Exposure to opposing views on social media can increase political polarization

Christopher A. Bail[a,1], Lisa P. Argyle[b], Taylor W. Brown[a], John P. Bumpus[a], Haohan Chen[c], M. B. Fallin Hunzaker[d], Jaemin Lee[a], Marcus Mann[a], Friedolin Merhout[a], and Alexander Volfovsky[e]

[a]Department of Sociology, Duke University, Durham, NC 27708; [b]Department of Political Science, Brigham Young University, Provo, UT 84602; [c]Department of Political Science, Duke University, Durham, NC 27708; [d]Department of Sociology, New York University, New York, NY 10012; and [e]Department of Statistical Science, Duke University, Durham, NC 27708

Edited by Peter S. Bearman, Columbia University, New York, NY, and approved August 9, 2018 (received for review March 20, 2018)

There is mounting concern that social media sites contribute to political polarization by creating "echo chambers" that insulate people from opposing views about current events. We surveyed a large sample of Democrats and Republicans who visit Twitter at least three times each week about a range of social policy issues. One week later, we randomly assigned respondents to a treatment condition in which they were offered financial incentives to follow a Twitter bot for 1 month that exposed them to messages from those with opposing political ideologies (e.g., elected officials, opinion leaders, media organizations, and nonprofit groups). Respondents were resurveyed at the end of the month to measure the effect of this treatment, and at regular intervals throughout the study period to monitor treatment compliance. We find that Republicans who followed a liberal Twitter bot became substantially more conservative posttreatment. Democrats exhibited slight increases in liberal attitudes after following a conservative Twitter bot, although these effects are not statistically significant. Notwithstanding important limitations of our study, these findings have significant implications for the interdisciplinary literature on political polarization and the emerging field of computational social science.

political polarization | computational social science | social networks | social media | sociology

Political polarization in the United States has become a central focus of social scientists in recent decades (1–7). Americans are deeply divided on controversial issues such as inequality, gun control, and immigration—and divisions about such issues have become increasingly aligned with partisan identities in recent years (8, 9). Partisan identification now predicts preferences about a range of social policy issues nearly three times as well as any other demographic factor—such as education or age (10). These partisan divisions not only impede compromise in the design and implementation of social policies but also have far-reaching consequences for the effective function of democracy more broadly (11–15).

America's cavernous partisan divides are often attributed to "echo chambers," or patterns of information sharing that reinforce preexisting political beliefs by limiting exposure to opposing political views (16–20). Concern about selective exposure to information and political polarization has increased in the age of social media (16, 21–23). The vast majority of Americans now visit a social media site at least once each day, and a rapidly growing number of them list social media as their primary source of news (24). Despite initial optimism that social media might enable people to consume more heterogeneous sources of information about current events, there is growing concern that such forums exacerbate political polarization because of social network homophily, or the well-documented tendency of people to form social network ties to those who are similar to themselves (25, 26). The endogenous relationship between social network formation and political attitudes also creates formidable challenges for the study of social media echo chambers and political polarization, since it is notoriously difficult to establish whether social media networks shape political opinions, or vice versa (27–29).

Here, we report the results of a large field experiment designed to examine whether disrupting selective exposure to partisan information among Twitter users shapes their political attitudes. Our research is governed by three preregistered hypotheses. The first hypothesis is that disrupting selective exposure to partisan information will decrease political polarization because of intergroup contact effects. A vast literature indicates contact between opposing groups can challenge stereotypes that develop in the absence of positive interactions between them (30). Studies also indicate intergroup contact increases the likelihood of deliberation and political compromise (31–33). However, all of these previous studies examine interpersonal contact between members of rival groups. In contrast, our experiment creates virtual contact between members of the public and opinion leaders from the opposing political party on a social media site. It is not yet known whether such virtual contact creates the

Significance

Social media sites are often blamed for exacerbating political polarization by creating "echo chambers" that prevent people from being exposed to information that contradicts their preexisting beliefs. We conducted a field experiment that offered a large group of Democrats and Republicans financial compensation to follow bots that retweeted messages by elected officials and opinion leaders with opposing political views. Republican participants expressed substantially more conservative views after following a liberal Twitter bot, whereas Democrats' attitudes became slightly more liberal after following a conservative Twitter bot—although this effect was not statistically significant. Despite several limitations, this study has important implications for the emerging field of computational social science and ongoing efforts to reduce political polarization online.

