Anonymous Cross-Party Conversations Can Decrease Political Polarization: A Field Experiment on a Mobile Chat Platform

Aidan Combs,[1†] Graham Tierney,[2†] Brian Guay[5], Friedolin Merhout[6], Christopher A. Bail[1,3], D. Sunshine Hillygus[3,4], and Alexander Volfovsky[2]

[1]Department of Sociology, Duke University,
[2]Department of Statistical Science, Duke University
[3]Sanford School of Public Policy, Duke University
[4]Department of Political Science, Duke University
[5]Department of Political Science, Massachusetts Institute of Technology
[6]Department of Sociology, University of Copenhagen
[†]Shared first authorship.

September 22, 2022

Abstract

There is widespread concern that social media is driving political polarization. However, there is also increasing evidence that conversation between people with opposing political views—which social media can enable—may decrease animus between them. Do anonymous online conversations between people of different parties exacerbate or mitigate partisan polarization? We created a mobile chat platform to study the impact of such discussions. Our study recruited Republicans and Democrats in the United States to complete a survey about their political views. We later randomized them into treatment conditions where they were offered financial incentives to use our platform to discuss a contentious policy issue with an opposing partisan. We find that people who engage in anonymous cross-party conversations about political topics exhibit substantial decreases in polarization compared to a placebo group that was asked to write an essay using the same conversation prompts. These depolarizing effects were particularly strong among Republicans and correlated with the civility of dialogue between study participants. Our findings demonstrate the potential for well-designed social media platforms to mitigate political polarization and underscore the need for a flexible platform for scientific research on social media.

Introduction

Political polarization is one of the most pressing social problems of our era (Voelkel, Stagnaro, Chu, Pink, and Mernyk, Voelkel et al.; Baldassarri and Bearman, 2007; Boxell et al., 2020; DellaPosta et al., 2015; Finkel et al., 2020; Fiorina and Abrams, 2008; Iyengar et al., 2019; Mason, 2018a). Though scholars were once optimistic that social media could help bridge partisan divides by allowing people to connect with a broader range of others, many now worry that current popular platforms have instead increased ideological segregation and incivility instead (Bakshy et al., 2015; Barberá, 2015; Sunstein, 2002; Levy, 2021; Settle, 2018). Understanding if and under what conditions online interactions will either exacerbate or mitigate partisan divisions is critical to identifying effective pathways for addressing the challenges facing American democracy.

Past research has led to mixed expectations about the role of cross-party interactions on political polarization. Some studies suggest these interactions can exacerbate polarization and incivility (Papacharissi, 2004; Price, 2009). For instance, recent work concludes that Facebook use may increase political polarization (Settle, 2018) and exposure to ideologically uncongenial information can push partisans further apart (Bail et al., 2018). However, other research suggests that people will moderate their views when they engage with those with different perspectives because they come to recognize the value of alternative viewpoints (Fishkin and Luskin, 2005; Mutz, 2006; Zhang, 2019; Broockman and Kalla, 2016; Fishkin et al., 2021). For example, recent studies have found significant reductions in partisan animosity from engaging in cross-party conversations in-person (Levendusky and Stecula, 2021; Fishkin et al., 2021), on video chat (Santoro and Broockman, 2022), and over text (Rossiter, 2020).

In addition to these conflicting conclusions about the consequence of discourse in general, the literature is particularly unclear about the impact of cross-party conversations on social media. A longstanding concern about social media is that it removes the social pressures that maintain civility in in-person discourse, leaving people feeling less encumbered by social norms that guide physical interaction (Cheng et al., 2014; Kiesler et al., 1984). This might be especially the case when a platform allows for anonymity in online exchanges, which some research suggests exacer-

bates incivility and animus (Lowry et al., 2016; Lapidot-Lefler and Barak, 2012; Suler, 2004). This research suggests that interactions between opposing partisans on social media would inevitably increase political polarization. On the other hand, the lack of identity information available on many social media platforms could instead encourage people to focus on the content of conversation rather than the identity of the people they engage with (Berg, 2016; De Choudhury and De, 2014; Guilbeault et al., 2018; Strandberg and Berg, 2015) and may also make people feel more comfortable discussing alternative viewpoints honestly without fear of social repercussions (Mansbridge, 1983; Price, 2009; Sanders, 1997). It thus remains unknown if the observed depolarization effects of cross-party conversations in recent studies would replicate in the context of anonymous online exchanges.

Unfortunately, studying the causal effects of cross-party conversations in a social media environment presents significant methodological challenges. Observational analyses of cross-party deliberations on social media platforms are poorly suited to identify such effects because the processes that lead people into such interactions are not random (Wu et al., 2011; Eady et al., 2019; Guess, 2020; Bail et al., 2018). Moreover, platforms such as Facebook or Twitter are typically unwilling to randomize their users into the experiments necessary to test hypotheses about the potential effects because of corporate priorities to protect user privacy and ensure consistent user experience (King and Persily, 2019; Lazer et al., 2020; Mynatt et al., 2020; Hosseinmardi et al., 2021).

To address these issues, we developed our own mobile chat app to conduct a field experiment testing the impact of anonymous cross-party conversations on controversial topics. As described below, this platform mimicked the look of popular social media platforms, but allowed for manipulation of specific design features. We randomly assigned participants discussion partners from the opposite party and had them complete a sustained, text-based conversation about a political topic using our app. Using a mobile communication app adds external validity to our experiment because it allows participants to have conversations asynchronously on their mobile devices, as they would if they were to send a direct message via text, Facebook, or Twitter (Rossiter, 2020). Creating our

own app also allowed us to customize—-and randomly assign—features such as the information participants are given about their discussion partner. We varied whether participants were shown with a partisan label, thereby shaping the conversation context and assumptions(Groenendyk and Krupnikov, 2021; Mason, 2018b). Additionally, our design allows us to separate the conversation experience from our pre- and post- surveys. This ostensibly-unrelated design is an advance on other recent work on conversations that minimizes demand effects.

The research described below was approved by the Institutional Review Board at Duke University (protocol #2020-0326) and pre-registered on the Open Science Framework: https://osf.io/g97z5/. Recruiting respondents to download an app for an unrecognized mobile chat platform is difficult. We describe these challenges in our appendix, as well as deviations from our pre-registration statement which became necessary to address these challenges.

Research Design

In early February 2020, we hired the survey firm YouGov to recruit self-identified Democrats and Republicans for our field experiment. Respondents started with a survey about their political views, which included multiple questions measuring both ideological polarization and affective polarization—negative sentiment between members of rival political parties (Iyengar et al., 2019). Similar to previous research (Allcott et al., 2020), our outcome of interest is a global index, constructed from all twenty-one measures of polarization (Chronbach's α =.72), an approach that improves measurement precision (Ansolabehere et al., 2008) and recognizes the theoretical and empirical interaction of policy disagreement and affective polarization (Dias and Lelkes, 2021; Druckman et al., 2021a). This depolarization index is coded such that more positive values indicate people expressed less polarized views. We provide additional details about question wording and scale construction in the Appendix, where we also present treatment effects on the policy and affective depolarization sub-indices separately. The effects on each of these sub-indices are similar to the effects on the combined depolarization index.

After participants completed the survey, we randomly assigned them to treatment and con-

trol conditions and sent the treatment group a seemingly-unrelated invitation to download and test a mobile app for a new social media platform called DiscussIt for financial compensation. Our seemingly-unrelated design is an advance over previous work in that it both obscures the political nature of the experimental treatment and guards against demand effects. We made the app available for both iOS and Android devices via the Apple App Store and the Google Play Store. The invitation informed participants that DiscussIt is a chat platform where people anonymously discuss various topics, but instructed them not to disclose their name or personal information about themselves in order to allow conversations to "develop freely." There was no mention of politics in the recruitment dialogue. Respondents who downloaded the app ($n = 1201$) were assigned an "invite" code that—unbeknownst to them—automatically paired them with an opposing partisan. We provide details of the recruitment process and challenges, including a lower than expected yield of respondents, and evaluations of the resulting sample in the Appendix. Figure 1 shows the on-boarding app screens welcoming participants, which instructed them to complete 14 thoughtful replies with another DiscussIt user over the course of one week in order to receive the financial incentive. Next, they were randomly assigned to one of the two discussion topic areas—immigration or gun control—and given a gender-neutral pseudonym. After they were matched with an opposing partisan, respondents advanced to a chat screen to begin the conversation about the designated policy topic.

Respondents were further randomized into one of three sub-conditions in the treatment group: the discussion partner was either 1) correctly identified as an opposing partisan; 2) not labelled with a party; or 3) mislabelled to have the same party identification as the respondent. These sub-conditions were meant to provide further insight about how information about partisan identity might shape anonymous conversation about politics (Cohen, 2003; Dias and Lelkes, 2021; Mason, 2018a). In the placebo condition, respondents were asked to write an essay on immigration or gun control in response to the same prompts provide in the app. The aim of this baseline condition is to ensure all individuals in the study (both treated and not treated) have given roughly equivalent thought to the specific policy topic (Arceneaux and Wielen, 2017). See our Appendix for analyses

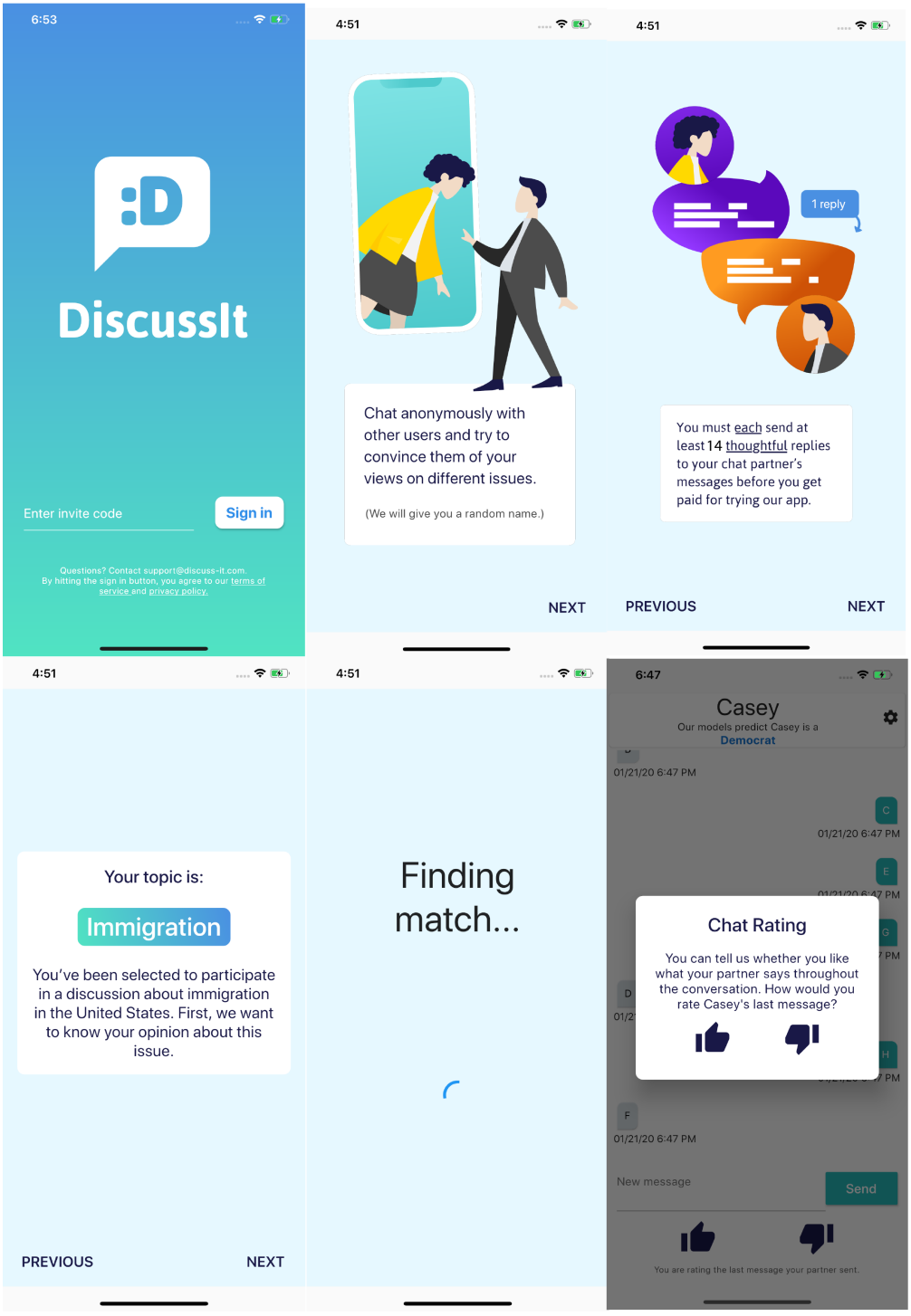

Figure 1: **Onboarding Images from Social Media Platform Created for Study.** After downloading the app, participants logged in and were guided through several on-boarding screens. They were then shown their randomly assigned discussion prompt about either immigration or gun control and matched with a partner from the opposing political party. Whether their partner's party affiliation was displayed correctly, incorrectly, or not at all was randomized at the time of matching. After matching, respondents entered the chat interface and could begin their conversation.

using a separate control condition where respondents were not asked to engage in any activity about the policy topic.

Several days after respondents in the treatment condition completed their chats, all respondents received another invitation to an ostensibly-unrelated survey measuring the key outcomes, thereby enabling within-subject assessment of the impact of our intervention. This survey included the same measures used in the pre-treatment survey but began with a set of distractor questions designed to mask the purpose of our study and discourage demand effects based on respondents' interpretation of the aims of our study. Finally, our app also collected the full text of all conversations, allowing further analysis of the potential mechanisms that shape anonymous cross-party interaction on social media.

Findings

Our analysis estimates the change between our pre-treatment and post-treatment depolarization index via a two-stage-least squares model designed to assess the Complier Average Causal Effect (CACE) of our intervention (Angrist et al., 1996). Comparisons between compliant and non-compliant respondents are reported in Appendix Section 4.3, and Intent-to-treat (ITT) results are presented in Appendix Section 4.4. There are no differences in effect statistical significance between the CACE and ITT results. Compliance was defined as installing our study's app and completing at least 10 exchanges with a member of the opposing party. As Figure 2 shows, we found respondents in our treatment condition exhibited sizable increases in our depolarization index even after relatively short conversations on our platform—equivalent to .22 standard deviations ($p < 0.05$).

Figure 3 reports the treatment effects by labelling condition (i.e. whether or how the partisanship of the respondent's conversation partner was labelled). As this figure shows, we observed the largest treatment effects when respondents' had information about their partner's political affiliation, regardless of whether it was accurate information. Respondents in these conditions exhibited a .25 (correct labels) and .26 (incorrect labels) standard deviation increase in our depolarization in-

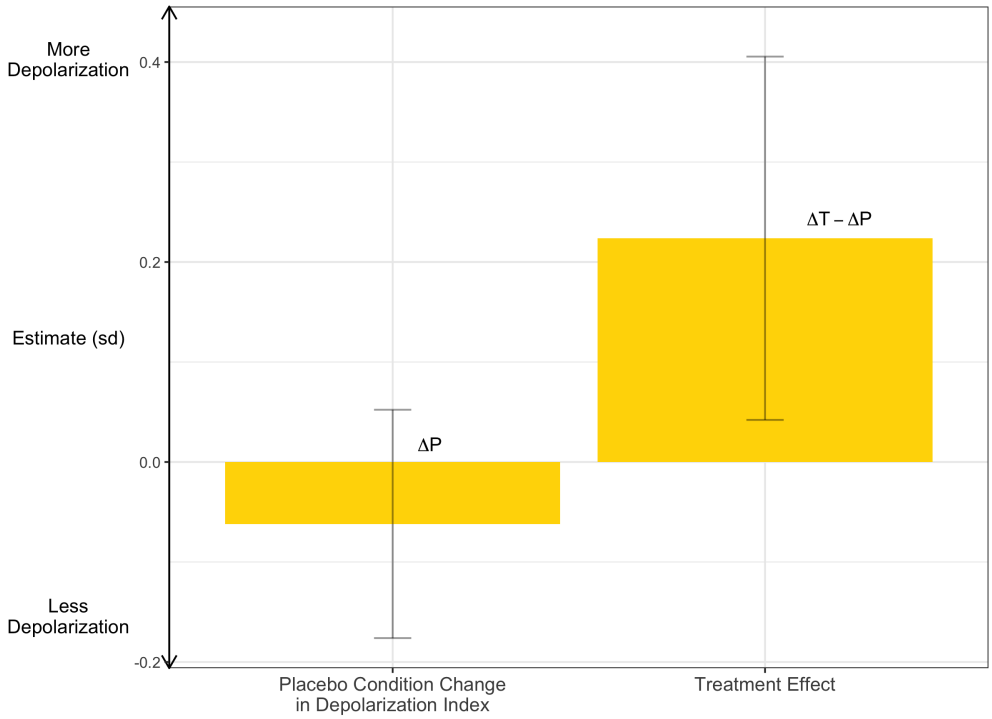

Figure 2: **Effect of Cross-Party Interaction on an Anonymous Chat Platform on Political Polarization**. Research participants (n=598 Democrats, n=603 Republicans) were randomized to a treatment condition where they were invited to test a new communication app that—unbeknownst to them—paired them with an opposing partisan to discuss immigration or gun control policy. The figure above describes the effect of this intervention compared to a placebo condition (n=218) in which respondents wrote an essay using the same conversation prompts on a depolarization index that is coded to be positive if respondents expressed less polarized attitudes, including issue positions and affective attitudes towards the opposing party (CACE=.224, $p < .05$). This finding indicates that using the study's anonymous chat platform to discuss a political issue with an opposing partisan depolarized political attitudes (equivalent to approximately .22 standard deviations on the depolarization index). Standard errors are clustered at the conversation-level. See Appendix Section 3 for full model details.

dex ($p < 0.05$ for both). Respondents in the sub-condition where political parties were unlabelled exhibited the smallest treatment effects (.15 standard deviations, $p = 0.15$). Figure 4 reports the treatment effects by the political party of the respondent. Republicans were more likely to depolarize than Democrats; the treatment effect was approximately .36 standard deviations ($p < 0.01$) for Republicans compared to .09 for Democrats ($p = 0.48$). In the appendix, we show that Republicans are *less* polarized before treatment than Democrats, suggesting that this difference does not occur because Republicans have more room to improve. We are unable to reject the null hypothesis that the treatment effect was the same for Republicans and Democrats ($p = 0.14$), but the point estimate of the treatment effect is nearly 4 times larger for Republicans and statistically significant, while the treatment effect for Democrats is not statistically significant.

