Unfollowing hyperpartisan social media influencers durably reduces out-party animosity

Steve Rathje[1]*, Clara Pretus[2], James K. He[3], Trisha Harjani[3], Jon Roozenbeek[3,4], Kurt Gray[5], Sander van der Linden[3], Jay J. Van Bavel[1,6]*

Affiliations:

[1]New York University, Department of Psychology

[2]Universitat Autònoma de Barcelona, Department of Psychobiology and Methodology of Health Sciences

[3]University of Cambridge, Department of Psychology

[4] King's College London, Department of War Studies

[5] University of North Carolina, Chapel Hill, Department of Psychology

[6] Norwegian School of Economics

*Corresponding authors. Email: srathje@alumni.stanford.edu and jay.vanbavel@nyu.edu.

Abstract Word Count: 292 words
Word Count: 4,971

Steve Rathje: 0000-0001-6727-571X
Clara Pretus: 0000-0003-2172-1184
Trisha Harjani: 0000-0002-5829-9485
Jon Roozenbeek: 0000-0002-8150-9305
James K. He: 0000-0002-1859-4914
Kurt Gray: 0000-0001-5816-2676
Sander van der Linden: 0000-0002-7144-806X
Jay J. Van Bavel: 0000-0002-2520-0442

Abstract: There is considerable debate over whether and how social media contributes to polarization[1–3]. Research suggests that a small number of hyperpartisan "influencers," or highly followed accounts, produce the vast majority of misinformation and toxic content[4–6]. Yet, little is known about the long-term causal effects of exposure to these influencers. In a correlational study (n_1 = 1,447) and two digital field experiments (n_2 = 494, n_3 = 1,133), we examined whether (un)following hyperpartisan social media influencers contributes to polarization and misinformation sharing. We found that incentivizing Twitter/X users to unfollow hyperpartisan social media influencers improved their recent feelings toward the out-party by 23.5% compared to the control group, with effects persisting for at least six months. Unfollowing also led participants to engage with more accurate news accounts, increased satisfaction with their Twitter/X feeds, and reduced the amount of political content they reported seeing a full year later—without reducing engagement. By contrast, incentivizing users to follow accounts that tweeted about science improved well-being. Additionally, we found that, after Elon's Musk purchased Twitter/X and made several platform changes, participants used Twitter/X less frequently, viewed their feeds as less reliable, and posted lower quality news. Our results demonstrate the long-term, causal impact of repeated exposure to hyper-partisan influencers on attitudes and behavior. They also illustrate that the behavior and experience of Twitter/X users changed substantially after Elon Musk's purchase of the platform, revealing the potential impact of social media design changes. Our work has implications for interventions that can be made by platforms or by individuals seeking to curate their social media experience. Unlike other social media reduction interventions, unfollowing is a targeted approach: like a scalpel, it surgically removes a few harmful parts of one's feed, allowing the beneficial aspects to remain.

Introduction

With more than 5 billion social media users worldwide[7], there is widespread concern about the potential role of social media in fostering partisan animosity and intergroup conflict[1,2,8,9]. Yet, researchers are divided about whether and how social media plays a causal role in shaping affective polarization, or negative feelings toward the out-party[3,10]. Experiments testing the causal effect of reducing social media usage on out-group attitudes have yielded mixed results[11–13], which makes it difficult to draw broad conclusions about the consequences of social media use[1]. The current paper aims to understand which specific aspects of social media might foster partisan animosity and how this can be mitigated.

Rather than looking at overall social media usage, it may be more productive to examine how specific ways of using social media contribute to polarization. Some have proposed that social media may increase polarization by sorting people into "echo chambers" that predominantly showcase content from like-minded sources[14–16]. However, the evidence for the polarizing impact of online echo chambers is mixed[17–20]. For instance, one study found that breaking open echo chambers by having people follow a bot that retweeted accounts from the opposing political party *increased* ideological polarization[21].