Author contributions: C.A.B., L.P.A., T.W.B., J.P.B., H.C., M.B.F.H., J.L., M.M., F.M., and A.V. designed research; C.A.B., L.P.A., T.W.B., H.C., M.B.F.H., J.L., M.M., and F.M. performed research; C.A.B., T.W.B., H.C., J.L., and A.V. contributed new reagents/analytic tools; C.A.B., L.P.A., T.W.B., H.C., M.B.F.H., J.L., M.M., F.M., and A.V. analyzed data; and C.A.B., L.P.A., T.W.B., M.B.F.H., M.M., F.M., and A.V. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: All data, code, and the markdown file used to create this report will be available at this link on the Dataverse: https://dataverse.harvard.edu/dataverse.xhtml?alias=chris_bail.

[1]To whom correspondence should be addressed. Email: christopher.bail@duke.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1804840115/-/DCSupplemental.

Published online August 28, 2018.

same type of positive mutual understanding—or whether the relative anonymity of social media forums emboldens people to act in an uncivil manner. Such incivility could be particularly rife in the absence of facial cues and other nonverbal gestures that might prevent the escalation of arguments in offline settings.

Our second hypothesis builds upon a more recent wave of studies that suggest exposure to those with opposing political views may create backfire effects that exacerbate political polarization (34–37). This literature—which now spans several academic disciplines—indicates people who are exposed to messages that conflict with their own attitudes are prone to counterargue them using motivated reasoning, which accentuates perceived differences between groups and increases their commitment to preexisting beliefs (34–37). Many studies in this literature observe backfire effects via survey experiments where respondents are exposed to information that corrects factual inaccuracies—such as the notion that Saddam Hussein possessed weapons of mass destruction prior to the 2003 US invasion of Iraq—although these findings have failed to replicate in two recent studies (38, 39). Yet our study is not designed to evaluate attempts to correct factual inaccuracies. Instead, we aim to assess the broader impact of prolonged exposure to counterattitudinal messages on social media.

Our third preregistered hypothesis is that backfire effects will be more likely to occur among conservatives than liberals. This hypothesis builds upon recent studies that indicate conservatives hold values that prioritize certainty and tradition, whereas liberals value change and diversity (40, 41). We also build upon recent studies in cultural sociology that examine the deeper cultural schemas and narratives that create and sustain such value differences (34, 26). Finally, we also build upon studies that observe asymmetric polarization in roll call voting wherein Republicans have become substantially more conservative whereas Democrats exhibit little or no increase in liberal voting positions (42). Although a number of studies have found evidence of this trend, we are not aware of any that examine such dynamics among the broader public—and on social media in particular.

Research Design

Fig. 1 provides an overview of our research design. We hired a professional survey firm to recruit self-identified Republicans and Democrats who visit Twitter at least three times each week to complete a 10-min survey in mid-October 2017 and 1.5 mo later. These surveys measure the key outcome variable: change in political ideology during the study period via a 10-item survey instrument that asked respondents to agree or disagree with a range of statements about policy issues on a seven-point scale ($\alpha = .91$) (10). Our survey also collected information about other political attitudes, use of social media and conventional media sources, and a range of demographic indicators that we describe in *SI Appendix*. Finally, all respondents were asked to report their Twitter ID, which we used to mine additional information about their online behavior, including the partisan background of the accounts they follow on Twitter. Our research was approved by the Institutional Review Boards at Duke University and New York University. All respondents provided informed consent before participating in our research.

We ran separate field experiments for Democratic and Republican respondents, and, within each group, we used a block randomization design that further stratified respondents according to two variables that have been linked to political polarization: (*i*) level of attachment to political party and (*ii*) level of interest in current events. We also randomized assignment according to respondents' frequency of Twitter use, which we reasoned would influence the amount of exposure to the intervention we describe in the following paragraph and thereby the overall likelihood of opinion change.

We received 1,652 responses to our pretreatment survey (901 Democrats and 751 Republicans). One week later, we randomly assigned respondents to a treatment condition, thus using an "ostensibly unrelated" survey design (43). At this time, respondents in the treatment condition were offered $11 to follow a Twitter bot, or automated Twitter account, that they were told would retweet 24 messages each day for 1 mo. Respondents were not informed of the content of the messages the bots would retweet. As Fig. 2 illustrates, we created a liberal Twitter bot and a conservative Twitter bot for each of our experiments. These bots retweeted messages randomly sampled from a list of 4,176 political Twitter accounts (e.g., elected officials, opinion leaders, media organizations, and nonprofit groups). These accounts were identified via a network-sampling technique that assumes those with similar political ideologies are more likely to follow each other on Twitter than those with opposing political ideologies (44). For further details about the design of the study's bots, please refer to *SI Appendix*.