Discussion & Conclusion

Our findings show that anonymous online cross-party conversations can help to depolarize the public. This finding is particularly noteworthy since there are currently so few examples of interventions that successfully reduce political tribalism on social media (Iyengar et al., 2019). Moreover, in contrast to many of the depolarization interventions being undertaken recently, we find discussions about contentious policy issues not only changed affect towards the other side (affective polarization), but also moderated attitudes on the issue being discussed (ideological polarization). And this depolarization occurred without explicit appeals for deliberation, empathy, or cooperation by the intervention–more closely reflecting the current social media environment.

The final sample sizes in our experiment allow for comparisons of each sub-condition to the placebo condition, but do not make possible precise comparisons across sub-conditions without strong parametric assumptions. Nonetheless, the observed patterns offer some suggestions as to the potential mechanisms underlying our findings. The smallest treatment effects are found among those whose partisan identifications were unlabeled. This pattern suggests that respondents in conversations with explicit partisan labels may more easily draw connections between the conversation and existing partisan stereotypes, experiences, and attitudes—a necessary precondition for

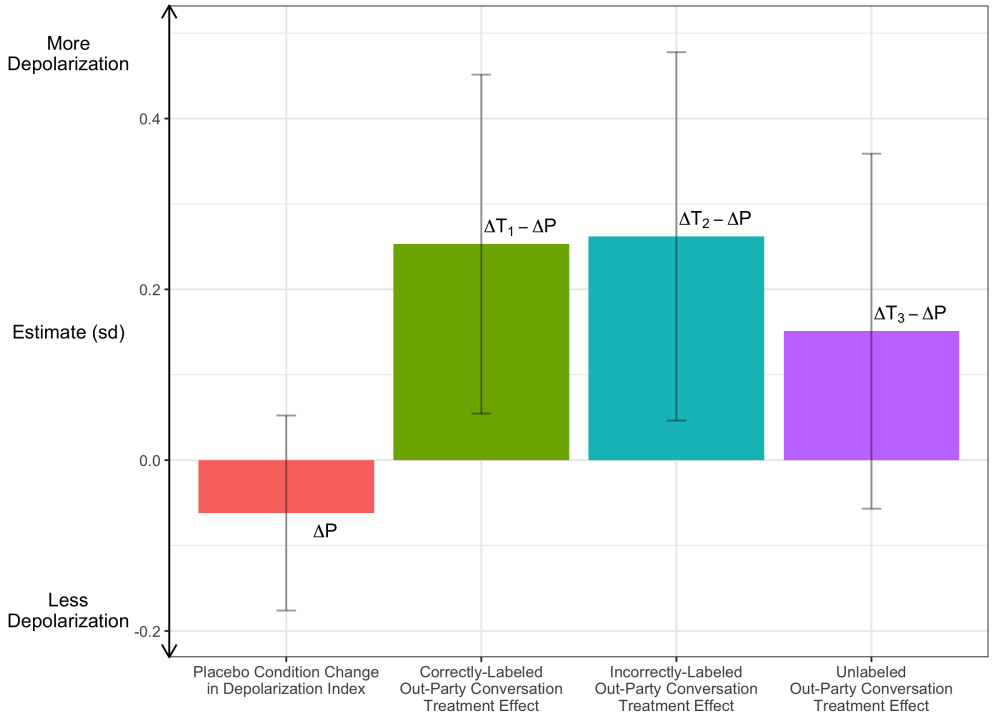

Figure 3: **Effect of Cross-Party Interaction on an Anonymous Chat Platform on Political Polarization According to Different Identity Cues**. This figure describes the treatment effects by different labelling conditions. The green bar describes the treatment effect for those whose discussion partner was accurately labelled as an opposing partisan. The teal bar describes those whose discussion partners were mislabelled as a co-partisan. The purple bar describes those whose party affiliation was unlabelled. As this figure shows, we observed large and significant effects for those in the correctly labelled and incorrectly labeled conditions (CACE=.253 and .262 respectively, $p < 0.05$ for both), and insignificant effects for those in the unlabelled condition (CACE=.151, $p = 0.15$). Standard errors are clustered at the conversation-level. See Appendix Section 3 for full model details.

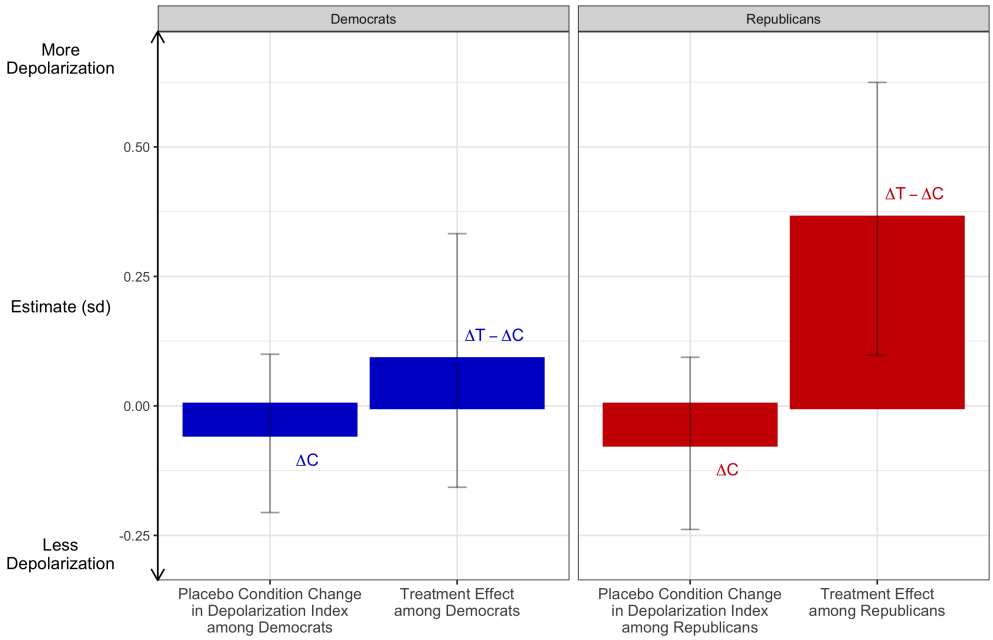

Figure 4: **Effect of Cross-Party Interaction on an Anonymous Chat Platform on Political Polarization, by Party**. This figure breaks down the treatment effects by the political party of respondents on our depolarization index. This panel shows the effect of our intervention was considerably stronger for Republicans (CACE=.361, $p < 0.01$) than Democrats (CACE=.088, $p = 0.48$). Traditional standard errors are reported because members of the same party do not interact. See Appendix Section 3 for full model specification details.

attitudinal change. This finding is consistent with recent research that finds polarization is fueled by wildly overestimated partisan stereotypes (Levendusky and Malhotra, 2016; Ahler and Sood, 2018; Moore-Berg et al., 2020; Paluck et al., 2019; Enders and Armaly, 2019; Ruggeri et al., 2021), so that a conversation with a member of the other party that contradicts prevailing stereotypes—revealing the actual extent of heterogeneity in partisan views—should help to depolarize (Rossiter, 2020; Fishkin et al., 2021; Levendusky and Stecula, 2021; Wojcieszak and Warner, 2020; Druckman et al., 2021b). It is telling that this is the case whether a respondent is told they are talking to someone from their outgroup or ingroup—consistent with recent research finding that ingroup social pressures contribute to polarization (White and Laird, 2020).

Initial analysis of the content of the conversations with respect to the civility of the conversation further reinforce this interpretation. We analyzed the civility of the messages exchanged using natural language processing techniques (Yeomans et al., 2018), an approach that has been used in previous research examining interpersonal exchanges about political disagreements (Yeomans et al., 2020). Figure 5 describes the "civility index" created via this analysis for respondent's chat partners over time. As this figure shows, people who experienced significant increases in our depolarization index tended to have conversation partners who used more civil language—particularly during the beginning of the conversation. In our appendix, we present models that show the relationship between chat partner's civility and depolarization is statistically significant—though civility was not incorporated into our randomization design and therefore cannot be considered an unambiguous causal factor.

It is also noteworthy that we observed significant heterogeneity in the treatment effects for Republicans and Democrats. A growing number of studies indicate political polarization in the United States has evolved in an asymmetric manner, driven primarily by Republicans. For example, Republican elected officials have increasingly taken more extreme positions in legislative votes than their counterparts in the Democratic party over the last forty years (Grossmann and Hopkins, 2016). Other studies indicate exposing Republicans to Democrats can make them more polarized, not less (Guilbeault et al., 2018; Bail et al., 2018). Yet these studies exposed people to

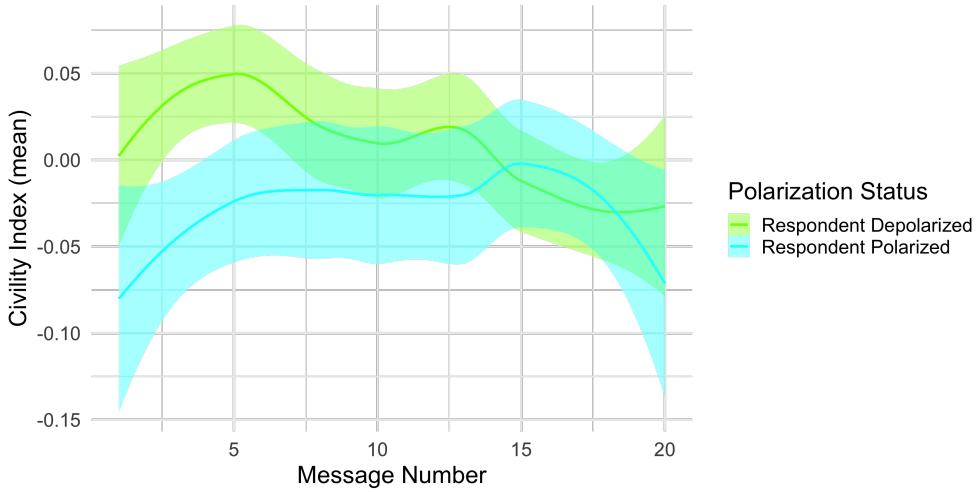

Figure 5: **Civility by Treatment Outcome (Over Time)**. Each chat produced on our platform was analyzed using natural language processing software to identify the frequency of civil exchanges. This figure plots the resultant "civility index" for each respondent's chat partner's messages according to their polarization status. Depolarized respondents are those whose depolarization index was more than one standard deviation above the mean; polarized respondents' depolarization index was more than one standard deviation below the mean. Respondents who depolarized tended to have chat partners who used more civil language—particularly at the beginning of the conversation. Time series data smoothed with LOESS function.

high-profile elites—or studied the effect of exposure in larger group settings or with less sustained interactions—than those used here. It may be that anonymous dyadic conversation between non-elites is a particularly important counterweight to asymmetric polarization by providing Republicans with the space to encounter views and stereotypes distinct from elite rhetoric and conservative media (Friedkin and Johnsen, 1999). In other words, people might find it easier to find common ground with a regular person than a political elite (Levendusky and Stecula, 2021)—about whom they have strong stereotypes generated by partisan media.

Our research has some notable limitations. It is possible that anonymous conversations might evolve quite differently on well-established social media platforms such as Facebook or Twitter where uncivil behavioral norms have already been established. It is also possible that anonymous conversation might unfold quite differently in a non-dyadic setting, where larger numbers of users interacting may generate peer influence dynamics that are quite different. For example, members of one party teaming up on the other—or feeling compelled to do so because of social norms. In addition, respondents in our study engaged in a substantial number of back and forth replies to each other. Cross-party interactions on Facebook or Twitter are often much shorter—and it may be that sustained conversation is necessary for depolarization. Finally, we asked users to engage in a focused discussion about a particular issue. Thus, our intervention may more closely resemble a platform such as Reddit, where discussions might be organized around a particular topic.

Nevertheless, our research demonstrates the promise of creating an open-source social media platform for scientific research. Such a tool could not only be used to conduct high quality field experiments that examine many other design features—but also avoid the many challenges of collaborating with social media platforms to conduct research on the increasingly urgent topic of political polarization. Perhaps most importantly, this research paradigm may inspire scientists, entrepreneurs, or existing social media companies to explore entirely new design features. Most of the dominant platforms evolved in a chaotic manner where interventions are tested in order to address emerging threats. In contrast, a new research agenda focused on scientifically testing the impact of social media design on polarization has the potential to test and develop a fuller range of

design features that might incentivize more positive behavior.

References

Ahler, D. J. and G. Sood (2018, jul). The Parties in Our Heads: Misperceptions about Party Composition and Their Consequences. *The Journal of Politics 80*(3), 964–981.

Allcott, H., L. Braghieri, S. Eichmeyer, and M. Gentzkow (2020). The welfare effects of social media. *American Economic Review 110*(3), 629–76.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association 91*(434), 444–455.

Ansolabehere, S., J. Rodden, and J. M. Snyder (2008). The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting. *American Political Science Review 102*(2), 215–232.

Arceneaux, K. and R. J. V. Wielen (2017). *Taming Intuition: How Reflection Minimizes Partisan Reasoning and Promotes Democratic Accountability*. Cambridge, United Kingdom ; New York, NY: Cambridge University Press.

Bail, C., L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. B. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky (2018, sep). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences 115*(37), 9216–9221.

Bakshy, E., S. Messing, and L. A. Adamic (2015, June). Exposure to ideologically diverse news and opinion on Facebook. *Science 348*(6239), 1130–1132.

Baldassarri, D. and P. Bearman (2007). Dynamics of Political Polarization. *American Sociological Review 72*, 784–811.

Barberá, P. (2015). Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis 23*(1), 76–91.

Berg, J. (2016, January). The impact of anonymity and issue controversiality on the quality of online discussion. *Journal of Information Technology & Politics 13*(1), 37–51.

Boxell, L., M. Gentzkow, and J. Shapiro (2020). Cross-country trends in affective polarization. *National Bureau of Economic Research (NBER Working Paper No. 26669)*.

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

Cheng, J., C. Danescu-Niculescu-Mizil, and J. Leskovec (2014). How Community Feedback Shapes User Behavior. *Proceedings of the Association for the Advancement of Artificial Intelligence*.

Cohen, G. L. (2003). Party Over Policy: The Dominating Impact of Group Influence on Political Beliefs. *Journal of Personality and Social Psychology 85*(5), 808–822.

De Choudhury, M. and S. De (2014). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. *Proceedings of the International AAAI Conference on Web and Social Media 8*(1).

DellaPosta, D., Y. Shi, and M. Macy (2015). Why Do Liberals Drink Lattes? *American Journal of Sociology 120*(5), 1473–1511.

Dias, N. and Y. Lelkes (2021). The Nature of Affective Polarization: Disentangling Policy Disagreement from Partisan Identity. *The American Journal of Political Science*.

Druckman, J. N., S. Klar, Y. Krupnikov, M. Levendusky, and J. B. Ryan (2021a, jan). Affective polarization, local contexts and public opinion in America. *Nature Human Behaviour 5*(1), 28–38.

Druckman, J. N., S. Klar, Y. Krupnikov, M. Levendusky, and J. B. Ryan (2021b, jun). (Mis-)Estimating Affective Polarization. *The Journal of Politics 40*(January).

Eady, G., J. Nagler, A. Guess, J. Zalinsky, and J. A. Tucker (2019). How Many People Live in Political Bubbles on Social Media? Evidence From Linked Survey and Twitter Data. *SAGE Open 9*, 1–21.

Enders, A. M. and M. T. Armaly (2019). *The Differential Effects of Actual and Perceived Polarization*, Volume 41.

Finkel, E. J., C. A. Bail, M. Cikara, P. H. Ditto, S. Iyengar, S. Klar, L. Mason, M. C. McGrath, B. Nyhan, D. G. Rand, L. J. Skitka, J. A. Tucker, J. J. Van Bavel, C. S. Wang, and J. N. Druckman (2020, oct). Political sectarianism in America. *Science 370*(6516), 533–536.

Fiorina, M. P. and S. J. Abrams (2008). Political Polarization in the American Public. *Annual Review of Political Science 11*(1), 563–588.

Fishkin, J. S. and R. C. Luskin (2005). Experimenting with a Democratic Ideal: Deliberative Polling and Public Opinion. *Acta Politica 40*(3), 284–298.

Fishkin, J. S., A. Siu, L. Diamon, and N. Bradburn (2021, nov). Is Deliberation an Antidote to Extreme Partisan Polarization? Reflections on "America in One Room". *American Political Science Review 115*(4), 1464–1481.

