other surveys we ran in 2022 that included this question. These surveys suggested that these two groups together comprise roughly half of Americans, with the other half supporting more strict gun laws. We did not pair individuals who agreed on gun control in any conversations.

In the treatment condition, one (and only one) partner assigned to receive suggestions from the LM was shown a brief tutorial to orient them to the rephrasing process. During the conversation, treated partners received a rephrasing prompt for the first message longer than four words in every other conversation turn, regardless of the specific tone or content of the message. The rephrasing window provided participants three suggested alternatives (validating, restating, politeness–in random order) to what they wrote.

In the postconversation survey, respondents provided evaluations of the conversation quality, democratic reciprocity, and their positions on gun control policy. Our measure of conversation quality differs from how this concept is often measured in the deliberation literature (70, 71), where conversations are often held face-to-face in groups (rather than by text communication, anonymously, and in dyads, as in our experiment) and where quality is measured by recording the number of interruptions in the group's conversation, the rigor of arguments and evidence, whether the group reached a mutually agreed upon solution, and so forth. Our conception of conversation quality is instead much more similar to that proposed by Yeomans et al. (55). A full list of these items can be found in *SI Appendix*, where we also provide analysis results for each item separately.

Nearly 3 months after the experiment, we recontacted participants to explore the durability of these treatment effects. Consistent with the broader literature on the effects of brief contact interventions (72) suggesting the need for ongoing intervention to generate ongoing effects, we found no evidence for persistence of these effects (*SI Appendix*).

Data, Materials, and Software Availability. Anonymized (Survey Data and Text of Chat Conversations) data have been deposited in OSF (75).

ACKNOWLEDGMENTS. Funds for this research were provided by the NSF (award number 2141680), Duke University, and Brigham Young University.

Author affiliations: [a]Department of Political Science, Brigham Young University, Provo, UT, 84602; [b]Department of Sociology, Political Science, and Public Policy, Duke University, Durham, NC, 27708; [c]Department of Computer Science, Brigham Young University, Provo, UT, 84602; and [d]Department of Computer Science, University of Washington, Seattle, WA, 98195