To monitor treatment compliance, respondents were offered additional financial incentives (up to $18) to complete weekly surveys that asked them to answer questions about the content of the tweets produced by the Twitter bots and identify a picture of an animal that was tweeted twice a day by the bot but deleted immediately before the weekly survey. At the conclusion of the study period, respondents were asked to complete a final survey with the same questions from the initial (pretreatment) survey. Of those invited to follow a Twitter bot, 64.9% of Democrats and 57.2% of Republicans accepted our invitation. Approximately 62% of Democrats and Republicans who followed the bots were able to answer all substantive questions about the content of messages retweeted each week, and 50.2% were able to identify the animal picture retweeted each day.

Results

Fig. 3 reports the effect of being assigned to the treatment condition, or the Intent-to-Treat (ITT) effects, as well as the Complier Average Causal Effects (CACE) which account for the differential rates of compliance among respondents we observed. These estimates were produced via multivariate models that predict respondents' posttreatment scores on the liberal/conservative scale described above, controlling for pretreatment scores on this scale as well as 12 other covariates described in *SI Appendix*. We control for respondents' pretreatment liberal/conservative scale score to mitigate the influence of period effects. Negative scores indicate respondents became more liberal in response to treatment, and positive scores indicate they became more conservative. Circles describe unstandardized point estimates, and the horizontal lines in Fig. 3 describe 90% and 95% confidence intervals. We measured compliance with treatment in three ways. "Minimally Compliant Respondents" describes those who followed our bot throughout the entire study period. "Partially Compliant Respondents" are those who were able to answer at least one—but not all—questions about the content of one of the bots' tweets administered each week during the survey period. "Fully Compliant Respondents" are those who successfully answered all of these questions. These last two categories are mutually exclusive.

Although treated Democrats exhibited slightly more liberal attitudes posttreatment that increase in size with level of compliance, none of these effects were statistically significant. Treated Republicans, by contrast, exhibited substantially more conservative views posttreatment. These effects also increase with level of compliance, but they are highly significant. Our most cautious estimate is that treated Republicans increased 0.12 points on a seven-point scale, although our model that estimates the effect of treatment upon fully compliant respondents indicates this effect

Initial Survey

Respondents were offered $11 to provide their Twitter ID and complete a 10-minute survey about their political attitudes, social media use, and media consumption habits (demographics provided by survey firm).

Randomization

One week later, respondents were assigned to treatment and control conditions within strata created using pre-treatment covariates that describe attachment to party, frequency of Twitter use, and overall interest in current events.

Weekly Surveys

Respondents in treatment conditions informed they are eligible to receive up to $6 each week during the study period for correctly answering questions about the content of messages retweeted by Twitter .Bots.

Post-Survey

Respondents were offered $12 to repeat the pre-treatment survey one month after initial survey.