Exposure to polarizing content may be more harmful than echo chambers alone. A growing body of work suggests that out-group animosity[22–24] and other forms of hostile content[25–29] tend to go "viral" online, even though most people do not want this to be the case[24,30,31]. Hostile online content may reinforce negative stereotypes about one's out-group[20,23,32,33], which could in turn increase out-group animosity. This may be why breaking open online echo chambers can sometimes be unproductive: exposure to polarizing messages—whether they come from inside or outside one's echo chamber—may foster partisan animosity.

A growing body of research has found that toxic and false content disproportionately stems from a very small number of highly influential social media accounts[4–6], including political elites, media sources, and other highly prominent accounts (often called "influencers")[34]. While there has been correlational work on the effects of following "hyperpartisan" influencers[15], or accounts that frequently share politically-biased, misleading, or false information[35,36], the causal effects of exposure to these influencers on beliefs and behavior are unclear. Since prior work suggests that a decent amount of online polarization is merely driven by people self-selecting into polarizing online communities[37–39], and several high-powered experiments that altered aspects of one's social media feed have yielded null results[40,41], it is not obvious that following these influencers has a meaningful causal impact. Testing the effect of unfollowing hyperpartisan influencers allows us to explore a potential process by which social media may causally contribute to partisan animosity, and may also inform interventions for reducing partisan animosity and misinformation spread[42–46].

Overview

We conducted three studies to test the impact of (un)following hyperpartisan influencers on partisan animosity and news-sharing behavior. Blatant misinformation on social media is rare [47,48], and biased or misleading information may be more harmful because of its broad reach[44,49].

Thus, we focused on "hyperpartisan" influencers who share a combination of politically-biased, misleading, and false news. We selected a list of hyperpartisan influencers for participants to unfollow. Then, in a correlational study, we found that following these hyperpartisan Twitter/X influencers was associated with partisan animosity, and that these hyperpartisan influencers tend to share low-quality news and use toxic language.

Afterwards, in a pilot field experiment (*Experiment 1*), we randomly assigned Twitter/X users to unfollow the hyperpartisan influencers we validated in our correlational analysis and follow non-political accounts that tweet about science and nature for one month. This follow/unfollow treatment reduced out-party animosity, improved well-being, and led to increased satisfaction with participants' Twitter/X feed.

Next, we replicated and extended these results in a pre-registered, large-scale field experiment with a more targeted sample of political Twitter/X users who unfollowed a greater proportion of hyperpartisan accounts (*Experiment 2*). This study replicated the effects of unfollowing on partisan animosity and satisfaction with one's Twitter/X feed, and separately found that following science accounts improved well-being, primarily by increasing feelings of awe. The effect of unfollowing on out-party animosity persisted for at least six months. Unfollowing also improved the quality of news accounts participants shared and reduced the amount of political content participants reported seeing one year later. Both experiments were pre-registered, and data, code, materials, and pre-registrations can be found on OSF: https://osf.io/e3jfk.

Finally, since *Experiment 2* was conducted during a time in which Twitter/X underwent a number of design changes as a result of Elon Musk's purchase of the platform, we conducted exploratory analysis to investigate how people's online experiences and behavior shifted after this purchase. After Elon Musk's purchase of Twitter/X, we found that participants used the platform less frequently, viewed their feeds as less reliable, and engaged with lower-quality news.

Experiment 1 - Follow/Unfollow Experiment

Selecting Accounts to Follow/Unfollow. For both experiments, we selected hyperpartisan accounts for participants to unfollow (60 accounts in *Experiment 1*, and 112 accounts in *Experiment 2*). We selected these accounts using a combination of existing lists of media bias and news source quality (such as *AllSides*, *MediaBiasFactCheck*, and *NewsGuard*), focusing on both biased and low-quality sources. We also selected political "influencers" (e.g, highly followed accounts)[34] and politicians, focusing primarily on politicians who had higher "falsity" scores (e.g., had made a number of false fact-checked claims) according to an existing dataset[50]. We selected a politically-balanced set of accounts for participants to unfollow (e.g., 30 conservative-leaning accounts and 30 liberal-leaning accounts in *Experiment 1*). We also prioritized selecting accounts that tweeted frequently and were highly followed at the time of the experiment.