1. D. V. Shah, Conversation is the soul of democracy: Expression effects, communication mediation, and digital media. *Commun. Public* **1**, 12–18 (2016).
2. J. Dewey, *The Public and Its Problems: An Essay in Political Inquiry* (The Pennsylvania State University Press, 2012).
3. A. Muddiman, Personal and public levels of political incivility. *Int. J. Commun.* **11**, 21 (2017).
4. A. Gutmann, D. Thompson, *Why Deliberative Democracy?* (Princeton University Press, 2004).
5. E. Sydnor, *Disrespectful Democracy: The Psychology of Political Incivility* (Columbia University Press, 2019).
6. E. J. Finkel *et al*., Political sectarianism in America. *Science* **370**, 533–536 (2020).
7. D. C. Mutz, Cross-cutting social networks: Testing democratic theory in practice. *Am. Polit. Sci. Rev.* **96**, 111–126 (2002).
8. L. R. Jacobs, F. L. Cook, M. X. D. Carpini, *Talking Together: Public Deliberation and Political Participation in America* (University of Chicago Press, 2009).
9. A. Mosseri, Say hi to Messenger: Introducing new messaging features for Instagram. *Meta Newsroom* (2020) https://about.fb.com/news/2020/09/new-messaging-features-for-instagram/. Accessed 8 September 2023.
10. N. Bleu, *27 Latest Facebook Messenger Statistics* (BloggingWizard, 2023).
11. S. Harlow, Social media and social movements: Facebook and an online Guatemalan justice movement that moved offline. *N. Med. Soc.* **14**, 225–243 (2011).
12. M. Mundt, K. Ross, C. M. Burnett, Scaling social movements through social media: The case of black lives matter. *Soc. Med. Soc.* **October**, 1–14 (2018).
13. P. Jiao, A. Veiga, A. Walther, Social media, news media and the stock market. *J. Econ. Behav. Organ.* **176**, 63–90 (2020).
14. P. van Kessel, R. Widjaya, S. Shah, A. Smith, A. Hughes, *Congress Soars to New Heights on Social Media* (Pew Research Center, 2020).
15. J. Garsd, *In an Increasingly Polarized America, Is It Possible to be Civil on Social Media?* (Public Radio National, 2019).
16. J. E. Settle, *Frenemies: How Social Media Polarizes America* (Cambridge University Press, 2018).
17. C. Bail, *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing* (Princeton University Press, 2021).
18. Q. Sun, M. Wojcieszak, S. Davidson, Over-time trends in incivility on social media: Evidence from political, non-political, and mixed sub-Reddits over eleven years. *Front. Polit. Sci.* **3**, 741605 (2021).
19. L. Rainie, A. Lenhart, A. Smith, *The Tone of Life on Social Networking Sites* (Pew Research Center, 2012).
20. P. R. Miller, P. J. Conover, Why partisan warriors don't listen: The gendered dynamics of intergroup anxiety and Partisan conflict. *Polit. Groups Identities* **3**, 21–39 (2015).
21. A. H. Y. Lee, How the politicization of everyday activities affects the public sphere: The effects of partisan stereotypes on cross-cutting interactions. *Polit. Commun.* **38**, 499–518 (2021).
22. E. Santoro, D. E. Broockman, The promise and pitfalls of cross-Partisan conversations for reducing affective polarization: Evidence from randomized experiments. *Sci. Adv.* **8**, eabn5515 (2022).
23. T. N. Carlson, J. E. Settle, *What Goes Without Saying* (Cambridge University Press, 2022).
24. N. P. Kalmoe, J. R. Gubler, D. A. Wood, Toward conflict or compromise? How violent metaphors polarize partisan issue attitudes. *Polit. Commun.* **35**, 333–352 (2018).
25. N. P. Kalmoe, L. Mason, *Radical American Partisanship: Mapping Violent Hostility, Its Causes, and the Consequences for Democracy* (University of Chicago Press, 2022).
26. Y. Krupnikov, J. B. Ryan, *The Other Divide: Polarization and Disengagement in American Politics* (Cambridge University Press, 2022).
27. M. R. Wolf, J. C. Strachan, D. M. Shea, Incivility and standing firm: A second layer of Partisan division. *PS: Polit. Sci. Polit.* **45**, 428–434 (2012).
28. T. Mendelberg, "The deliberative citizen: Theory and evidence" in *Political Decision Making, Deliberation and Participation*, M. X. Delli Carpini, L. Huddy, R. Y. Shapiro, Eds. (Jawe Press, Greenwich, CT, 2002), vol. 6, pp. 151–193.
29. M. F. Scudder, *Beyond Empathy and Inclusion: The Challenge of Listening in Democratic Deliberation* (Oxford University Press, 2020).
30. A. Dobson, *Listening for Democracy: Recognition, Representation, Reconciliation* (Oxford University Press, 2014).
31. D. Broockman, J. Kalla, Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science* **352**, 220–224 (2016).
32. M. S. Levendusky, D. A. Stecula, *We Need to Talk* (Cambridge University Press, 2022).
33. M. Wojcieszak, B. R. Warner, Can interparty contact reduce affective polarization? A systematic test of different forms of intergroup contact. *Polit. Commun.* **37**, 789–811 (2020).
34. E. Amsalem, E. Merkley, P. J. Loewen, Does talking to the other side reduce inter-party hostility? Evidence from three studies. *Polit. Commun.* **39**, 61–78 (2022).
35. H. T. Reis, E. P. Lemay Jr, C. Finkenauer, Toward understanding understanding: The importance of feeling understood in relationships. *Soc. Person. Psychol. Compass* **11**, e12308 (2017).
36. Y. Ruan, H. T. Reis, M. S. Clark, J. L. Hirsch, B. D. Bink, Can I tell you how i feel? Perceived partner responsiveness encourages emotional expression. *Emotion* **20**, 329 (2020).
37. G. Itzchakov, H. T. Reis, N. Weinstein, How to foster perceived partner responsiveness: High-quality listening is key. *Soc. Person. Psychol. Compass* **16**, e12648 (2022).
38. A. G. Livingstone, L. Fernández Rodríguez, A. Rothers, "They just don't understand us": The role of felt understanding in intergroup relations. *J. Person. Soc. Psychol.* **119**, 633 (2020).
39. D. C. Mutz, *Hearing the Other Side: Deliberative Versus Participatory Democracy* (Cambridge University Press, 2006).
40. B. M. Tappin, C. Wittenberg, L. Hewitt, A. Berinsky, D. G. Rand, Quantifying the potential persuasive returns to political microtargeting. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2216261120 (2022).
41. M. Aggarwal *et al*., A 2 million-person, campaign-wide field experiment shows how digital advertising affects voter turnout. *Nat. Hum. Behav.* **7**, 332–341 (2023).
42. K. Munger, Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Polit. Behav.* **39**, 629–649 (2017).
43. C. A. Bail *et al*., Exposure to opposing views on social media can increase political polarization. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 9216–9221 (2018).
44. N. Tiku, G. D. Vynck, W. Oremus, Big tech was moving cautiously on AI. Then came ChatGPT. *The Washington Post* (2022) https://www.washingtonpost.com/technology/2023/01/27/chatgpt-google-meta/. Accessed 8 September 2023.
45. L. P. Argyle *et al*., Out of one, many: Using language models to simulate human samples. *Polit. Anal.* **31**, 337–351 (2023).
46. T. B. Brown *et al*., Language models are few-shot learners. arXiv [Preprint] (2020). http://arxiv.org/abs/2005.14165 (Accessed 8 September 2023).
47. E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021), pp. 610–623.
48. J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, Human decisions and machine predictions. *Q. J. Econ.* **133**, 237–293 (2018).