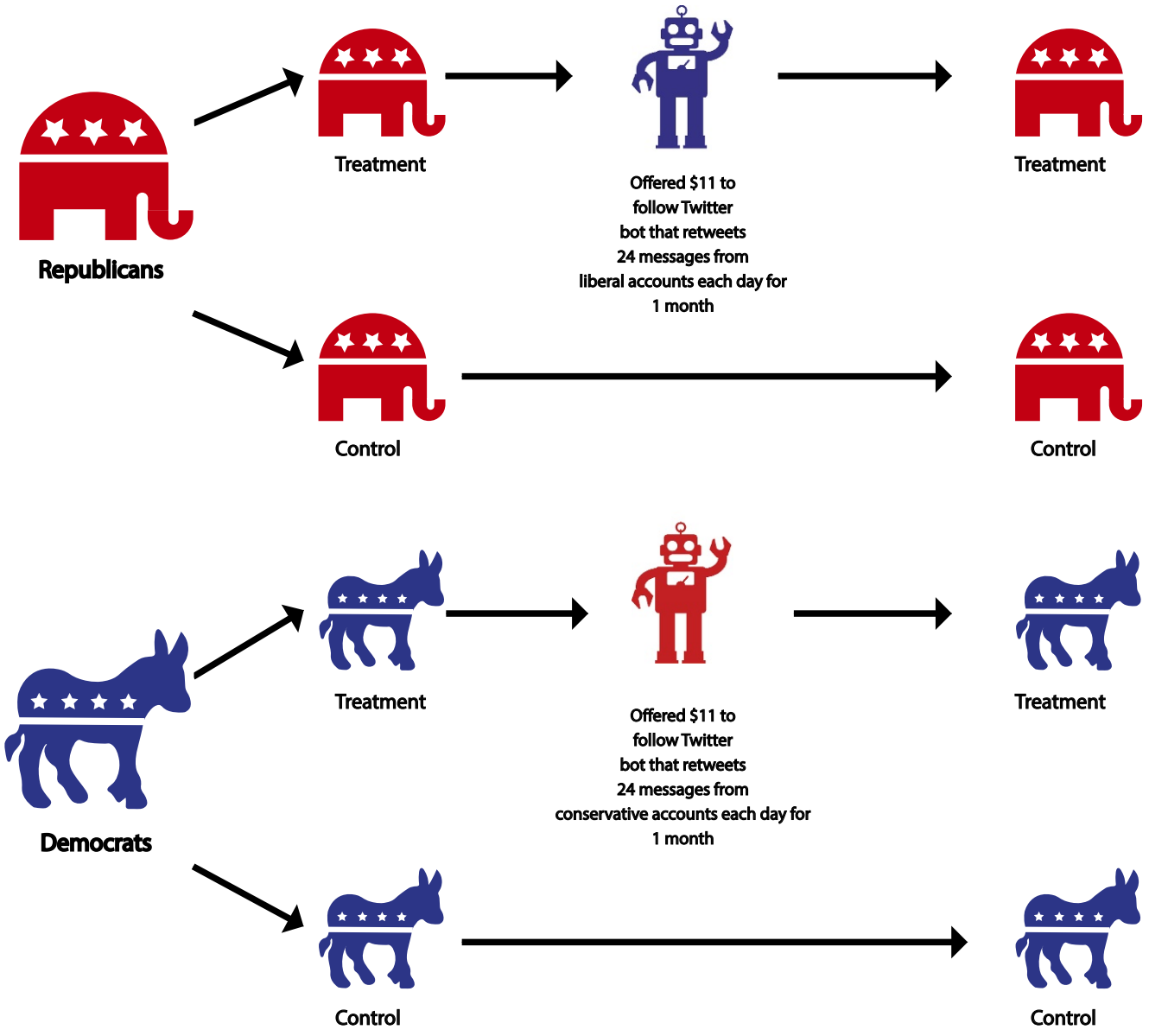

Fig. 1. Overview of research design.

is substantially larger (0.60 points). These estimates correspond to an increase in conservatism between 0.11 and 0.59 standard deviations.

Discussion and Conclusion

Before discussing the implications of these findings, we first note important limitations of our study. Readers should not interpret