Friedkin, N. and E. Johnsen (1999, 01). Social influence networks and opinion change. *Advances in Group Processes 16*.

Groenendyk, E. and Y. Krupnikov (2021). What motivates reasoning? a theory of goal-dependent political evaluation. *American Journal of Political Science 65*(1), 180–196.

Grossmann, M. and D. A. Hopkins (2016, September). *Asymmetric Politics: Ideological Republicans and Group Interest Democrats* (1 edition ed.). New York, NY: Oxford University Press.

Guess, A. (2020). (Almost) Everything in moderation: New evidence on americans' online media diets. *American Journal of Political Science*.

Guilbeault, D., J. Becker, and D. Centola (2018, September). Social learning and partisan bias in the interpretation of climate trends. *Proceedings of the National Academy of Sciences 115*(39), 9714–9719.

Hosseinmardi, H., A. Ghasemian, A. Clauset, M. Mobius, D. M. Rothschild, and D. J. Watts (2021). Examining the consumption of radical content on youtube. *Proceedings of the National Academy of Sciences 118*(32).

Iyengar, S., Y. Lelkes, M. Levendusky, N. Malhotra, and S. J. Westwood (2019). The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science 22*(1), 129–146.

Kiesler, S., J. Siegel, and T. W. McGuire (1984). Social psychological aspects of computer-mediated communication. *American psychologist 39*(10), 1123.

King, G. and N. Persily (2019, 2019). A new model for industry-academic partnerships. *PS: Political Science and Politics 53*(4), 703–709.

Lapidot-Lefler, N. and A. Barak (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in human behavior 28*(2), 434–443.

Lazer, D. M. J., A. Pentland, D. J. Watts, S. Aral, S. Athey, N. Contractor, D. Freelon, S. Gonzalez-Bailon, G. King, H. Margetts, A. Nelson, M. J. Salganik, M. Strohmaier, A. Vespignani, and C. Wagner (2020, August). Computational social science: Obstacles and opportunities. *Science 369*(6507), 1060–1062.

Levendusky, M. S. and N. Malhotra (2016). Does Media Coverage of Partisan Polarization Affect Political Attitudes? *Political Communication 33*(2), 283–301.

Levendusky, M. S. and D. A. Stecula (2021, may). *We Need to Talk*. Cambridge University Press.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American economic review 111*(3), 831–70.

Lowry, P. B., J. Zhang, C. Wang, and M. Siponen (2016). Why do adults engage in cyberbullying on social media? an integration of online disinhibition and deindividuation effects with the social structure and social learning model. *Information Systems Research 27*(4), 962–986.

Mansbridge, J. J. (1983, June). *Beyond Adversary Democracy*. Chicago,: University Of Chicago Press.

Mason, L. (2018a). *Uncivil Agreement*. Chicago: University of Chicago Press.

Mason, L. (2018b). *Uncivil agreement: How politics became our identity*. University of Chicago Press.

Moore-Berg, S. L., L.-O. Ankori-Karlinsky, B. Hameiri, and E. Bruneau (2020, jun). Exaggerated meta-perceptions predict intergroup hostility between American political partisans. *Proceedings of the National Academy of Sciences 117*(26), 14864–14872.

Mutz, D. (2006). *Hearing the Other Side: Deliberative versus Participatory Democracy*. Cambridge, United Kingdom ; New York, NY: Cambridge University Press.

Mynatt, E., D. Watts, N. Bliss, A. Nelson, W. Pearson, and R. Rutenbar (2020). Harnessing the Computational and Social Sciences to Solve Critical Social Problems. Technical report, National Science Foundation, Alexandria, VA.

Paluck, E. L., S. A. Green, and D. P. Green (2019). The contact hypothesis re-evaluated. *Behavioural Public Policy 3*(2), 129–158.

Papacharissi, Z. (2004, June). Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society 6*(2).

Price, V. (2009). Citizens Deliberating Online: Theory and Some Evidence. In T. Davies and S. P. Gangadharan (Eds.), *Online deliberation: Design, research, and practice*, Chapter 2, pp. 37–58. Chicao, IL: Chicao University Press.

Rossiter, E. (2020). The consequences of interparty conversation on outparty affect and stereotypes. *Working Paper*.

Ruggeri, K., B. Većkalov, L. Bojanić, T. L. Andersen, S. Ashcroft-Jones, N. Ayacaxli, P. Barea-Arroyo, M. L. Berge, L. D. Bjørndal, A. Bursalıoğlu, V. Bühler, M. Čadek, M. Çetinçelik, G. Clay, A. Cortijos-Bernabeu, K. Damnjanović, T. M. Dugue, M. Esberg, C. Esteban-Serna, E. N. Felder, M. Friedemann, D. I. Frontera-Villanueva, P. Gale, E. Garcia-Garzon, S. J. Geiger, L. George, A. Girardello, A. Gracheva, A. Gracheva, M. Guillory, M. Hecht, K. Herte, B. Hubená, W. Ingalls, L. Jakob, M. Janssens, H. Jarke, O. Kácha, K. N. Kalinova, R. Karakasheva, P. R. Khorrami, Ž. Lep, S. Lins, I. S. Lofthus, S. Mamede, S. Mareva, M. F. Mascarenhas, L. McGill, S. Morales-Izquierdo, B. Moltrecht, T. S. Mueller, M. Musetti, J. Nelsson, T. Otto, A. F. Paul, I. Pavlović, M. B. Petrović, D. Popović, G. M. Prinz, J. Razum, I. Sakelariev, V. Samuels, I. Sanguino, N. Say, J. Schuck, I. Soysal, A. L. Todsen, M. R. Tünte, M. Vdovic, J. Vintr, M. Vovko, M. A. Vranka, L. Wagner, L. Wilkins, M. Willems, E. Wisdom, A. Yosifova, S. Zeng, M. A. Ahmed, T. Dwarkanath, M. Cikara, J. Lees, and T. Folke (2021, apr). The general fault in our fault lines. *Nature Human Behaviour*.

Sanders, L. M. (1997, June). Against Deliberation. *Political Theory 25*(3), 347–376.

Santoro, E. and D. E. Broockman (2022). The promise and pitfalls of cross-partisan conversations for reducing affective polarization: Evidence from randomized experiments. *Science Advances 8*(25), eabn5515.

Settle, J. E. (2018). *Frenemies: How social media polarizes America*. Cambridge University Press.

Strandberg, K. and J. Berg (2015, April). Impact of Temporality and Identifiability in Online

Deliberations on Discussion Quality: An Experimental Study. *Javnost-The Public 22*(2), 164–180.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior 7*(3), 321–326.

Sunstein, C. R. (2002, April). *Republic.com*. Princeton, N.J.: Princeton University Press.

Voelkel, J., M. Stagnaro, J. Chu, S. Pink, and J. Mernyk. Megastudy identifying successful interventions to strengthen americans' democratic attitudes.

White, I. K. and C. N. Laird (2020, June). *Beyond Adversary Democracy*. Princeton: Princeton University Press Press.

Wojcieszak, M. and B. R. Warner (2020, jun). Can Interparty Contact Reduce Affective Polarization? A Systematic Test of Different Forms of Intergroup Contact. *Political Communication*, 1–23.

Wu, S., J. M. Hofman, W. A. Mason, and D. J. Watts (2011). Who says what to whom on Twitter. *Proceedings of the 20th international conference on World wide web*, 705–714.

Yeomans, M., A. Kantor, and D. Tingley (2018). The politeness Package: Detecting Politeness in Natural Language. *The R Journal 10*(2), 489–502.

Yeomans, M., J. Minson, H. Collins, F. Chen, and F. Gino (2020). Conversational receptiveness: Improving engagement with opposing views. *Organizational Behavior and Human Decision Processes 160*, 131–148.

Zhang, K. (2019, apr). Encountering Dissimilar Views in Deliberation: Political Knowledge, Attitude Strength, and Opinion Change. *Political Psychology 40*(2), 315–333.

Acknowledgments

We thank Chris Goode for assistance with software development. This research was funded by the Provost's Office at Duke University and a Facebook Foundational Research Award.

Appendix for "Anonymous Cross-Party Conversations Can Decrease Political Polarization: A Field Experiment on a Mobile Chat Platform"

Contents

1 Introduction

This document describes all materials and methods for the article "Anonymous Cross-Party Conversations can Decrease Polarization," by Combs et al. All data, code, and the markdown file used to create this report will be available at this link on the Dataverse.

2 Additional Details about Research Design

2.1 Description of Mobile Chat Platform Developed for Study

Respondents assigned to the treatment group who consented to help test a new social media platform were sent instructions about how to install the DiscussIt app on their phone or tablet. DiscussIt was made available for both iOS and Android devices and was available in the Apple App Store and the Google Play Store. Respondents were assigned an invite code by the YouGov survey firm (See Section 2.2) and told to input it into the app's welcome screen. Unbeknownst to respondents, this code was used to pair them with a member of the opposing party for a conversation on the platform via a real-time matching system.

After inputting their invite code, users completed a short onboarding process in which they were told they would be assigned a pseudonym and had their task explained (see Figures 1-2). All respondents were assigned one of five androgynous names—"Jamie," "Jessie," "Taylor,"

"Quinn," or "Casey"—in order to prevent gender attribution effects. We chose these names by compiling a set of 26 androgynous names identified from prior literature (Lieberson et al., 2000) and publicly available Social Security Administration data. We then surveyed 135 people through Amazon Turk Prime about their perceptions of the typical gender, race, and age of people with that name. We chose names that were the most ambiguous, particularly with respect to gender.

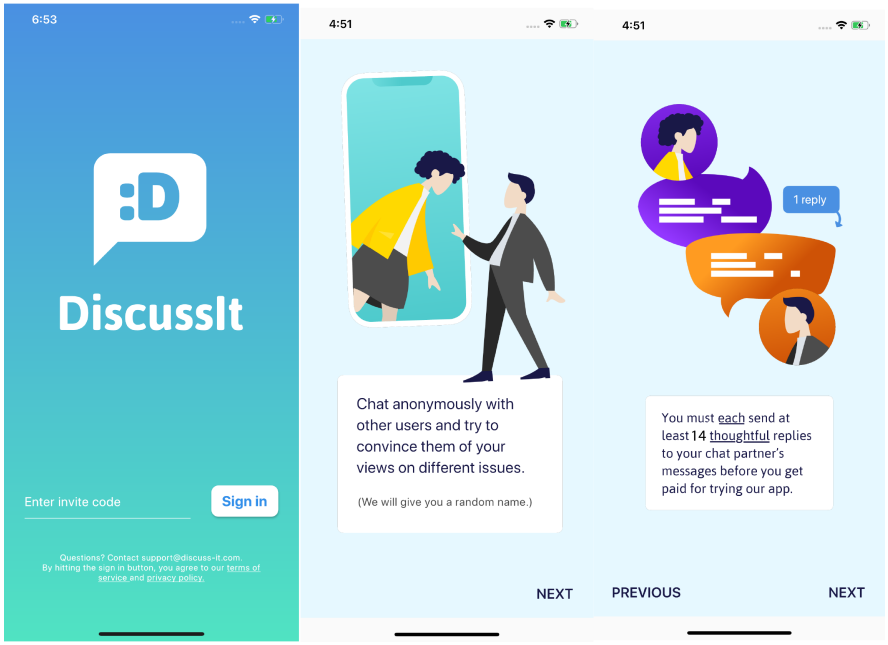

Figure 1: Onboarding Screenshots from Study's App.

After on-boarding, respondents were randomly-assigned to discuss either immigration or gun control. Respondents answered an in-app question asking their general view on the issue and were then matched to a conversation partner of the opposing party. If a match was immediately available, the respondent saw a message that presented the pseudonym of the person with whom they have been matched. If no match was immediately available, the screen instead said,"No match found yet! We haven't found a chat partner for you yet. But don't worry! We will keep searching and notify you when we do." When a match became available, the user

received a push notification on their phone and, upon opening the app, they were taken to the screen that shows the pseudonym assigned to the person with whom they have been matched.

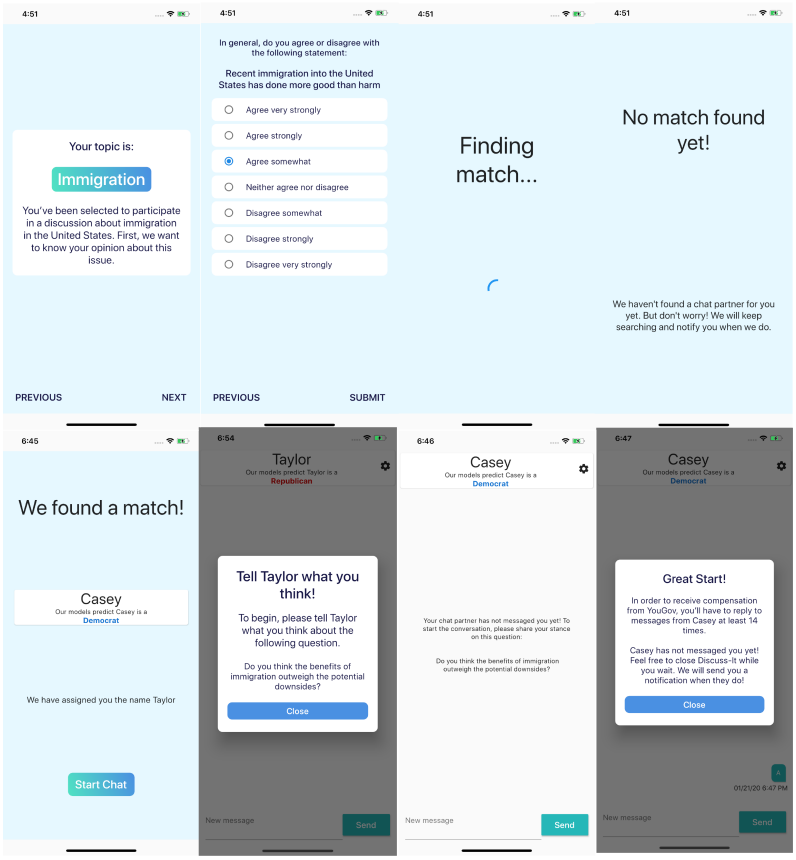

Figure 2: Chat Messaging Interface

Respondents in the treatment group were further randomized to one of three sub-conditions. Conditions varied (no party label, correct party label, incorrect party label) as to whether respondents were shown—in addition to the partner pseudonym—the political party of their discussion partner. As Figure 2 shows, the app used the following language to cue partisanship: "Our models predict [partner name] is a [Republican/Democrat]." We used this phrasing to make the nature of our intentional mislabelling more plausible in case respondents directly ask each other about their party background. Information about the chat partner's party was continually provided during the chat at the top of the screen (see Figure 2).

After being matched with a partner, users were directed to the main chat interface. The first to enter the chat received a pop-up window titled "Tell [partner name] what you think!" Below this title read the following text: "To begin, please tell [partner name] what you think about the following question:" Respondents then saw one of two questions designed to stimulate conversation. 1) "Do you think the benefits of immigration outweigh the potential downsides?"; or, 2) "Do you think the benefits of gun control outweigh the downsides?" After the window was closed, this question was shown in gray text on the background of the messenger window until one user sent a message. After closing the popup window, respondents could enter their response to this question in a text window at the bottom of the screen. They then received another pop-up message titled "Great start! In order to receive compensation from YouGov, you'll have to reply to messages from [partner name] at least 14 times." Below this was the text "[Partner name] has not messaged you yet! Feel free to close DiscussIt while you wait. We will send you a notification when they do!" The user received a pop-up notification on their phone when their discussion partner replied to them. When they tapped on this message, they were redirected to the chat window.

The second user to log in received a popup window that said "To begin, we asked [partner name] to share their stance on this question:" followed by the prompt question and then the text "Write back to complete your first interaction on DiscussIt!" After the user sent a message, they were shown a popup window titled "Great start! In order to receive compensation from YouGov, you'll have to reply to messages from [partner name] at least 14 times." When the users reached 3 exchanges with each other, they saw a third pop up window titled "Chat Rating" with text "You can tell us what you think about what your partner says throughout the conversation. How was [partner name]'s last message?" followed by the rating buttons (thumbs up/thumbs down).

While using the chat interface, study participants had access to a settings page via a gear icon

at the top of the page, which gave them the option to block their partner, contact the DiscussIt support team (staffed by us), see how many exchanges they have completed, view chat prompts, or view the discussion prompt question again. We also add a "chat prompts" window, which could be accessed via a button titled "Need a conversation starter?" and displayed the following text: "Think about some questions you could ask [partner name]. For example: How does this issue affect your life? What's one of the key reasons you have your position? Have you always had this position? Have you had any personal experiences that make this issue important to you?" These prompts were meant to serve as conversation starters and were optional for participants to use. If users had not completed 14 replies with their discussion partner with only two days left in the week in which they were meant to do so, the app sent a notification to the last respondent who did not reply to their conversation partner as follows: "We noticed you haven't replied to [name of partner] for a while and the deadline is coming up! If you don't respond soon, we'll have to assign them a new partner. Need ideas to keep the conversation going? Visit the settings menu for suggestions." Meanwhile, the other chat partner received the following message "We've noticed that [partner name] hasn't responded to you in a while and the deadline is coming up. Don't worry, we sent them a reminder. We will assign you a new partner if they don't respond soon."

We monitored the platform for harassment throughout the entire study period. We used Google's Perspective API to assign each message a toxicity score ranging between 0 and 1. This score can be interpreted as the probability a user would perceive the message as toxic. Users also had the option to report their conversation partner. When a message's toxicity score exceeded 0.8 or when a user submitted a report, the research team was notified and manually reviewed the conversation for abusive language. One conversation was shut down as a result.