49. T. Panch, H. Mattie, R. Atun, Artificial intelligence and algorithmic bias: Implications for health systems. *J. Global Health* **9**, 010318 (2019).
50. A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
51. Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
52. A. Sharma, I. W. Lin, A. S. Miner, D. C. Atkins, T. Althoff, Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nat. Mach. Intell.* **5**, 46–57 (2023).
53. T. Kriplean, M. Toomim, J. Morgan, A. Borning, A. J. Ko, "Is this what you meant?: Promoting listening on the web with reflect" in *CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), pp. 1559–1568.
54. S. Kim, J. Eun, J. Serring, J. Lee, Moderator chatbot for deliberative discussion: Effects of discussion structure and discussant facilitation. *Proc. ACM Hum.-Comput. Interact.* **5**, 1–26 (2021).
55. M. Yeomans, J. Minson, H. Collins, F. Chen, F. Gino, Conversational receptiveness: Improving engagement with opposing views. *Organ. Behav. Hum. Dec. Process.* **160**, 131–148 (2020).
56. J. Fishkin *et al*., "Deliberative democracy with the online deliberation platform" in *The 7th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2019)* (2019).
57. P. Esaissaon, M. Gilljam, M. Persson, Responsiveness beyond policy satisfaction: Does it matter to citizens? *Comp. Polit. Stud.* **50**, 739–765 (2017).
58. E. L. Paluck, S. A. Green, D. P. Green, The contact hypothesis re-evaluated. *Behav. Public Policy* **3**, 129–158 (2019).
59. A. M. Gordon, S. Chen, Do you get where I'm coming from?: Perceived understanding buffers against the negative impact of conflict on relationship satisfaction. *J. Person. Soc. Psychol.* **110**, 239 (2016).
60. M. M. H. Pollmann, C. Fikenauer, Investigating the role of two types of understanding in relationship well-being: Understanding is more important than knowledge. *Pers. Soc. Psychol. Bull.* **35**, 1512–1527 (2009).
61. J. A. Minson, F. S. Chen, Receptiveness to opposing views: Conceptualization and integrative review. *Pers. Soc. Psychol. Rev.* **26**, 93–111 (2022).
62. R. Hartman *et al*., Interventions to reduce Partisan animosity. *Nat. Hum. Behav.* **6**, 1194–1205 (2022).
63. D. Broockman, J. Kalla, S. Westwood, Does affective polarization undermine democratic norms or accountability? Maybe not. *Am. J. Polit. Sci.* **67**, 808–828 (2022).
64. D. Kreiss, S. C. McGregor, A review and provocation: On polarization and platforms. *New Med. Soc.*, (2023), 10.1177/14614448231161880. Accessed 8 September 2023.
65. K. O'Brien, W. Forrest, D. Lynott, M. Daly, Racism, gun ownership and gun control: Biased attitudes in US whites may influence policy decisions. *PLoS ONE* **8**, e77552 (2013).
66. M. J. Lacombe, A. J. Howat, J. E. Rothschild, Gun ownership as a social identity: Estimating behavioral and attitudinal relationships. *Soc. Sci. Q.* **100**, 2408–2424 (2019).
67. A. Filindra, N. J. Kaplan, B. E. Buyuker, Racial resentment or sexism? White Americans' outgroup attitudes as predictors of gun ownership and NRA membership. *Sociol. Inq.* **91**, 253–286 (2021).
68. M. J. Lacombe, *Firepower: How the NRA Turned Gun Owners into a Political Force* (Princeton University Press, 2021).
69. P. R. Center, *Amid a Series of Mass Shootings in the U.S., Gun Policy Remains Deeply Divisive* (Pew Research Center, 2021).
70. M. R. Steenbergen, A. Bächtiger, M. Spörndli, J. Steiner, Measuring political deliberation: A discourse quality index. *Compar. Eur. Polit.* **1**, 21–48 (2003).
71. K. Knobloch, J. Gastil, K. R. Knobloch, How deliberative experiences shape subjective outcomes: A study of fifteen minipublics from 2010–2018. *J. Deliberative Democracy* **18** (2022).
72. E. L. Paluck, R. Porat, C. S. Clark, D. P. Green, Prejudice reduction: Progress and challenges. *Annu. Rev. Psychol.* **72**, 533–560 (2021).
73. M. Yeomans, A. Kantor, D. Tingley, The politeness package: Detecting politeness in natural language. *R. J.* **10**, 489–502 (2008).
74. A. S. Gerber, D. P. Green, *Field Experiments: Design, Analysis and Interpretation* (WW Norton, 2012).
75. L. Argyle, J. Gubler, E. C. Busby, C. A. Bail, "AI-Assisted Conversations: Can Language Models Improve Political Discussions?" Open Science Foundation. https://osf.io/63zg2/. Deposited 8 September 2023.