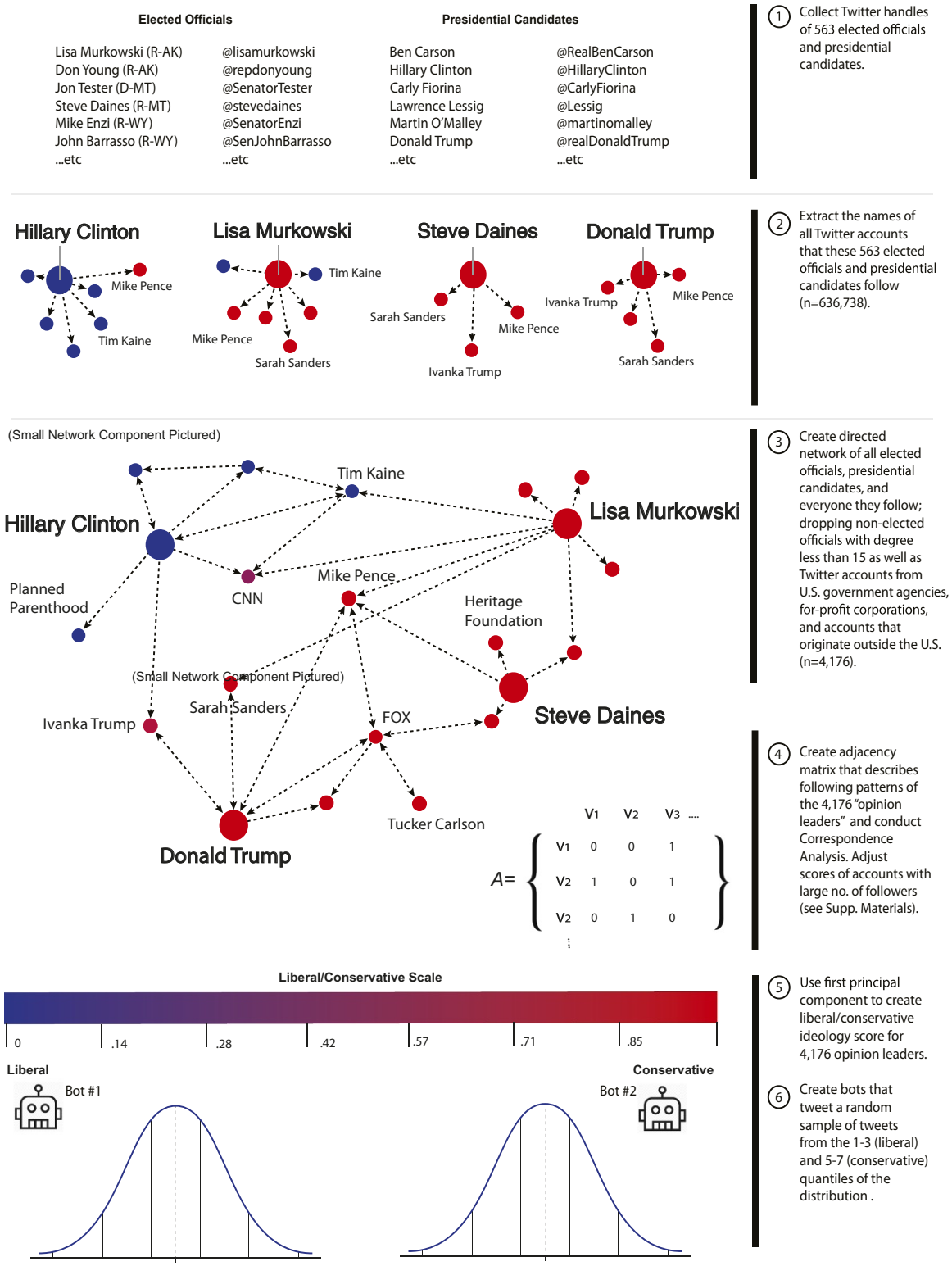

Fig. 2. Design of study's Twitter bots.

our findings as evidence that exposure to opposing political views will increase polarization in all settings. Although ours is among the largest field experiments conducted on social media to date, the findings above should not be generalized to the entire US population, because a majority of Americans do not use Twitter (24). It is also unclear how exposure to opposing views might shape political polarization in other parts of the world. In addition, we did not study people who identify as independents, or those who use Twitter but do so infrequently. Such individuals might exhibit quite different reactions to an intervention such as ours. Future studies are needed to further evaluate the external validity of our findings, because we offered our respondents