Respondents were instructed to complete 14 replies to receive financial compensation, but we preregistered the compliance threshold at 10 replies to be considered treated and to receive

full compensation. If one or both of the conversation partners reached 10 replies but did not complete the final four replies within the time frame, they were still sent the completion message and in-app questions. If, however, one of the two did not reach 10 replies and their discussion partner was non-responsive, they were rematched with another person in the study who also had a non-compliant discussion partner, was a member of the opposite party, and was assigned to the same treatment condition as they were in before. The rematched pair were told they must complete a number of additional exchanges equal to 14 minus the minimum number the two respondents had before being rematched.

2.2 Respondent Recruitment

We hired the survey research firm YouGov to recruit a pool of potential respondents who were U.S. citizens at least 18 years of age, self-identified as either Republican or Democrat (including Independents who said they 'leaned' toward one party), used an iOS or Android smartphone or tablet, and self-reported a willingness to install an app on their phone or tablet. YouGov draws survey respondents from a large national non-probability panel using a combination of quota sampling and weighting to provide a sample that matches the demographic composition of the U.S. population. The study instruments, consent documentation and design were approved by the Duke IRB (protocol #2020-0326). We note that they language we used in recruiting respondents to download the app did not mention Duke university (to keep the post-treatment survey ostensibly unrelated to the pre-treatment survey) and did not mention politics.

From January 24-28, 2020, YouGov identified more than 7,000 panelists to screen for the eligibility criteria described above and invite to a 20-minute pre-treatment survey for an incentive of $12.50 in YouGov's online point system. Unfortunately, YouGov made a mistake in implementing the screening; they only asked those assigned to the treatment conditions whether they were willing to install an app on their phone or tablet (yielding 2514 positive responses in

the treatment group). Although those in the control and placebo conditions would not later be invited to actually download the app, a self-reported willingness to do so was a key inclusion criterion. This mistake created a risk of selection bias if not corrected and also meant that YouGov had not assigned sufficient number of individuals to those conditions to account for the need to exclude those unwilling to download an app. This error was not identified until after the completion of the field experiment (in late February), and we considered several approaches to address it. We initially considered recruiting additional individuals for the placebo and control conditions and correctly screening them for the inclusion criterion. Unfortunately, the political context had changed by that point—President Trump was acquitted in his impeachment trial and the COVID pandemic had emerged—creating potential confounders. We thus decided to recontact the individuals in our control and placebo conditions to ask their willingness to download an app, though it was several weeks after they had completed the initial survey. Those who did not express willingness to install an app or did not respond) were excluded to ensure equivalent inferential populations. With this correction, we ended up with the following number of respondents in the non-treatment arms who self-reported a willingness to download the app, completed the pre- and post-treatment surveys, and did not fail a data quality check—see Section 2.3: 218 respondents in the placebo condition and 144 respondents in the pure control condition. Unfortunately, the small number of respondents who self-reported a willingness to download the app among those assigned to the pure control condition leaves an insufficient sample size to make a precise comparison to this group. Results comparing treatment to the control are reported in Section 4.2, but we focus in the main text on comparisons to the placebo condition.

One day after they completed the pre-treatment survey, respondents in the treatment condition were informed they had been randomly selected for an opportunity to receive $17.25 in compensation for "testing a new social media platform called DiscussIt where people can

discuss issues." Of those invited, 52% followed through in downloading the app by the one-week deadline. Those who downloaded the app were further randomized into treatment sub-conditions and matched to an opposing partisan, as discussed in Section 2.1; 827 completed at least 10 exchanges with their conversation partner to be considered compliant with treatment.

On February 9, all respondents who completed the pre-treatment survey were invited to a post-treatment survey (and had until February 13th to complete it) for an additional $12.00 in YouGov's point system. The post-treatment survey to appear unrelated to the app invitation dialogue in order to avoid demand effects. Though the study included the same outcome measures from the pre-treatment survey, it was called a "health care study" and began with a series of "distractor" questions about health care and other issues not covered in the pretreatment survey in order to further mask the purpose of the study. The attrition rate between the pre- and post-treatment surveys was just 10.7%. One week after respondents completed the post-treatment survey, they were sent a debriefing statement that explained the IRB-approved deception used in the study, as well as the broader purpose of the study. This debriefing dialogue also informed respondents that some of their party affiliations had been intentionally mislabelled during the chats on our social media platform. Importantly, those in the placebo and control conditions were resurveyed about their self-reported willingness to download an app prior to the debriefing.

2.3 Sample Characteristics

In this section, we evaluate the characteristics of our sample of respondents. To summarize, the final sample on which our primary analysis is based are self-identified Democrats or Republicans in the YouGov panel who completed a pre- and post-treatment survey, expressed a willingness to download an app, and—for those in the treatment group—actually downloaded the app and completed at least 10 exchanges by the one-week deadline. Twenty-five respondents

were dropped from our analysis because of data quality issues. As preregistered, respondents were considered poor quality if they exhibited two of the following behaviors: 1) completed either the pre or post-treatment survey in less than five minutes (median survey completion time was 23.5 minutes); 2) straightlined more than six survey answers with a grid-style response; or 3) provided nonsensical responses to open-ended survey questions. Our final analytic sample is 1201 respondents in the treatment condition, 218 respondents in the placebo condition, and 144 respondents in the pure control condition.

Table 1 compares the demographic makeup of our sample with that of the population of adult U.S. citizens, as measured by the 2019 American Community Survey. Compared to the adult population, our sample is somewhat older and more educated, but similar in terms of gender and geographic distribution. We note, however, that our target population was self-identified partisans with an Android or iOS phone or tablet who expressed a willingness to download an app.

We next evaluate how the people who actually downloaded the app might have differed from the larger set of panelists invited to do so. First, we evaluate whether there are differences in the characteristics of those who self-reported willingness to download an app compared to those who were unwilling. Second, among those self-reported willingness, we evaluate any differences between those who actually downloaded the app and those who did not. The findings largely mirror the differences that have been observed between social media users and non-users (Marengo et al., 2020).

Column 1 of Table 2 provides coefficients of logistic regressions (for individuals for whom there are no missing demographics) where the response variable is whether an individual is willing to download the app. Column 1 combines the respondents in all conditions and includes an indicator for whether the question about willingness to install an app was asked originally or during the re-survey. Our second analysis in Column 2 of Table 2 reports the coefficients

Table 1: Comparison of sample to US adult citizen population. Population weighted means are calculated using the 2019 American Community Survey (noncitizens, those under 18, and those residing in Puerto Rico are excluded; $n = 2,449,277$). Sample means include all respondents who are used in our analysis; $n = 1,563$.

Variable	National Mean	Study Mean	p
Age	48.23	52.24	0.00
Female	0.51	0.55	0.01
No HS	0.10	0.02	0.00
High school	0.28	0.19	0.00
Some college	0.23	0.27	0.00
Two year degree	0.09	0.15	0.00
Four year degree	0.20	0.24	0.00
Postgraduate	0.11	0.13	0.11
New England	0.05	0.04	0.04
Middle Atlantic	0.13	0.13	0.37
East North Central	0.15	0.14	0.19
West North Central	0.07	0.07	0.76
South Atlantic	0.20	0.21	0.41
East South Central	0.06	0.07	0.04
West South Central	0.12	0.10	0.16
Mountain	0.08	0.08	0.29
Pacific	0.15	0.15	0.70

predicting who downloaded the app among those who reported willingness and were assigned to do so. The significant predictors—age, education, political knowledge, and missingness of family income—are not surprising. Older individuals are often more hesitant about technology, those who do not provide their income information often do so because of privacy concerns, and political knowledge and education are predictive of cooperativeness. We control for these predictors in subsequent analysis. These results offer reassurance that our population is defined by self-reported willingness to download an app rather than the behavior of doing so. This conclusion is further bolstered by a follow-up question asked of those respondents who said they were willing to download an app when asked during the pre-treatment survey did not ultimately do so. Common explanations included issues like memory limitations on phone and other technical difficulties.

2.4 Covariate Balance Check

Table 3 reports the mean levels of demographic variables for each of the 5 conditions. To determine whether differences between the placebo and treatment conditions are statistically significant, we regressed each demographic variable on an variable representing treatment assignment using OLS, where the reference level is the placebo condition. We observe statistically significant differences ($p < .05$) for 2 of the 18 variables: the placebo condition had higher levels of conscientiousness and openness to new experiences. However, when accounting multiple comparisons with a Bonferroni correction ($p = .05/18 = .0028$), neither of these differences are statistically significant. As a robustness check we present models estimating the effect of the treatment with demographic controls in Section 4.6.

Table 2: Response to app download questions (complete case analysis)

	Outcome in the logistic regression:	
	Consented to download	Downloaded following request
	(1)	(2)
treatment_assigned	−0.12 (0.08)	
gender	−0.08 (0.06)	0.11 (0.09)
age	−0.02*** (0.002)	−0.02*** (0.003)
college	−0.02 (0.06)	−0.23* (0.09)
white	−0.34*** (0.07)	0.10 (0.11)
knowledge_counter	−0.15*** (0.03)	0.18*** (0.05)
extraversion	−0.08*** (0.02)	0.01 (0.03)
agreeable	0.02 (0.03)	−0.07 (0.04)
consecientious	0.11*** (0.03)	−0.06 (0.04)
emotionally_stable	0.02 (0.02)	−0.001 (0.04)
open_new_experiences	−0.17*** (0.02)	−0.05 (0.04)
political_interest	0.12** (0.05)	−0.06 (0.07)
strong_partisan	0.10 (0.06)	−0.11 (0.09)
percent_other_party_exposure	0.001 (0.001)	0.002 (0.002)
democrat	0.10 (0.09)	0.07 (0.14)
sevenpoint_ideology	−0.002 (0.02)	−0.02 (0.04)
regionMidwest	−0.13 (0.09)	−0.08 (0.14)
regionSouth	0.01 (0.08)	0.10 (0.12)
regionWest	−0.10 (0.08)	−0.03 (0.13)
faminc_val	0.0000 (0.0000)	−0.0000 (0.0000)
faminc_notsay	−0.42*** (0.09)	−0.34* (0.16)
Constant	1.80*** (0.31)	1.42** (0.46)
Observations	6,217	2,294
Log Likelihood	−4,033.93	−1,549.10
Akaike Inf. Crit.	8,111.87	3,140.20

Note: *p<0.05; **p<0.01; ***p<0.001

Table 3: Demographic Covariate Balance Check

Covariate	Placebo	Correct Label	Incorrect Label	No Label	Control
age	52.06	51.52	52.90	50.92	51.09
agreeable	2.78	2.62	2.69	2.68	2.79
college	0.38	0.38	0.35	0.38	0.35
conscientious	2.57	2.34	2.36	2.39	2.50
democrat	0.53	0.50	0.50	0.50	0.53
emotionally_stable	2.94	3.03	3.04	2.88	3.08
extraversion	4.39	4.25	4.16	4.19	4.20
faminc_val	64554	67145	72076	73052	63198
faminc_notsay	0.08	0.08	0.06	0.09	0.12
gender	1.55	1.55	1.55	1.56	1.57
knowledge_counter	2.00	2.09	2.07	2.04	1.80
open_new_experiences	3.02	2.81	2.77	2.73	2.93
percent_other_party_exposure	26.99	26.75	28.91	29.65	27.48
political_interest	3.61	3.55	3.56	3.58	3.43
region	2.70	2.66	2.67	2.72	2.69
sevenpoint_ideology	3.90	4.10	4.16	4.04	4.03
strong_partisan	0.58	0.55	0.55	0.57	0.61
white	0.80	0.78	0.78	0.77	0.78

Cell values represent mean levels of demographic covariates in each condition.

2.5 Deviations from Pre-Registration Statement

As is perhaps inevitable given the complexity of the field experiment, we encountered some implementation issues that necessitated deviation from the preregistration. As previously noted, YouGov mistakenly failed to screen respondents in the control condition for their willingness to download an app (expressed willingness was an inclusion criterion even though they did not need to actually download the app). More generally, despite running an initial pilot, we (and YouGov) overestimated overall respondent yield; many of those who self-reported a willingness to download an app in the screener, did not actual do so when invited. Because screened respondents completed a pre-treatment survey outside the app prior to being issued an invitation to download the app, we were left with a smaller than anticipated sample size when we had a lower yield of survey respondents following through on the app download.

As a consequence, we focus attention on our overarching research question about the impact of anonymous cross-party engagement on the DiscussIt app on polarization. The final sample sizes for each sub-condition sufficiently powers a comparison to the baseline, but offer limited precision for comparisons to one another. We nonetheless report the results for all sub-conditions in the text and note that the smallest treatment effects are for those in the unlabeled condition. While there is too much uncertainty in the estimates to draw clear conclusions, the point estimates suggest fairly consistent depolarization effects across the board, with no indication of the backlash hypothesized for the correctly-labeled condition.

We would also note that the preregistration outlines other analyses/outcomes still to be explored from the DiscussIT app—for example, evaluating the nature and quality of the discourse in exchanges beyond the civility assessments discussed in Section 4.10. Our outcome of interest in the current analysis is a general polarization measure, as described below.

2.6 Dependent variable construction

The analysis reported in the manuscript relies on a global depolarization index ($\alpha = 0.72$), created as the average of an issue polarization and affective polarization indices. The affective polarization index was computed as the average of three sub-indices: affect towards the other party, outgroup trait stereotypes, social distance from the other party. All indices were computed as the average of observed 0-1 standardized responses for an individual and coded such that more positive values indicate people expressed less polarized views. We computed changes in each of the indices between the pre-treatment survey and the post-treatment survey. Difference measures were top-coded at the 95th percentile and bottom-coded at the 5th percentile. It is important to note that the issue questions used for an individual's issue polarization measure are based on the issue they were assigned (gun control or immigration).

We focus on a global depolarization index for a number of reasons. Following previous research in this area (Allcott et al., 2020), a global measure offers the broadest possible conceptualization for evaluating the impact of anonymous cross-party interactions on polarization. Recent research has shown there exists an interaction between affective polarization and policy disagreement (Dias and Lelkes, 2021; Druckman et al., 2021). For readers interested in the sub-indices, we report the results in Section 4.5, which shows the observed effects are not very different from one another. Future analyses of individual sub-indices will also necessitate engagement with the rich and nuanced literature associated with each of the individual measures (e.g., social distance, stereotypes, thermometer). For example, recent work has debated the use of partisan stereotypes as measures of affective polarization (Druckman et al., 2018). Our approach also improves measurement precision (Ansolabehere et al., 2008), which is especially helpful in light of measurement differences across the items. For example, the social distance items use 5 response categories that are asymmetric ("extremely well" to "not at all well"), whereas the issue attitude items rely on a 7 response options that are symmetric ("strongly

support" to"strongly oppose").

The full pre-treatment and post-treatment survey instruments are available on the project Dataverse page. The question wording for the individual items as well as the Chronbach's α for each of the sub-indices are reported below.

The Ideological Polarization Index was created based on the particular issue topic assigned ($\alpha = 0.92$ for immigration; $\alpha = 0.92$ for gun control) and included the following individual survey items:

Please indicate whether you would support or oppose the following proposals about gun policy. [Support Strongly, Support Moderately, Support Slightly, Oppose Slightly, Oppose Moderately, Oppose Strongly]

- Requiring background checks for all gun sales
- Preventing people with mental illnesses from purchasing guns
- Banning assault-style weapons
- Allowing people to carry concealed guns in more places
- Allowing teachers and school officials to carry guns in K-12 schools
- Barring gun purchases by people on the federal no-fly or watch lists

How much do you agree or disagree with the following statements? [Agree very Strongly, Agree Strongly, Agree Somewhat, Neither Agree nor Disagree, Disagree Somewhat, Disagree Strongly, Disagree very Strongly]

- The benefits of gun control outweigh the potential downsides.
- The federal government should make it more difficult for people to buy a gun.

- More restrictions on handguns would decrease violent crime by making it harder for criminals to get handguns.
- More restrictions on handguns would increase violent crime by making it harder for law-abiding citizens to defend themselves with handguns.

Please indicate whether you would support or oppose the following proposals about immigration policy:

- Changing the U.S. constitution so that children of unauthorized immigrants do not automatically get citizenship if they are born in this country
- Building a wall on the U.S. border with Mexico
- Separating the children from those parents caught crossing the border illegally
- Allowing unauthorized immigrants currently living in the United States to remain in the country and eventually qualify for citizenship?
- Requiring that all immigrants to the United States learn to speak English

How much do you agree or disagree with the following statements? [Agree very Strongly, Agree Strongly, Agree Somewhat, Neither Agree nor Disagree, Disagree Somewhat, Disagree Strongly, Disagree very Strongly]

- The benefits of immigration outweigh the potential downsides
- The federal government should permit fewer immigrants from foreign countries to come to live in the United States
- Having an increasing number of people of many different races, ethnic groups and nationalities in the United States makes this country a better place to live

- Immigrants mostly hurt the economy by driving wages down for many Americans
- Illegal immigration increases the crime rate in the U.S.

The Affective Polarization index was computed as the average of three sub-indices: affect towards the other party, outgroup trait stereotypes, social distance from the other party. The question wording for each sub-index is included here:

Social Distance ($\alpha = 0.79$): How well do each of the following statements describe you? [Extremely Well, Very Well, Moderately Well, Slightly Well, Not Well at all]

- I would be unhappy if someone in my immediate family married a [other party]
- I would be unhappy if I had to spend time socializing with a [other party]
- I would be unhappy if I had to work closely with a [other party] on a job.