We then conducted a correlational analysis to test whether following these hyperpartisan accounts was associated with partisan animosity. Specifically, we analyzed a dataset of Twitter/X data linked to survey data (n = 1,477, M_{age} = 39.34, SD_{age} = 12.53, 768M, 458F, 30

non-binary/other, 1011 liberal, 254 conservative), collected via an app called "Have I Shared Fake News" that was widely shared online[51]. Using the now discontinued free Twitter/X API, we downloaded the follower lists of the Twitter/X users who used this app and answered questions about their partisan animosity. The number of conservative-leaning hyperpartisan accounts a participant followed in this sample was negatively associated with favorability toward the Democratic party, r = -0.25, 95% CI = [-0.29, -0.20], $p < 0.001$, and the number of liberal-leaning hyperpartisan accounts a participant followed was negatively associated with favorability toward the Republican party, r = -0.18, 95% CI = [-0.23, -0.13], $p < 0.001$. In *Experiment 2*, we conducted a more detailed correlational analysis to examine the content posted by these hyperpartisan accounts.

We also selected 18 accounts for people to follow that primarily tweeted about science, nature, and space and 8 "placebo" accounts (e.g., @Walmart) for the control conditions to unfollow. See ***Supplementary Appendix S1*** for more details on account selection, and ***Supplementary Appendices S2-S7*** for a full list of accounts people were asked to follow and unfollow in both studies.

Procedure. For this pilot experiment, we recruited United States participants using Twitter/X ads targeting users who were following accounts similar to those in our list of accounts to unfollow using Twitter's "follower look-alike" settings, as well as via Twitter posts sent from the study authors. We randomly assigned 60% of participants to unfollow all hyperpartisan accounts that they were currently following *and* follow all science accounts they were not yet following for one month in exchange for an incentive of $8.50 (in addition to $21 total for completing all surveys). While the follow and unfollow treatments were bundled into one condition in the *Experiment 1* pilot, this was not true for the *Experiment 2* replication. 40% of participants were assigned to a *control condition*, in which they were instructed to unfollow the "placebo" accounts.

Participants filled out pre- and post-treatment surveys immediately before or immediately after the one-month experiment. These surveys included *1)* two questions that asked how participants felt toward the opposing party and one's own party over the past month using "feeling thermometers" [12], *2)* 10 questions that assessed participants' subjective well-being over the past month, and *3)* 12 questions assessing how positive people felt toward their Twitter/X feed over the past month. We focused on participants' feelings over the past month since this was the length of the intervention and since this phrasing was similar to phrasing from prior social media "deactivation" experiments[12]. See ***Supplementary Appendix S8*** for the specific wording of these measures. We also included a number of secondary outcome variables (the results for all secondary outcome variables are reported in ***Supplementary Appendix S9***). Our final sample consisted of 494 participants (308 male, 124 female, 24 non-binary/transgender/other, M_{age} = 28.14, SD_{age} = 9.06, 393 Democrat, 100 Republican) who filled out pre- and post-treatment surveys, passed at least one attention check, and provided valid Twitter/X handles. View ***Supplementary Appendix S1*** for the extended procedure and descriptive statistics.

Compliance. Compliance with the follow/unfollow treatment was measured each week via the Twitter/X Application Programming Interface (API). Twitter/X posting data was also retrieved each week from participants via the Twitter/X API. Averaging across the four weeks of the

experimental treatment, participants in the treatment condition followed 8.73 (95% CI = [7.73, 9.72]) more science accounts compared to the control condition, $t(370.78) = 17.24$, $p < .001$, $d = 1.79$, and followed 0.79 (95% CI = [0.09, 1.50]) fewer hyperpartisan accounts compared to the control condition, $t(391.65) = 2.23$, $p = 0.027$, $d = 0.23$. In other words, participants in this pilot experiment followed about nine accounts and unfollowed about one account (in *Experiment 2,* people unfollowed a much larger proportion of accounts due to more targeted recruiting).