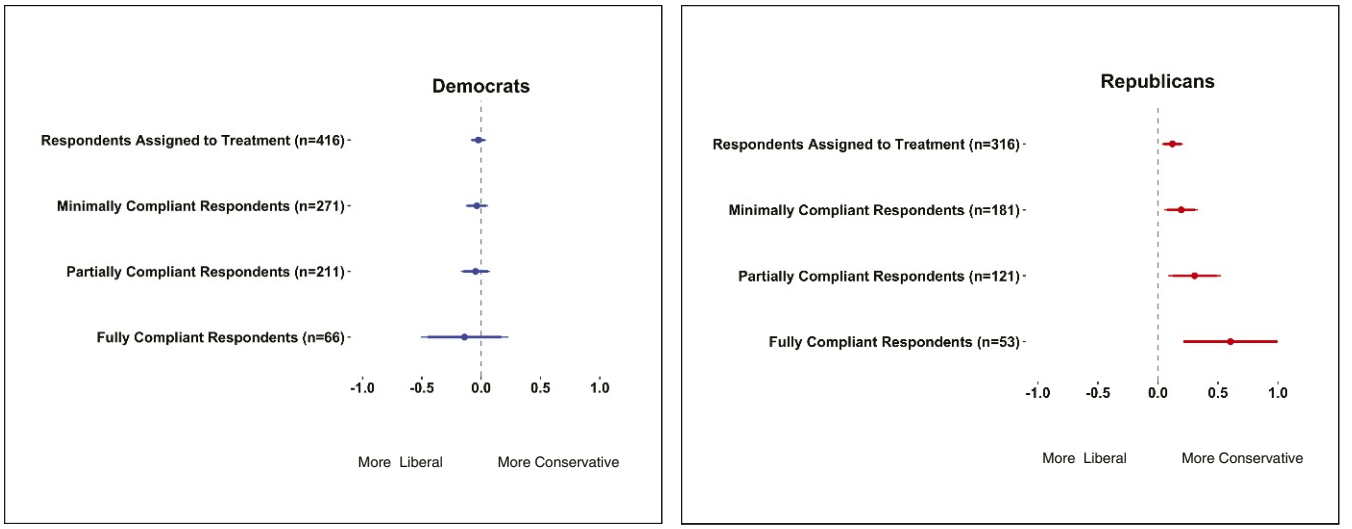

Fig. 3. Effect of following Twitter bots that retweet messages by elected officials, organizations, and opinion leaders with opposing political ideologies for 1 mo, on a seven-point liberal/conservative scale where larger values indicate more conservative opinions about social policy issues, for experiments with Democrats (n = 697) and Republicans (n = 542). Models predict posttreatment liberal/conservative scale score and control for pretreatment score on this scale as well as 12 other covariates described in *SI Appendix*. Circles describe unstandardized point estimates, and bars describe 90% and 95% confidence intervals. "Respondents Assigned to Treatment" describes the ITT effect for Democrats (ITT = −0.02, t = −0.76, p = 0.45, n = 416) and Republicans (ITT = 0.12, t = 2.68, p = 0.008, n = 316). "Minimally-Compliant Respondents" describes the CACE for respondents who followed one of the study's bots for Democrats (CACE = −0.04, t = −0.75, p = 0.45, n of compliant respondents = 271) and Republicans (CACE = 0.19, t = 2.73, $p < 0.007$, n of compliant respondents = 181). "Partially-Compliant Respondents" describes the CACE for respondents who correctly answered at least one question, but not all questions, about the content of a bot's tweets during weekly surveys throughout the study period for Democrats (CACE = −0.05, t = −0.75, p = 0.45, n of compliant respondents = 211) and Republicans (CACE = 0.31, t = 2.73, $p <.007$, n of compliant respondents = 121). "Fully-Compliant Respondents" describes the CACE for respondents who answered all questions about the content of the bot's tweets correctly for Democrats (CACE = −0.14, t = −0.75, p = 0.46, n of compliant respondents = 66) and Republicans (CACE = 0.60, t = 2.53, $p < 0.01$, n of compliant respondents = 53). Although treated Democrats exhibited slightly more liberal attitudes posttreatment that increase in size with level of compliance, none of these effects were statistically significant. In contrast, treated Republicans exhibited substantially more conservative views posttreatment that increase in size with level of compliance, and these effects are highly significant.

financial incentives to read messages from people or organizations with opposing views. It is possible that Twitter users may simply ignore such counterattitudinal messages in the absence of such incentives. Perhaps the most important limitation of our study is that we were unable to identify the precise mechanism that created the backfire effect among Republican respondents reported above. Future studies are thus urgently needed not only to determine whether our findings replicate in different populations or within varied social settings but to further identify the precise causal pathways that create backfire effects more broadly.

Future studies are also needed because we cannot rule out all alternative explanations of our findings. In *SI Appendix*, we present additional analyses that give us confidence that our results are not driven by Hawthorne effects, partisan "learning" processes, variation in the ideological extremity of messages by party, or demographic differences in social media use by age. At the same time, we are unable to rule out other alternative explanations discussed in *SI Appendix*. For example, it is possible that our findings resulted from increased exposure to information about politics, and not exposure to opposing messages per se. Similarly, increases in conservatism among Republicans may have resulted from increased exposure to women or racial and ethnic minorities whose messages were retweeted by our liberal bot. Finally, our intervention only exposed respondents to high-profile elites with opposing political ideologies. Although our liberal and conservative bots randomly selected messages from across the liberal and conservative spectrum, previous studies indicate such elites are significantly more polarized than the general electorate (45). It is thus possible that the backfire effect we identified could be exacerbated by an antielite bias, and future studies are needed to examine the effect of online intergroup contact with nonelites.