Outgroup trait stereotypes ($\alpha = 0.83$): How strongly do you agree or disagree with the following statements? [Agree very strongly, Agree strongly, Agree somewhat, Neither agree nor disagree, Disagree somewhat, Disagree strongly, Disagree very strongly][Name of Other Party Members] are Intelligent, Open-Minded, Generous, Hypocritical, Selfish, Mean

Thermometer Rating ($\alpha = 0.83$): Next, we'd like to ask you some questions about what you think of different political figures and groups. Please rate each of the following on a feeling thermometer that runs from 0 to 100 degrees, where a rating above 50 is favorable and a rating below 50 is unfavorable. Please enter a whole number that ranges from 0 to 100 for [Democrats, Republicans]

3 Model Specification

The base model specification that we use in main paper and appendix is the following:

$$Y_i = \beta_0 + \beta_1 Compliance_i + \boldsymbol{\beta}_2 \mathbf{X}_i + u_i$$

$$Compliance_i = \gamma_0 + \gamma_1 Treatment_i + \boldsymbol{\gamma}_2 \mathbf{X}_i + v_i$$

Y_i is the outcome variable, typically the depolarization index, $Compliance_i$ is a dummy variable with 1 indicating unit i is treated and fully compliant (completes all exchanges with their conversation partner), and $Treatment_i$ is a dummy variable where 1 indicates unit i was assigned to treatment. $\mathbf{X}_i$ is a vector of pre-treatment covariates for unit i. Greek letters are coefficients. We assume u_i and v_i are uncorrelated. Our main results omit the covariates $\mathbf{X}_i$ for ease of interpretation and due to missing values in select covariates, but the conclusions are invariant to their inclusion or exclusion; all results are reported in Tables 6 and 7.

In models that include both treated Democrats and treated Republicans, the standard independent error term assumption could be violated. Treated individuals interact with their partner, so their outcomes might be correlated. We thus use cluster-robust standard errors to account for this dependence. For details on the method, see reference (Cameron and Miller, 2015). This procedure widens confidence intervals (and increases p-values). As such, we do not report traditional standard errors for our results; all estimates that are statistically significant with clustered standard errors will remain significant without such clustering. When regressions are split by party, as in main paper Figure 4, traditional standard errors are reported because individuals in each regression do not interact.

However, we note that outcomes for treated units in fact appear quite independent and clustering may not be necessary. The Pearson's correlation coefficient between the Democrat's and the Republican's depolarization across all conversations is only 0.03 (95% CI from -0.06 to 0.12). Moreover, we are measuring individual outcomes several weeks after the conversations were completed. One person improving his or her views regarding the other party does not necessarily imply that his or her conversation partner also changed his or her views.

Models are estimated with the AER package (Kleiber and Zeileis, 2008), and 95% confidence intervals accounting for clustered standard errors are generated using the sandwich (Zeileis et al., 2020; Zeileis, 2006) and lmtest packages (Zeileis and Hothorn, 2002).

4 Additional Analyses

4.1 Willingness to download an app among placebo and pure control participants

YouGov's implementation mistake meant that we had to recontact respondents in our control and placebo conditions to get their self-reported willingness to download an app. Given this deviation from the planned design, we re-estimate the models including all respondents assigned to the placebo condition, irrespective of their reported willingness to download an app. The substantive conclusions do not change. We present these results for completeness and note that it is usually insufficient to control for covariates to adjust for potential confounding as might happen when ignoring an inclusion criterion question.

4.2 Comparison to Pure Control

Within the main text we report our treatment effects relative to a placebo condition, in which respondents were asked to discuss their opinions on their randomly-assigned issue (immigration or gun control) in essay form, rather than in an anonymous conversation with someone from the opposing party. The advantages of placebos are well-established in the literature (Porter and Velez, 2021; Nickerson, 2005). In the case of our field experiment, the placebo baseline ensures that any observed effects are not simply an artifact of respondents giving more thought to the specified policy topic than they would otherwise do if not in the experiment (Arceneaux and Wielen, 2017). For robustness, we preregistered two control conditions—the placebo as well as a pure control in which respondents completed the pre-treatment and post-treatment surveys

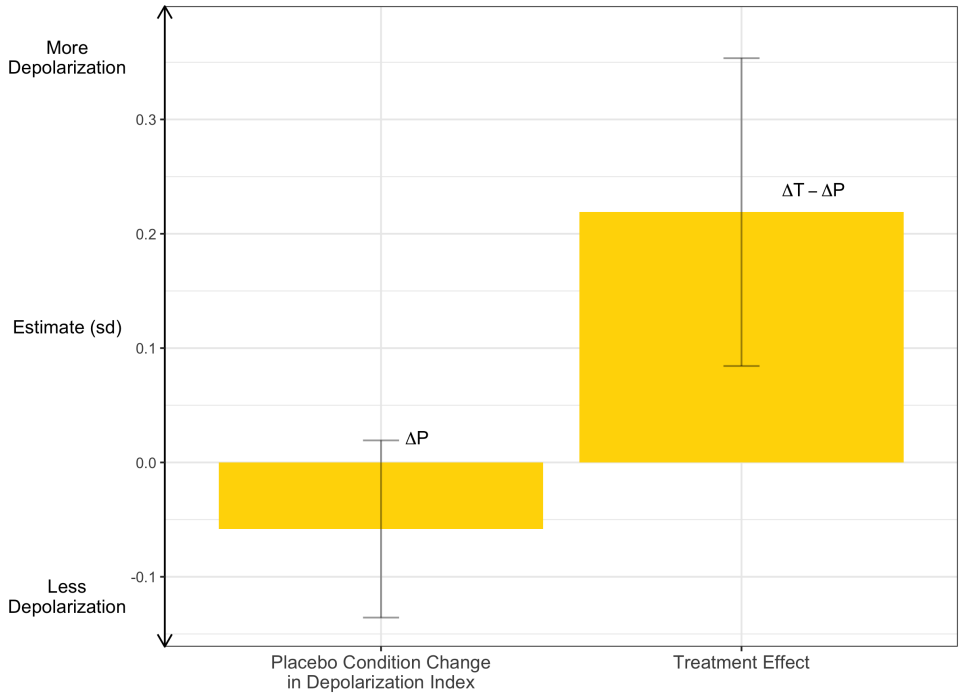

Figure 3: Out-Party Discussion Treatment Effects Compared to All Placebo Respondents

without any additional activities/discussion of the policy topic. Unfortunately, due to the error by our survey firm discussed above, the final sample size of those in the control condition is especially small. The treatment effects with that control group are reported in Figures 6 to 8 below, mirroring Figures 2 to 4 in the main text. The substantive pattern remains unchanged, but the results are noisier given the smaller sample size.

4.3 Addressing Non-Compliance

In the process of any complex intervention, some individuals do not fully comply with the treatment regime. In this section we evaluate whether specific groups were more likely to comply than others. Specifically, we searched for variables that might predict whether respondents 1) sent one message on the DiscussIt platform compared to none and 2) completed the required 10 exchanges to be considered fully compliant with treatment compared to less than 10.

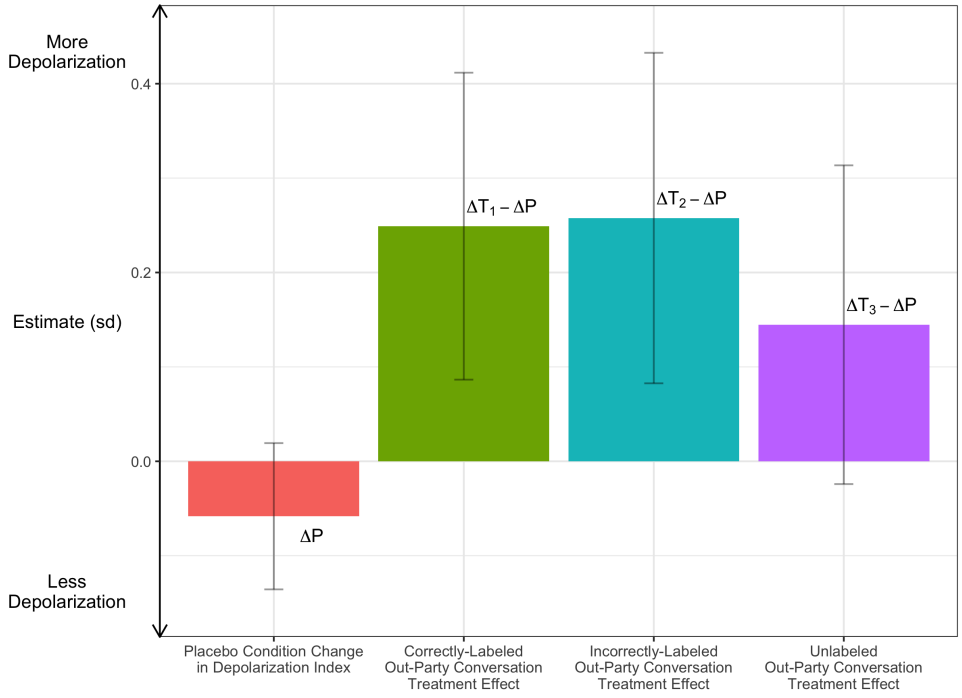

Figure 4: Labeling Treatment Effects Compared to All Placebo Respondents

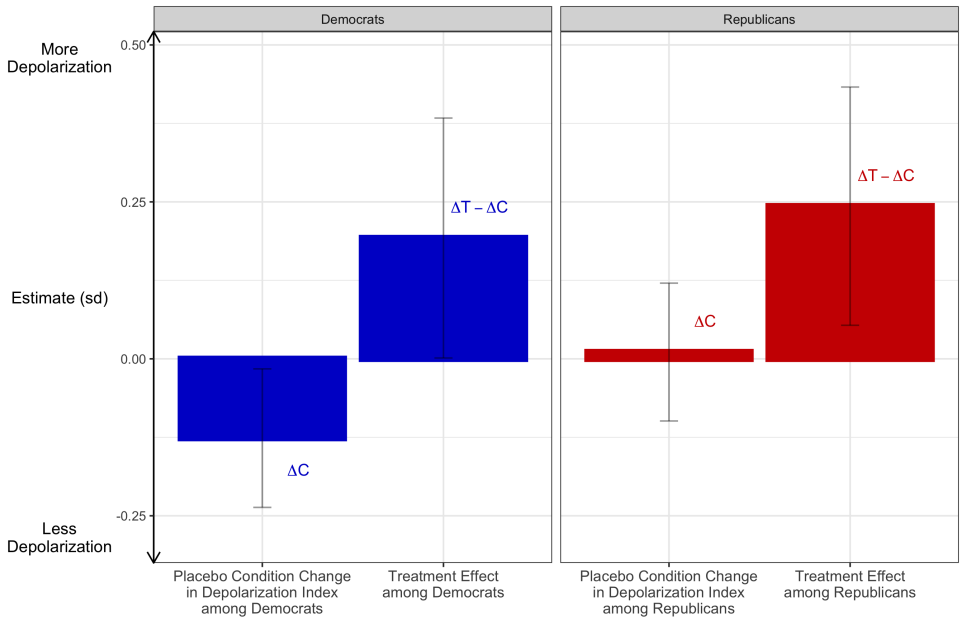

Figure 5: Out-Party Discussion Treatment Effects Compared to All Placebo Respondents by Party

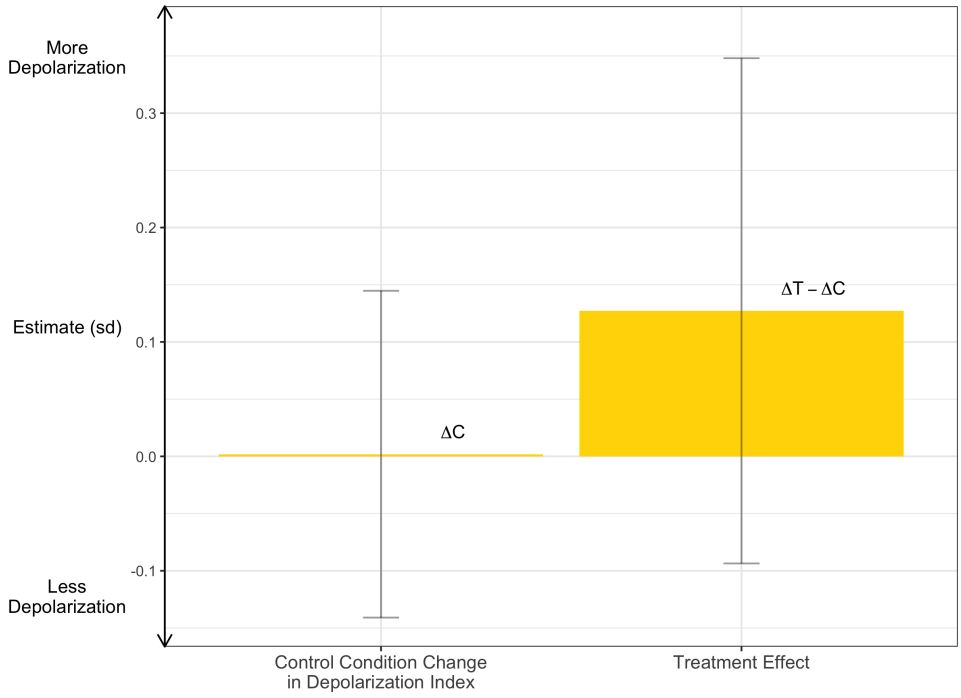

Figure 6: Treatment Effects Compared to Pure Control Condition

Table 4 predicts retention of respondents among users who downloaded the app and were randomized into a treatment condition. Column 1 predicts whether they completed at least 1 exchange with their partner. Column 2 predicts if they completed at least 10 exchanges (enough for compliance). These models indicate there are few differences among individuals who complete at least one exchange (that is, they downloaded the app and exchanged at least one round of messages with their partner) versus those who do not complete at least one message.

4.4 Intent to Treat Results

Here we present the intent to treat results (ITT) where the regression model is only the first equation in Section 3 with $Treatment_i$ replacing $Compliance_i$. Results are reported in Table 5. None of the statistical significance of results changes. Column 1 corresponds to main paper Figure 2, column 2 corresponds to Figure 3, and columns 3 and 4 correspond to Figure 4.

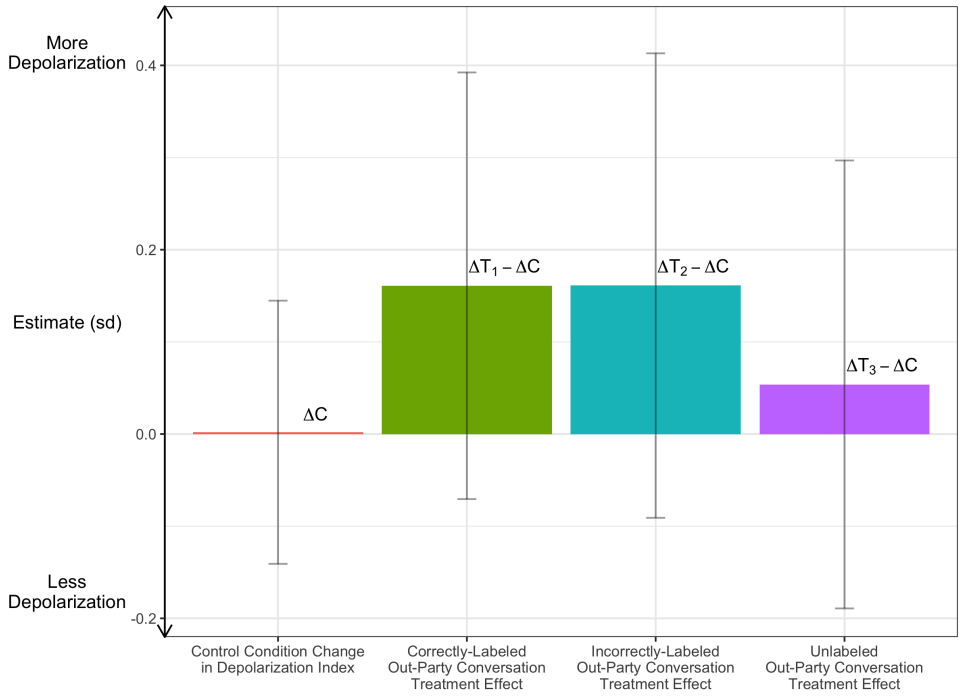

Figure 7: Treatment Effects Compared to Pure Control Condition by Labeling

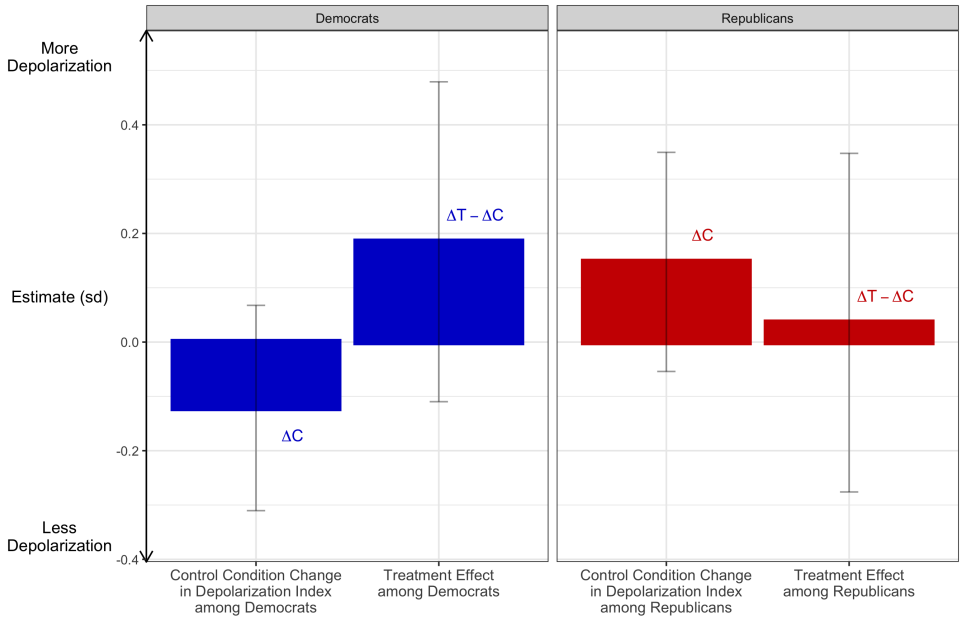

Figure 8: Treatment Effects Compared to Pure Control Condition by Party

Table 4: Attrition in App Usage

	Completed At Least 1 Exchange	Completed At Least 10 Exchanges
	(1)	(2)
correct_labels	−0.001 (0.02)	0.04 (0.03)
incorrect_labels	0.02 (0.02)	−0.02 (0.03)
gender	0.01 (0.02)	−0.001 (0.03)
age	0.001 (0.001)	0.0002 (0.001)
college	−0.01 (0.02)	0.001 (0.03)
white	0.01 (0.02)	0.02 (0.03)
knowledge_counter	0.002 (0.01)	−0.01 (0.02)
extraversion	−0.002 (0.005)	−0.01 (0.01)
agreeable	0.01* (0.01)	−0.01 (0.01)
consecientious	−0.01 (0.01)	−0.03* (0.01)
emotionally_stable	−0.004 (0.01)	0.04** (0.01)
open_new_experiences	0.01 (0.01)	0.01 (0.01)
political_interest	−0.004 (0.01)	0.01 (0.02)
strong_partisan	−0.02 (0.02)	−0.03 (0.03)
percent_other_party_exposure	0.0003 (0.0004)	−0.0002 (0.001)
democrat	−0.001 (0.03)	0.08 (0.05)
sevenpoint_ideology	0.01 (0.01)	0.02* (0.01)
regionMidwest	0.01 (0.03)	0.05 (0.04)
regionSouth	−0.01 (0.02)	0.002 (0.04)
regionWest	0.003 (0.02)	−0.01 (0.04)
faminc_val	0.0000 (0.0000)	0.0000 (0.0000)
faminc_notsay	0.05 (0.03)	0.01 (0.05)
Constant	0.81*** (0.08)	0.55*** (0.15)
Observations	1,216	1,216
Log Likelihood	−93.95	−775.72
Akaike Inf. Crit.	233.89	1,597.44

Note: *p<0.05; **p<0.01; ***p<0.001

Columns 1 and 2 cluster standard errors by conversation, columns 2 and 3 do not because no participants from the same conversation are in the same regression, following the same practice as the main paper figures.

Table 5: ITT Effects for Main Paper Figures

	All Respondents	Label Effects	Democrats	Republicans
	(1)	(2)	(3)	(4)
Any Treatment	0.153* (0.063)		0.060 (0.086)	0.243** (0.089)
Correct Labels		0.182* (0.073)		
Incorrect Labels		0.171* (0.071)		
No Labels		0.102 (0.071)		
Constant	−0.062 (0.058)	−0.062 (0.058)	−0.058 (0.078)	−0.065 (0.083)
Observations	1,419	1,419	715	704
R^2	0.004	0.006	0.001	0.010
Adjusted R^2	0.004	0.004	−0.001	0.009
Residual Std. Error	0.843	0.843	0.848	0.831
F Statistic	6.040*	2.697*	0.488	7.399**
Note:				*p<0.05; **p<0.01; ***p<0.001

4.5 Analysis of Component Indices

In the text, we report the results for a global polarization index. This section shows results for effects on the component indices. Figure 9 reports complier average causal effects estimated on the depolarization index (the same results presented in Figure 3 of the main text), affect and issue sub-indices that are averaged into the overall index, and the three components of the affect index. The top panel replicates the numbers reported in the main paper Figure 3. In scrutinizing the effects on the component indices, the most striking finding is that the sign changes (but it is not statistically significant) for the no label condition for outgroup trait stereotypes. While we are hesitant to read too much into it given the uncertainty in the estimate, this finding is consistent with recent research that finds that thermometer ratings, social distance, and trait

stereotypes are capturing different things (Druckman et al., 2018)

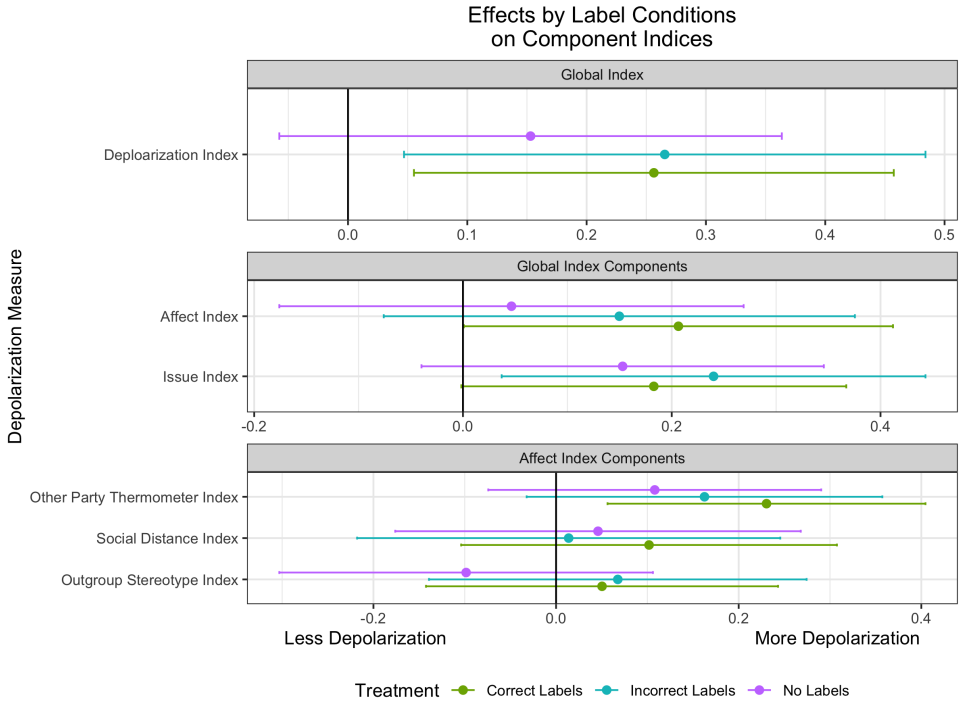

Figure 9: Treatment Effects on Component Indices. Effects are relative to the placebo condition. 95% confidence intervals are shown. Standard errors are clustered at the conversation-level.

4.6 Alternative Model Specifications

In this section we describe the results when including control variables in all regressions reported in the main paper and using various approaches for handling missing data in the variables.

Table 6 presents models predicting our key outcome (the depolarization index) with controls using list-wise deletion. The first column is the result reported in the main paper: column 1 considers the effect of the app when compared to participants in the placebo condition who were asked to write the essay, column 2 considers the reference group to be those who were not asked to do anything (pure control) and column 3 considers both these groups as a single

reference group. Columns 4-6 present the same comparisons while controlling for baseline covariates. We see that age and family income have significant negative effects. That is, older and more affluent individuals were less likely to depolarize no matter what condition they were in. Similarly, we see that those who identify as Democrats were also less likely to depolarize irrespective of their treatment status, as we reported in the main text of this article.

Tables 7-9 presents the same analysis as Table 6 Columns 4-6 but after imputing missing control variables via Multiple Imputation Chained Equations—a procedure that under the Missing At Random assumption generates several datasets where each missing value is probabilistically imputed. We generate 5 such datasets and report the regression on each of them in Tables 7-9. The results of this analysis are nearly identical to the models which employ list-wise deletion

Table 6: Treatment Effects by Reference Level and Control Variables with Complete Cases

	Placebo	Control	Both	Placebo (Comp. Cases)	Control (Comp. Cases)	Both (Comp. Cases)
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment Effect	0.224* (0.093)	0.127 (0.113)	0.186* (0.076)	0.259** (0.098)	0.126 (0.121)	0.204* (0.080)
Gender				0.079 (0.053)	0.073 (0.054)	0.078 (0.050)
Age				−0.005** (0.002)	−0.006*** (0.002)	−0.006*** (0.002)
College				0.049 (0.053)	0.028 (0.055)	0.027 (0.051)
White				−0.081 (0.063)	−0.070 (0.065)	−0.063 (0.060)
Political Knowledge				0.024 (0.032)	0.016 (0.032)	0.029 (0.030)
Extraversion				−0.014 (0.015)	−0.014 (0.015)	−0.008 (0.014)
Agreeableness				0.026 (0.024)	0.021 (0.025)	0.026 (0.023)
Conscientiousness				0.005 (0.024)	−0.006 (0.024)	−0.015 (0.022)
Emotional Stability				−0.015 (0.021)	−0.005 (0.022)	−0.008 (0.020)
Open to New Experiences				−0.003 (0.022)	0.0001 (0.023)	−0.0003 (0.021)
Interest in Politics				−0.059 (0.042)	−0.029 (0.042)	−0.060 (0.039)
Strong Partisan				0.068 (0.049)	0.046 (0.052)	0.047 (0.047)
Other Party Exposure				−0.001 (0.001)	−0.001 (0.001)	−0.001 (0.001)
Democrat				−0.223** (0.084)	−0.236** (0.089)	−0.208* (0.081)
Political Ideology				−0.022 (0.021)	−0.014 (0.022)	−0.018 (0.020)
Midwest				−0.038 (0.077)	−0.022 (0.079)	−0.052 (0.074)
South				−0.048 (0.071)	−0.044 (0.072)	−0.080 (0.068)
West				−0.054 (0.076)	−0.062 (0.078)	−0.060 (0.073)
Income				−0.070** (0.023)	−0.073*** (0.022)	−0.062** (0.021)
Constant	−0.062 (0.058)	0.002 (0.073)	−0.037 (0.046)	0.563* (0.261)	0.608* (0.264)	0.595* (0.244)
Observations	1,419	1,345	1,563	1,219	1,155	1,334
R^2	−0.010	−0.004	−0.005	0.021	0.031	0.025
Adjusted R^2	−0.011	−0.005	−0.006	0.005	0.013	0.011
Residual Std. Error	0.850	0.847	0.852	0.826	0.825	0.826

Note: *p<0.05; **p<0.01; ***p<0.001

Table 7: Treatment Effects with MICE Imputed Control Variables (Placebo as Reference Group)

	(1)	(2)	(3)	(4)	(5)
Treatment Effect	0.210* (0.095)	0.210* (0.095)	0.210* (0.095)	0.210* (0.095)	0.210* (0.095)
Gender	0.019 (0.049)	0.019 (0.049)	0.019 (0.049)	0.019 (0.049)	0.019 (0.049)
Age	−0.005** (0.002)	−0.005** (0.002)	−0.005** (0.002)	−0.005** (0.002)	−0.005** (0.002)
College	0.036 (0.050)	0.036 (0.050)	0.036 (0.050)	0.036 (0.050)	0.036 (0.050)
White	−0.065 (0.060)	−0.065 (0.060)	−0.065 (0.060)	−0.065 (0.060)	−0.065 (0.060)
Political Knowledge	0.016 (0.029)	0.016 (0.029)	0.016 (0.029)	0.016 (0.029)	0.016 (0.029)
Extraversion	−0.011 (0.014)	−0.011 (0.014)	−0.011 (0.014)	−0.011 (0.014)	−0.011 (0.014)
Agreeableness	0.007 (0.023)	0.007 (0.023)	0.007 (0.023)	0.007 (0.023)	0.007 (0.023)
Conscientiousness	−0.008 (0.023)	−0.008 (0.023)	−0.008 (0.023)	−0.008 (0.023)	−0.008 (0.023)
Emotional Stability	−0.009 (0.020)	−0.009 (0.020)	−0.009 (0.020)	−0.009 (0.020)	−0.009 (0.020)
Open to New Experiences	−0.009 (0.021)	−0.009 (0.021)	−0.009 (0.021)	−0.009 (0.021)	−0.009 (0.021)
Interest in Politics	−0.064 (0.039)	−0.064 (0.039)	−0.064 (0.039)	−0.064 (0.039)	−0.064 (0.039)
Strong Partisan	0.059 (0.047)	0.059 (0.047)	0.059 (0.047)	0.059 (0.047)	0.059 (0.047)
Other Party Exposure	−0.0003 (0.001)	−0.0003 (0.001)	−0.0003 (0.001)	−0.0003 (0.001)	−0.0003 (0.001)
Democrat	−0.264*** (0.079)	−0.264*** (0.079)	−0.264*** (0.079)	−0.264*** (0.079)	−0.264*** (0.079)
Political Ideology	−0.026 (0.020)	−0.026 (0.020)	−0.026 (0.020)	−0.026 (0.020)	−0.026 (0.020)
Midwest	−0.062 (0.073)	−0.062 (0.073)	−0.062 (0.073)	−0.062 (0.073)	−0.062 (0.073)
South	−0.076 (0.068)	−0.076 (0.068)	−0.076 (0.068)	−0.076 (0.068)	−0.076 (0.068)
West	−0.091 (0.073)	−0.091 (0.073)	−0.091 (0.073)	−0.091 (0.073)	−0.091 (0.073)
Income	−0.058** (0.021)	−0.058** (0.021)	−0.058** (0.021)	−0.058** (0.021)	−0.058** (0.021)
Constant	0.796** (0.244)	0.796** (0.244)	0.796** (0.244)	0.796** (0.244)	0.796** (0.244)
Observations	1,419	1,419	1,419	1,419	1,419
R^2	0.019	0.019	0.019	0.019	0.019
Adjusted R^2	0.005	0.005	0.005	0.005	0.005
Residual Std. Error (df = 1398)	0.843	0.843	0.843	0.843	0.843

Note: *p<0.05; **p<0.01; ***p<0.001

Table 8: Treatment Effects with MICE Imputed Control Variables (Pure Control as Reference Group)

	(1)	(2)	(3)	(4)	(5)
fully_compliant	0.130 (0.113)	0.130 (0.113)	0.130 (0.113)	0.130 (0.113)	0.130 (0.113)
gender	0.033 (0.050)	0.033 (0.050)	0.033 (0.050)	0.033 (0.050)	0.033 (0.050)
age	−0.005** (0.002)	−0.005** (0.002)	−0.005** (0.002)	−0.005** (0.002)	−0.005** (0.002)
college	0.014 (0.052)	0.014 (0.052)	0.014 (0.052)	0.014 (0.052)	0.014 (0.052)
white	−0.060 (0.062)	−0.060 (0.062)	−0.060 (0.062)	−0.060 (0.062)	−0.060 (0.062)
knowledge_counter	0.016 (0.030)	0.016 (0.030)	0.016 (0.030)	0.016 (0.030)	0.016 (0.030)
extraversion	−0.010 (0.015)	−0.010 (0.015)	−0.010 (0.015)	−0.010 (0.015)	−0.010 (0.015)
agreeable	0.009 (0.023)	0.009 (0.023)	0.009 (0.023)	0.009 (0.023)	0.009 (0.023)
consecientious	−0.007 (0.023)	−0.007 (0.023)	−0.007 (0.023)	−0.007 (0.023)	−0.007 (0.023)
emotionally_stable	−0.007 (0.020)	−0.007 (0.020)	−0.007 (0.020)	−0.007 (0.020)	−0.007 (0.020)
open_new_experiences	−0.001 (0.022)	−0.001 (0.022)	−0.001 (0.022)	−0.001 (0.022)	−0.001 (0.022)
political_interest	−0.032 (0.039)	−0.032 (0.039)	−0.032 (0.039)	−0.032 (0.039)	−0.032 (0.039)
strong_partisan	0.046 (0.049)	0.046 (0.049)	0.046 (0.049)	0.046 (0.049)	0.046 (0.049)
percent_other_party_exposure	−0.0002 (0.001)	−0.0002 (0.001)	−0.0002 (0.001)	−0.0002 (0.001)	−0.0002 (0.001)
democrat	−0.300*** (0.082)	−0.300*** (0.082)	−0.300*** (0.082)	−0.300*** (0.082)	−0.300*** (0.082)
sevenpoint_ideology	−0.025 (0.020)	−0.025 (0.020)	−0.025 (0.020)	−0.025 (0.020)	−0.025 (0.020)
regionMidwest	−0.042 (0.074)	−0.042 (0.074)	−0.042 (0.074)	−0.042 (0.074)	−0.042 (0.074)
regionSouth	−0.080 (0.069)	−0.080 (0.069)	−0.080 (0.069)	−0.080 (0.069)	−0.080 (0.069)
regionWest	−0.094 (0.075)	−0.094 (0.075)	−0.094 (0.075)	−0.094 (0.075)	−0.094 (0.075)
faminc_val	−0.055** (0.021)	−0.055** (0.021)	−0.055** (0.021)	−0.055** (0.021)	−0.055** (0.021)
Constant	0.684** (0.243)	0.684** (0.243)	0.684** (0.243)	0.684** (0.243)	0.684** (0.243)
Observations	1,345	1,345	1,345	1,345	1,345
R^2	0.023	0.023	0.023	0.023	0.023
Adjusted R^2	0.008	0.008	0.008	0.008	0.008
Residual Std. Error (df = 1324)	0.842	0.842	0.842	0.842	0.842