Despite these limitations, our findings have important implications for current debates in sociology, political science, social psychology, communications, and information science. Although we found no evidence that exposing Twitter users to opposing views reduces political polarization, our study revealed significant partisan differences in backfire effects. This finding is important, since our study examines such effects in an experimental setting that involves repeated contact between rival groups across an extended time period on social media. Our field experiment also disrupts selective exposure to information about politics in a real-world setting through a combination of survey research, bot technology, and digital trace data collection. This methodological innovation enabled us to collect information about the nexus of social media and politics with high granularity while developing techniques for measuring treatment compliance, mitigating causal interference, and verifying survey responses with behavioral data—as we discuss in *SI Appendix*. Together, we believe these contributions represent an important advance for the nascent field of computational social science (46).

Although our findings should not be generalized beyond party-identified Americans who use Twitter frequently, we note that recent studies indicate this population has an outsized influence on the trajectory of public discussion—particularly as the media itself has come to rely upon Twitter as a source of news and a window into public opinion (47). Although limited in scope, our findings may be of interest to those who are working to reduce political polarization in applied settings. More specifically, our study indicates that attempts to introduce people to a broad range of opposing political views on a social media site such as Twitter might be not only be ineffective but counterproductive—particularly if such interventions are initiated by liberals. Since previous studies have produced substantial evidence that intergroup contact produces compromise and mutual understanding in other contexts, however, future attempts to reduce political

polarization on social media will most likely require learning which types of messages, tactics, or issue positions are most likely to create backfire effects and whether others—perhaps delivered by nonelites or in offline settings—might be more effective vehicles to bridge America's partisan divides.

Materials and Methods

See *SI Appendix* for a detailed description of all materials and methods used within this study as well as links to our preregistration statement, replication materials, additional robustness checks, and an extended discussion of alternative explanations of our findings. Our research was approved by the Institutional Review Boards at Duke University and New York University.

ACKNOWLEDGMENTS. We thank Paul DiMaggio, Sunshine Hillygus, Gary King, Fan Li, Arthur Lupia, Brendan Nyhan, and Samantha Luks for helpful conversations about this study prior to our research. Our work was supported by the Carnegie Foundation, the Russell Sage Foundation, and the National Science Foundation.