Note: *p<0.05; **p<0.01; ***p<0.001

Table 9: Treatment Effects with MICE Imputed Control Variables (Pure Control and Placebo as Reference Group)

	(1)	(2)	(3)	(4)	(5)
Treatment Effect	0.175* (0.077)	0.175* (0.077)	0.175* (0.077)	0.175* (0.077)	0.175* (0.077)
Gender	0.030 (0.047)	0.030 (0.047)	0.030 (0.047)	0.030 (0.047)	0.030 (0.047)
Age	−0.005** (0.002)	−0.005** (0.002)	−0.005** (0.002)	−0.005** (0.002)	−0.005** (0.002)
College	0.015 (0.048)	0.015 (0.048)	0.015 (0.048)	0.015 (0.048)	0.015 (0.048)
White	−0.058 (0.057)	−0.058 (0.057)	−0.058 (0.057)	−0.058 (0.057)	−0.058 (0.057)
Political Knowledge	0.026 (0.028)	0.026 (0.028)	0.026 (0.028)	0.026 (0.028)	0.026 (0.028)
Extraversion	−0.006 (0.014)	−0.006 (0.014)	−0.006 (0.014)	−0.006 (0.014)	−0.006 (0.014)
Agreeableness	0.011 (0.022)	0.011 (0.022)	0.011 (0.022)	0.011 (0.022)	0.011 (0.022)
Conscientiousness	−0.015 (0.021)	−0.015 (0.021)	−0.015 (0.021)	−0.015 (0.021)	−0.015 (0.021)
Emotional Stability	−0.007 (0.018)	−0.007 (0.018)	−0.007 (0.018)	−0.007 (0.018)	−0.007 (0.018)
Open to New Experiences	−0.005 (0.020)	−0.005 (0.020)	−0.005 (0.020)	−0.005 (0.020)	−0.005 (0.020)
Interest in Politics	−0.059 (0.037)	−0.059 (0.037)	−0.059 (0.037)	−0.059 (0.037)	−0.059 (0.037)
Strong Partisan	0.038 (0.045)	0.038 (0.045)	0.038 (0.045)	0.038 (0.045)	0.038 (0.045)
Other Party Exposure	0.00005 (0.001)	0.00005 (0.001)	0.00005 (0.001)	0.00005 (0.001)	0.00005 (0.001)
Democrat	−0.250*** (0.076)	−0.250*** (0.076)	−0.250*** (0.076)	−0.250*** (0.076)	−0.250*** (0.076)
Political Ideology	−0.019 (0.019)	−0.019 (0.019)	−0.019 (0.019)	−0.019 (0.019)	−0.019 (0.019)
Midwest	−0.079 (0.070)	−0.079 (0.070)	−0.079 (0.070)	−0.079 (0.070)	−0.079 (0.070)
South	−0.113 (0.065)	−0.113 (0.065)	−0.113 (0.065)	−0.113 (0.065)	−0.113 (0.065)
West	−0.110 (0.071)	−0.110 (0.071)	−0.110 (0.071)	−0.110 (0.071)	−0.110 (0.071)
Income	−0.049* (0.020)	−0.049* (0.020)	−0.049* (0.020)	−0.049* (0.020)	−0.049* (0.020)
Constant	0.727** (0.228)	0.727** (0.228)	0.727** (0.228)	0.727** (0.228)	0.727** (0.228)
Observations	1,563	1,563	1,563	1,563	1,563
R^2	0.022	0.022	0.022	0.022	0.022
Adjusted R^2	0.009	0.009	0.009	0.009	0.009
Residual Std. Error (df = 1542)	0.846	0.846	0.846	0.846	0.846

Note: *p<0.05; **p<0.01; ***p<0.001

4.7 Heterogeneous Treatment Effects

The results of this section describe the group-level intent to treat (ITT) causal effects modeled using Bayesian Additive Regression Trees (Hahn et al., 2020). We note that since these are ITTs they are smaller than the complier average effects we report in the main text. We only report point estimates of ITTs for each individual, so clustered standard errors, which leave point estimates unchanged, are not necessary. In Figure 10 we note that Democrats and Republicans exhibit overall different responses to the treatment—with Republicans depolarizing at nearly twice the rate as Democrats. Figures 11 and 12 plot the individual estimated treatment effects and color them by party identification— respondents with greater partisan identification and lower political knowledge have larger treatment effects. Two observations are important: first, the magnitude of these differences is substantially smaller than those observed for political party overall. Second, no matter the level of political knowledge or strength of partisanship, Republicans depolarize at a greater rate than Democrats. Figure 15 illustrates heterogeneity in treatment effects by age, suggesting older respondents were slightly more likely to depolarize while still maintaining the order (of Republicans depolarizing at a slightly higher rate than Democrats) within each age demographic.

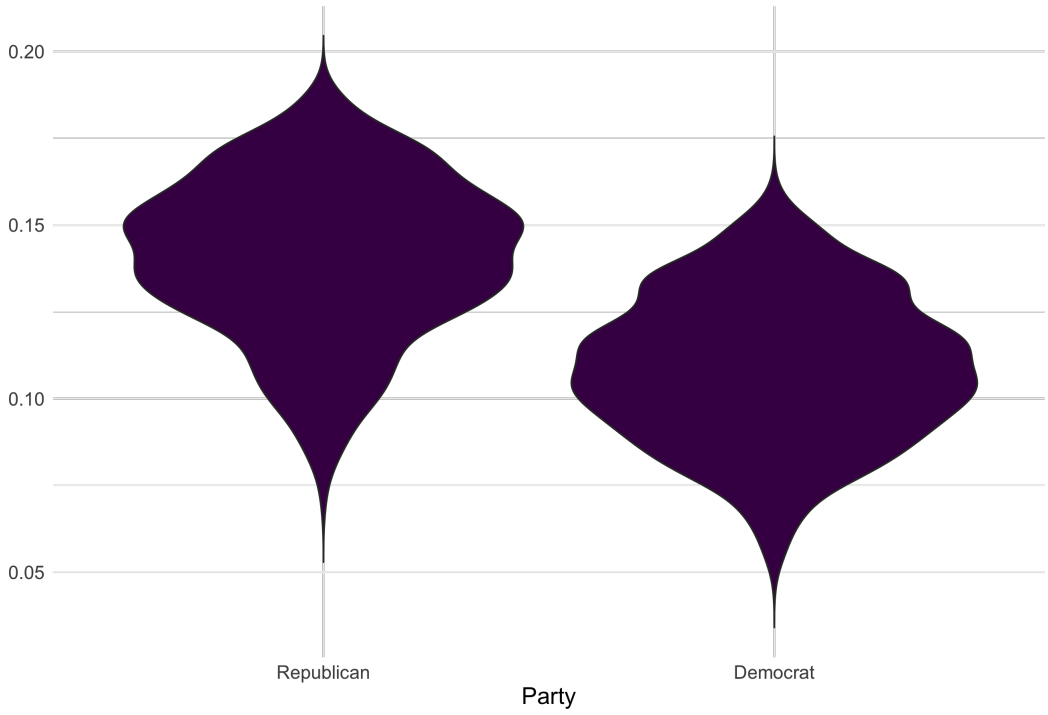

Figure 10: Heterogeneous Treatment Effects by Partisanship

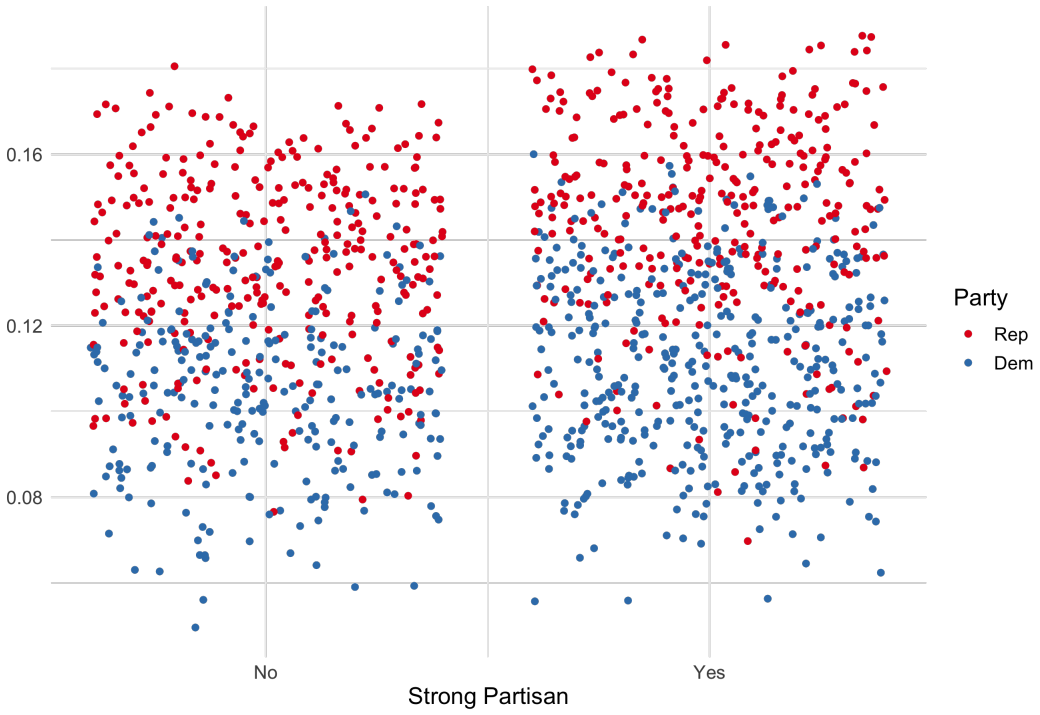

Figure 11: Heterogeneous Treatment Effects by Strength of Partisanship

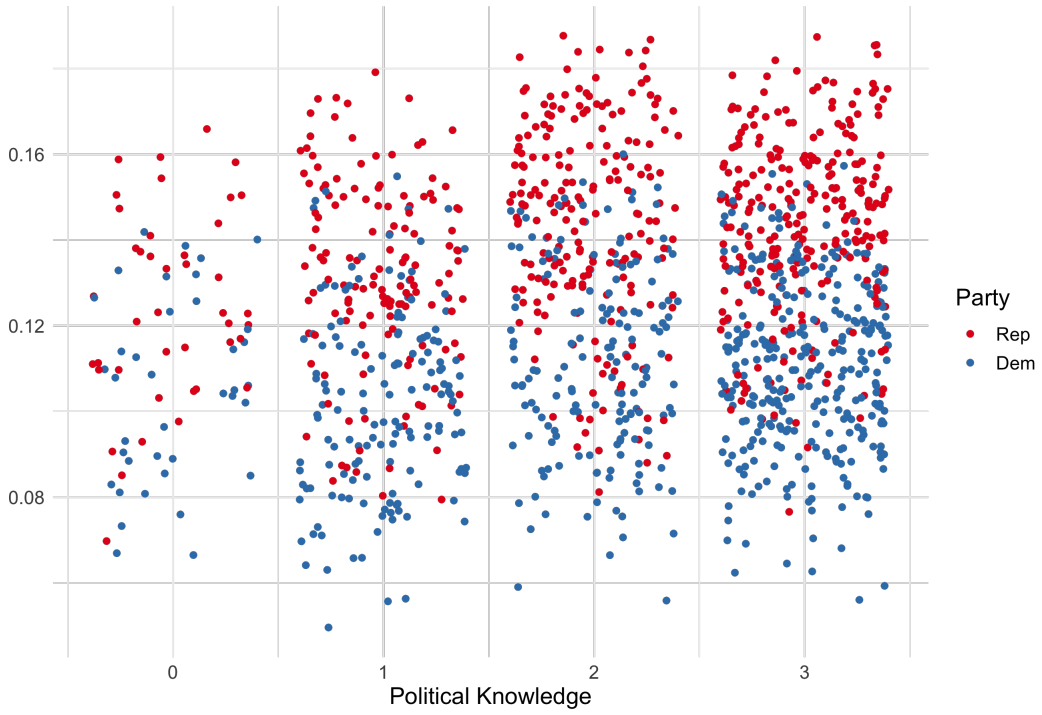

Figure 12: Heterogeneous Treatment Effects by Political Knowledge

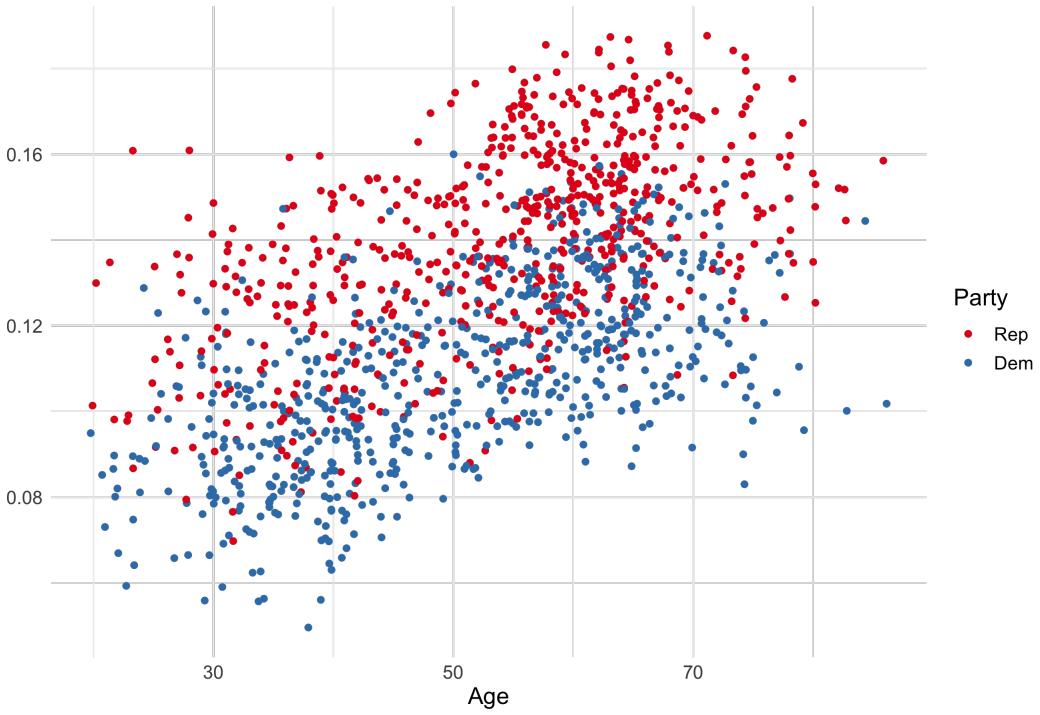

Figure 13: Heterogeneous Treatment Effects by Age

4.8 Pre-Treatment Level of Depolarization

A key highlight of Section 4.7 and main paper Figure 4 is the larger effect for Republicans. In this section we summarize the pre-treatment level of depolarization by party. The main conclusion is that Republicans actually begin slightly less polarized than Democrats, especially on issue-related polarization. So the larger treatment effect is not due to Republicans having more room to improve.

Figure 14 shows the distribution of component indices measuring depolarization (higher values indicate the respondent is less polarized) by party. Republicans are slightly less polarized on the Social Distance index and both issue component indices. See Section 2.6 for details on what questions are included in each index. Table 10 reports regressions of the level of pre-treatment depolarization on indicators for party, treatment, and the interaction between treatment and party. The treatment-related coefficients are all indistinguishable from zero, which confirms that our randomization balanced the pre-treatment level of depolarization across treatment arms. Except for other party thermometers (column 1), Republicans have significantly higher levels of depolarization (they are less polarized than Democrats) before treatment.

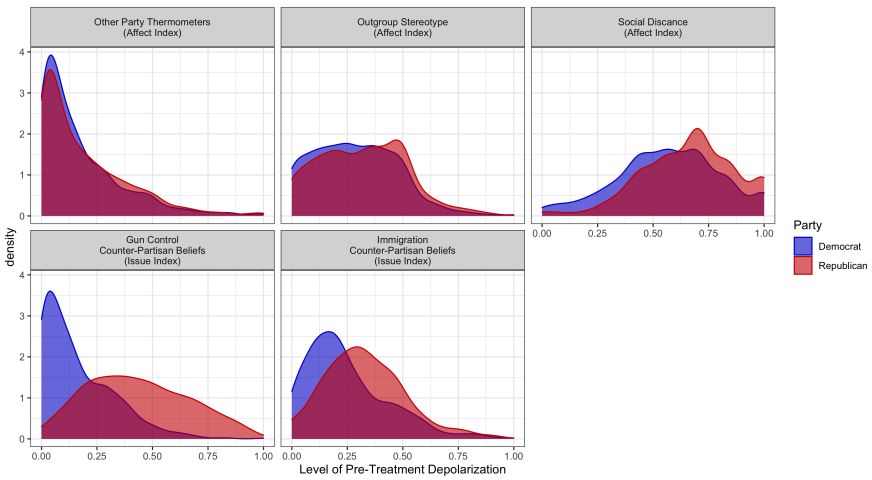

Figure 14: Distribution of Pre-Treatment Depolarization on Component Indices

Table 10: Pre-Treatment Levels of Depolarization

	Dependent variable:				
	Other Party Thermometers	Social Distance	Outgroup Stereotype	Immigration	Gun Control
	(1)	(2)	(3)	(4)	(5)
Republican	0.024 (0.026)	0.089*** (0.030)	0.045* (0.026)	0.049* (0.025)	0.271*** (0.027)
Treatment	−0.012 (0.019)	0.020 (0.023)	0.002 (0.020)	−0.022 (0.019)	−0.006 (0.020)
Republican x Treatment	−0.012 (0.029)	0.006 (0.033)	−0.013 (0.029)	0.041 (0.027)	−0.002 (0.029)
Constant	0.184*** (0.018)	0.551*** (0.021)	0.284*** (0.018)	0.266*** (0.017)	0.170*** (0.018)
Observations	1,385	1,417	1,418	1,414	1,411
R^2	0.002	0.044	0.008	0.050	0.319
Adjusted R^2	0.0003	0.042	0.006	0.048	0.317
Residual Std. Error	0.191	0.223	0.193	0.184	0.198
F Statistic	1.137	21.893***	3.719**	24.909***	219.387***

Note: *p<0.1; **p<0.05; ***p<0.01

4.9 Mislabelling Treatment Effectiveness

As we discussed in the main text of our article, some of the respondents who were assigned to the study's treatment condition intentionally received incorrect information about their discussion partner's party affiliation in order to disentangle the effect of partisan cues on our outcome. Respondents were always matched with a member of the opposing party, but respondents in this condition were informed that their discussion partner belonged to their own party. Respondents were informed of their partner's partisan affiliation in a way that made errors credible ("Our models predict [partner name] is a [Republican/Democrat]"), to mitigate the impact of respondent's discovering their partner's true party identification in the course of their conversation. Despite this deception, some respondents became aware that the party of their conversation partner had been mislabelled. After completing the required number of exchanges with their partner, respondents were asked to guess their partner's party affiliation. In the accurately labeled sub-treatment condition, this guess matched the information they were given about their partner 88% of the time. However, in the mislabeled treatment condition, this guess matched the (incorrect) information they were given only 44% of the time. When no partisan labels were provided, participants guessed their partner's party correctly 66% of the time.

One potential effect of this mislabeling is that feelings towards one's own party may change due to (incorrectly) believing one is talking to a co-partisan. While the most questions were asked only of the opposing party, we did collect thermometer ratings of same-party voters and politicians from each respondent. Figure 15 displays those results below. We only compare individual treatment conditions to the placebo group because we expect results to differ in the incorrect labels condition. The outcome is post-survey thermometer ratings of voters, politicians, or the average rating of the two minus pre-treatment survey ratings of the same entity. We use the same model specification as in the main results tables, using two-stage least squares to estimate complier average causal effects with standard errors clustered at the conversation

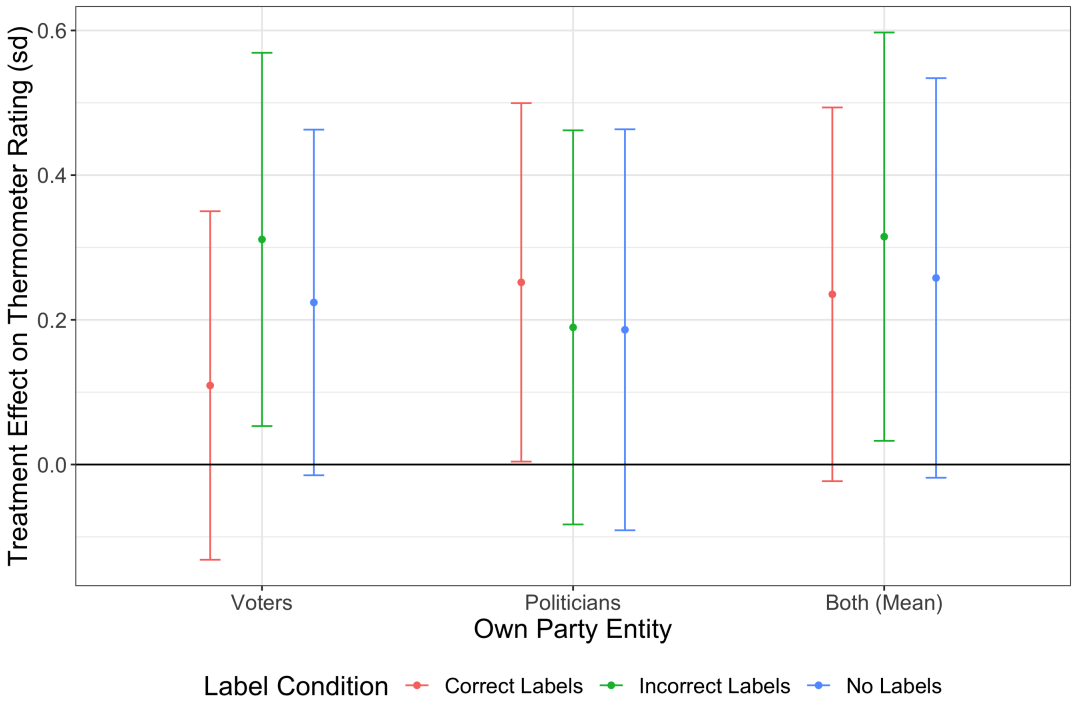

Figure 15: Treatment Effects on Thermometer Ratings of Own Party Entities. Outcome is the change in thermometer rating for same-party voters, politicians, or the average of both. Points and 95% CIs are for estimated treatment effects relative to the placebo condition. Standard Errors are clustered at the conversation level.

level. We indeed find a significant effect on the average feeling towards one's own party; those in the incorrect labels condition improved ratings of their own party by about 0.3 standard deviations. Thus, we have evidence that incorrectly labeling conversation partners improved opinions of one's own party. The result is driven mostly by improved ratings of voters from one's own party. Note, that the correct labels and no labels conditions also have positive point estimates for treatment effects, although these are only significant at the 5% level for the correctly labeled condition on own-party politicians.

4.10 Automated Text Analysis of Chat Data Generated Within App

An attractive feature of our field experiment is that it generated a large amount of text data that can be used to further analyze the impact of our treatment. In this section, we examine whether people in different study treatment conditions (and sub-conditions) used substantially different types of language. Though scholars are only beginning to understand how to use text-based data for causal inference (Mozer et al., 2020; Roberts et al., 2020), we followed an emerging consensus that summary measures derived from text can be used as mediators in causal analysis. More specifically, we hypothesized that the civility of an app user might increase the likelihood that their chat partner scores higher on our depolarization index. We created a measure of civility via natural language processing tools and include a global measure of civility derived from this approach about each respondent's chat partner to mediate the treatment effect on the respondent. Since respondents in our placebo and control conditions do not have partners, however, it is not possible to study the mediation effect of the text on the overall treatment effect we describe in the main text. Instead, we study the mediation of the text for differences among the three treatment sub-conditions.

We evaluate the civility of exchanges used the R package *politeness* (Yeomans et al., 2018). In the deliberation literature, civility is a construct that is more theoretically complex and nuanced than politeness (Papacharissi, 2004) but we treat the operationalization as interchangeable for examining cross-party interactions where the goal of the exchange is simple discussion rather than deliberation (Rossiter, 2020; Levendusky and Stecula, 2021). This package creates a scale intended to capture whether the language used "is bolstering the listener's self-image (showing gratitude, identifying as an in-group member, paying compliments) as well as not derogating that image (complaints, cursing, informal titles, and so on)...[and] respecting the listener's autonomy. This involves a general softening of statements, using hedges and adverbs. Requests may also be tempered, using indirect subjunctive language, and apologizing" (Yeo-

mans et al., 2018, p. 2).

For each message a respondent was exposed to, we use the package to compute all 40 text features the authors identify as being potentially related to politeness. For each feature, we standardize it to have mean zero and variance 1, then average across all 40 features to create a composite measure of civility. To calculate how civil a participant was, we sum the scores across messages sent by the user. We calculate the sum rather than average because we are interested in constructing a proxy for the total indicators of civility in the conversation, rather than the per-message average. As a validation exercise, we check whether respondents who said they found the conversation with their partner enjoyable also had more civil exchanges with their partners under this measure. Indeed, we confirm that the median civility score for respondents who liked their partner was about 0.21 standard deviations higher (more civil) than the median for respondents who did not like their partner.

We merge the civility index with our main analysis data only for treated participants who completed at least 10 exchanges. We thus have measurements of how civility for each participant and their partner. We focus on the partner civility as a mediator. Unobserved variables could confound the relationship between one's own civility and depolarization. However, if we observe that the (randomized) labeling treatment sub-conditions cause differences in experienced civility and that civility and depolarization are correlated, we can say with minimal assumptions that we have identified a mechanism for how the treatment effect works and a causal link between the civility of a conversation partner and depolarization.

Table 11 indicates that respondents with civil partners depolarize more, but we do not find that the labeling treatment sub-conditions affect the level of a partner's civility. While not strictly needed for this analysis we further note that a participant's own civility levels also do not vary with the labeling treatment sub-conditions. We can thus conclude that a correlational relationship exists—that is, more civil partners are associated with more depolarization—but we

cannot conclude a causal relationship exists. Yet future research could examine other features of the text beyond civility (e.g. displays of emotion or storytelling).

As a robustness check for this particular civility measure we also compute a civility index using the API for the Politeness Strategies module in Convokit (Zhang et al., 2018), which was built upon the features identified in previous research (Danescu-Niculescu-Mizil et al., 2013; Zhang et al., 2018). The module also produces, for each message, a list of features deemed by the authors to relate to politeness, with just slightly different features from the R package. As Table 12 shows, the direction and significance of the results remains mostly unchanged.

Finally, we show a scatterplot of the civility index and depolarization index at the individual level, with colors representing conversation topic and shapes representing labeling conditions in Figure 16. Effects look broadly similar across those populations. Note again that we present this figure only for illustrative purposes and cannot assess whether this is a causal relationship or not. Even though the correlation exists, partner civility was not randomized and the relationship could be confounded by ego characteristics that cause partners to be civil and depolarization.

Table 11: Civility measured with R package politeness

	Dependent variable:		
	Depolarization Index		
	All	Dem	Rep
	(1)	(2)	(3)
Partner's Civility	0.069*	0.097*	0.043
	(0.034)	(0.038)	(0.042)
CORRECT Labels	0.147	0.176	0.093
	(0.087)	(0.104)	(0.099)
INCORRECT Labels	0.129	0.134	0.111
	(0.090)	(0.105)	(0.104)
Republican	0.236**		
	(0.079)		
Constant	0.467	1.095**	−0.063
	(0.357)	(0.388)	(0.437)
Controls?	Yes	Yes	Yes
Observations	775	386	389
R^2	0.052	0.080	0.056
Adjusted R^2	0.026	0.033	0.008
Residual Std. Error	0.804	0.808	0.794
Note:	*p<0.05; **p<0.01; ***p<0.001		

Table 12: Civility measured with Python library convokit

	Dependent variable:		
	Depolarization Index		
	All	Dem	Rep
	(1)	(2)	(3)
Partner's Civility	0.470*	0.624	0.285
	(0.225)	(0.336)	(0.372)
CORRECT Labels	0.146*	0.178	0.091
	(0.070)	(0.104)	(0.099)
INCORRECT Labels	0.121	0.125	0.107
	(0.070)	(0.106)	(0.104)
Republican	0.233***		
	(0.066)		
Constant	0.462	1.081**	−0.069
	(0.294)	(0.390)	(0.437)
Controls?	Yes	Yes	Yes
Observations	775	386	389
R^2	0.048	0.072	0.055
Adjusted R^2	0.023	0.024	0.007
Residual Std. Error	0.805	0.811	0.795
Note:	*p<0.05; **p<0.01; ***p<0.001		

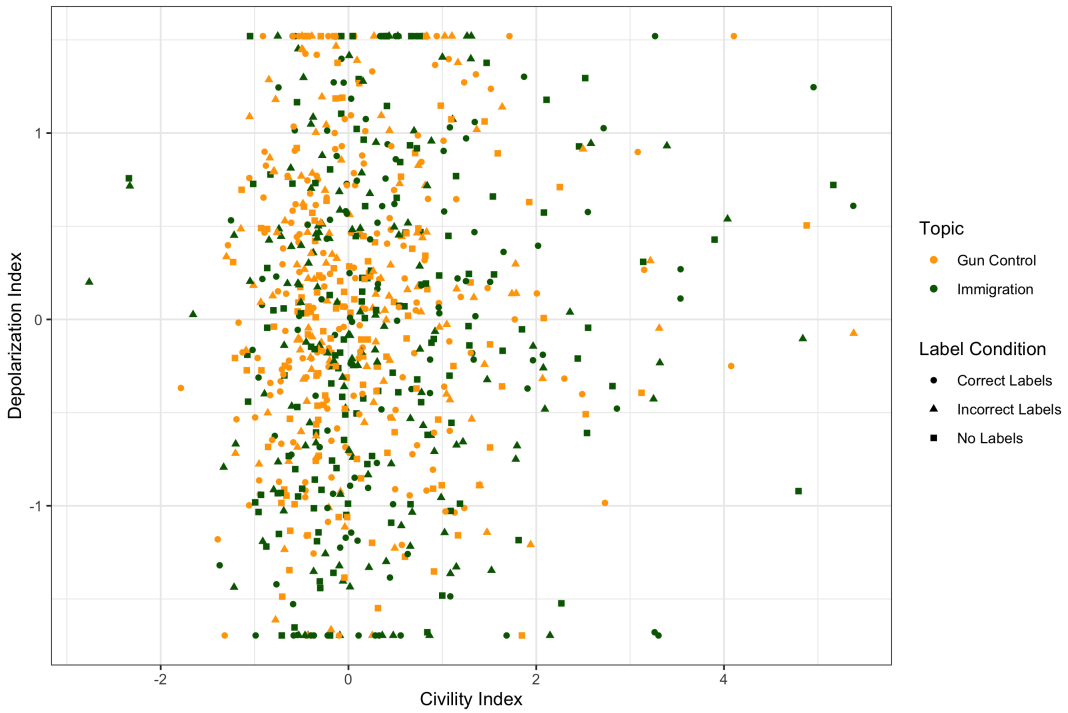

Figure 16: Depolarization and Partner Civility Indices by Conversation Topic and Labeling Condition

References

Allcott, H., L. Braghieri, S. Eichmeyer, and M. Gentzkow (2020). The welfare effects of social media. *American Economic Review 110*(3), 629–76.

Ansolabehere, S., J. Rodden, and J. M. Snyder (2008). The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting. *American Political Science Review 102*(2), 215–232.

Arceneaux, K. and R. J. V. Wielen (2017). *Taming Intuition: How Reflection Minimizes Partisan Reasoning and Promotes Democratic Accountability*. Cambridge, United Kingdom ; New York, NY: Cambridge University Press.

Cameron, A. C. and D. L. Miller (2015). A practitioner's guide to cluster-robust inference. *Journal of human resources 50*(2), 317–372.

Danescu-Niculescu-Mizil, C., M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts (2013). A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*.

Dias, N. and Y. Lelkes (2021). The Nature of Affective Polarization: Disentangling Policy Disagreement from Partisan Identity. *The American Journal of Political Science*.

Druckman, J. N., S. Klar, Y. Krupnikov, M. Levendusky, and J. B. Ryan (2021, jan). Affective polarization, local contexts and public opinion in America. *Nature Human Behaviour 5*(1), 28–38.

Druckman, J. N., M. S. Levendusky, and A. McLain (2018). No need to watch: How the effects of partisan media can spread via interpersonal discussions. *American Journal of Political Science 62*(1), 99–112.

Hahn, P. R., J. S. Murray, and C. M. Carvalho (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis 15*(3), 965–1056.

Kleiber, C. and A. Zeileis (2008). *Applied Econometrics with R*. New York: Springer-Verlag. ISBN 978-0-387-77316-2.

Levendusky, M. S. and D. A. Stecula (2021, may). *We Need to Talk*. Cambridge University Press.

Lieberson, S., S. Dumais, and S. Baumann (2000). The Instability of Androgynous Names: The Symbolic Maintenance of Gender Boundaries. *American Journal of Sociology 105*(5), 1249–1287. Publisher: University of Chicago Press.

Marengo, D., C. Sindermann, J. D. Elhai, and C. Montag (2020). One social media company to rule them all: associations between use of facebook-owned social media platforms, sociodemographic characteristics, and the big five personality traits. *Frontiers in psychology 11*, 936.

Mozer, R., L. Miratrix, A. R. Kaufman, and L. J. Anastasopoulos (2020). Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality. *Political Analysis 28*(4), 445–468.

Nickerson, D. W. (2005). Scalable protocols offer efficient design for field experiments. *Political Analysis 13*(3), 233–252.

Papacharissi, Z. (2004, June). Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society 6*(2).

Porter, E. and Y. R. Velez (2021). Placebo selection in survey experiments: An agnostic approach. *Political Analysis*, 1–14.

Roberts, M. E., B. M. Stewart, and R. A. Nielsen (2020). Adjusting for confounding with text matching. *American Journal of Political Science 64*(4), 887–903.

Rossiter, E. (2020). The consequences of interparty conversation on outparty affect and stereotypes. *Working Paper*.

Yeomans, M., A. Kantor, and D. Tingley (2018). The politeness package: Detecting politeness in natural language. *R Journal 10*(2).

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software 16*(9), 1–16.

Zeileis, A. and T. Hothorn (2002). Diagnostic checking in regression relationships. *R News 2*(3), 7–10.

Zeileis, A., S. Köll, and N. Graham (2020). Various versatile variances: An object-oriented implementation of clustered covariances in R. *Journal of Statistical Software 95*(1), 1–36.

Zhang, J., J. P. Chang, C. Danescu-Niculescu-Mizil, L. Dixon, Y. Hua, N. Thain, and D. Taraborelli (2018). Conversations gone awry: Detecting early signs of conversational failure. *arXiv preprint arXiv:1805.05345*.