1. DiMaggio P, Evans J, Bryson B (1996) Have American's social attitudes become more polarized? *Am J Sociol* 102:690–755.
2. Iyengar S, Westwood SJ (2015) Fear and loathing across party lines: New evidence on group polarization. *Am J Polit Sci* 59:690–707.
3. Baldassarri D, Gelman A (2008) Partisans without constraint: Political polarization and trends in American public opinion. *Am J Sociol* 114:408–446.
4. Sides J, Hopkins DJ (2015) *Political Polarization in American Politics* (Bloomsbury, London).
5. Baldassarri D, Bearman P (2007) Dynamics of political polarization. *Am Sociol Rev* 72:784–811.
6. Fiorina MP, Abrams SJ (2008) Political polarization in the American public. *Annu Rev Polit Sci* 11:563–588.
7. DellaPosta D, Shi Y, Macy M (2015) Why do liberals drink lattes? *Am J Sociol* 120: 1473–1511.
8. Levendusky M (2009) *The Partisan Sort: How Liberals Became Democrats and Conservatives Became Republicans* (Univ Chicago Press, Chicago).
9. Mason L (2018) *Uncivil Agreement: How Politics Became Our Identity* (Univ Chicago Press, Chicago).
10. Dimock M, Carroll D (2014) Political polarization in the American public: How increasing ideological uniformity and partisan antipathy affect politics, compromise, and everyday life (Pew Res Cent, Washington, DC).
11. Achen CH, Bartels LM (2016) *Democracy for Realists: Why Elections Do Not Produce Responsive Government* (Princeton Univ Press, Princeton).
12. Erikson RS, Wright GC, McIver JP (1993) *Public Opinion and Policy in the American States* (Cambridge Univ Press, Cambridge, UK).
13. Fishkin JS (2011) *When the People Speak: Deliberative Democracy and Public Consultation* (Oxford Univ Press, Oxford).
14. Bennett WL, Iyengar S (2008) A new era of minimal effects? The changing foundations of political communication. *J Commun* 58:707–731.
15. Sunstein C (2002) *Republic.com* (Princeton Univ Press, Princeton).
16. Bakshy E, Messing S, Adamic LA (2015) Political science. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348:1130–1132.
17. Sunstein C (2001) *Echo Chambers, Bush v. Gore, Impeachment, and Beyond* (Princeton Univ Press, Princeton).
18. King G, Schneer B, White A (2017) How the news media activate public expression and influence national agendas. *Science* 358:776–780.
19. Berry JM, Sobieraj S (2013) *The Outrage Industry: Political Opinion Media and the New Incivility* (Oxford Univ Press, Oxford).
20. Prior M (2013) Media and political polarization. *Annu Rev Polit Sci* 16: 101–127.
21. Pariser E (2011) *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think* (Penguin, New York).
22. Conover M, Ratkiewicz J, Francisco M (2011) Political polarization on twitter. *ICWSM* 133:89–96.
23. Boxell L, Gentzkow M, Shapiro JM (2017) Greater Internet use is not associated with faster growth in political polarization among us demographic groups. *Proc Natl Acad Sci USA* 114:10612–10617.
24. Perrin A (2015) Social media usage: 2005-2015 (Pew Res Cent, Washington, DC).
25. McPherson M, Smith-lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. *Annu Rev Sociol* 27:415–444.
26. Edelmann A, Vaisey S (2014) Cultural resources and cultural distinction in networks. *Poetics* 46:22–37.
27. Lazer D, Rubineau B, Chetkovich C, Katz N, Neblo M (2010) The coevolution of networks and political attitudes. *Polit Commun* 27:248–274.
28. Centola D (2011) An experimental study of homophily in the adoption of health behavior. *Science* 334:1269–1272.
29. Vaisey S, Lizardo O (2010) Can cultural worldviews influence network composition?. *Social Forces* 88:1595–1618.
30. Pettigrew TF, Tropp LR (2006) A meta-analytic test of intergroup contact theory. *J Pers Soc Psychol* 90:751–783.
31. Huckfeldt R, Johnson PE, Sprague J (2004) *Political Disagreement: The Survival of Diverse Opinions Within Communication Networks* (Cambridge Univ Press, Cambridge, UK).
32. Mutz DC (2002) Cross-cutting social networks: Testing democratic theory in practice. *Am Polit Sci Rev* 96:111–126.
33. Grönlund K, Herne K, Setälä M (2015) Does enclave deliberation polarize opinions? *Polit Behav* 37:995–1020.
34. Bail C (2015) *Terrified: How Anti-Muslim Fringe Organizations Became Mainstream* (Princeton Univ Press, Princeton).
35. Lord CG, Ross L, Lepper MR (1979) Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *J Pers Soc Psychol* 37:2098–2109.
36. Nyhan B, Reifler J (2010) When corrections fail: The persistence of political misperceptions. *Polit Behav* 32:303–330.
37. Taber CS, Lodge M (2006) Motivated skepticism in the evaluation of political beliefs. *Am J Polit Sci* 50:755–769.
38. Wood T, Porter E (2016) The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Polit Behav*, 1–29.
39. Wood T, Porter E (2018) The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Polit Behav*, 10.1007/s11109-018-9443-y.
40. Graham J, Haidt J, Nosek BA (2009) Liberals and conservatives rely on different sets of moral foundations. *J Pers Social Psychol* 96:1029–1046.
41. Jost JT, et al. (2007) Are needs to manage uncertainty and threat associated with political conservatism or ideological extremity? *Pers Soc Psychol Bull* 33:989–1007.
42. Grossmann M, Hopkins DA (2016) *Asymmetric Politics: Ideological Republicans and Group Interest Democrats* (Oxford Univ Press, Oxford).
43. Broockman D, Kalla J (2016) Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science* 352:220–224.
44. Barberá P (2014) Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Polit Anal* 23:76–91.
45. Abramowitz AI, Saunders KL (2008) Is polarization a myth?. *J Polit* 70:542–555.
46. Lazer D, et al (2009) Life in the network: The coming age of computational social science. *Science* 323:721–723.
47. Faris R, et al. (2017) Partisanship, Propaganda, and disinformation: Online media and the 2016 U.S. presidential election, (Berkman Klein Cent Internet Soc Harvard Univ, Cambridge, MA).