Results. We ran mixed-effects models that probed for an interaction between condition (treatment vs. control) and time (pre-treatment vs. post-treatment) with random intercepts for participants for all main outcome variables. Standardized beta coefficients are reported for ease of interpretation. The follow/unfollow treatment improved participants' feelings toward the out-party over the past month, $\beta = 0.13$, 95% CI = [0.02, 0.24], $p = 0.025$. It did not significantly impact feelings toward the in-party over the past month ($p = 0.448$), but it did reduce affective polarization (in-party feelings - out-party feelings), $\beta = -0.16$, 95% CI = [-0.30, 0.03], $p = 0.020$. The treatment also improved well-being, $\beta = 0.18$, 95% CI = [0.04, 0.31], $p = 0.011$, and led participants to be more satisfied with their Twitter/X feeds, $\beta = 0.16$, 95% CI = [0.01, 0.31], $p = 0.039$. The treatment did not significantly improve the quality of news shared on Twitter/X during the treatment month (all *ps* > 0.101). See ***Supplementary Appendix S9*** for all regression models and robustness checks.

Experimental 2 - Large-Scale Replication with Longitudinal Results

We conducted a pre-registered replication study in a larger, more targeted sample of participants who followed a higher number of hyperpartisan accounts pre-treatment. This replication also included larger incentives for compliance, an expanded list of accounts to unfollow, and longitudinal measures that tested the long-term effects of the experiment. Furthermore, due to the bundled follow and unfollow treatments in *Experiment 1*, we separately tested the effects of following and unfollowing together versus just unfollowing in *Experiment 2*.

Selecting Accounts to Unfollow. For the large-scale replication, we expanded the list of hyperpartisan influencers to 112 accounts (56 conservative-leaning and 56 liberal-leaning). For this study, we closely focused on having participants unfollow accounts that were considered low-quality by NewsGuard, an organization that rates the news quality of an extensive list of sources[50,52]. For example, we had people unfollow left-leaning accounts such as @PalmerReport, which has a NewsGuard score of 34.5 out of 100 (with 100 indicating most trustworthy), and had people unfollow right-leaning accounts such as @BreitbartNews, which has a NewsGuard score of 49.5 out of 100. The average NewsGuard quality score of accounts we asked people to unfollow was 54.42 out of 100 (a score below 60 is considered "untrustworthy" according to NewsGuard). As in *Experiment 1*, following the new list of hyperpartisan accounts was negatively associated with favorability toward the opposing party for both the conservative ($r = -0.25$, 95% CI = [-0.30, -0.21], $p < .001$) and liberal-leaning accounts ($r = -0.17$, 95% CI = [-0.22, -0.12], $p < .001$). We also selected a list of 17 science accounts that was similar to the one in *Experiment 1* for participants to follow. The average NewsGuard score of these science accounts was 100.

Analysis of Content Posted by Accounts. To better understand the type of content removed from and added to participant's feeds, we analyzed all tweets posted by the 112 hyperpartisan

accounts we asked a group of participants to unfollow and the 17 science accounts we asked another group of participants to follow during the month of the experimental treatment. The average NewsGuard quality score of URLs posted by the hyperpartisan accounts was significantly lower (M = 59.17, SD = 27.47) than the average NewsGuard quality score of URLs posted by the science accounts (M = 99.87, SD = 1.49), $t(10561.21) = 145.26$, $p < .001$, $d = 2.09$. We also analyzed the toxicity of language posted by the hyperpartisan and science-based accounts using the Perspective API toxicity classifier[53], finding that the hyperpartisan accounts posted significantly more toxic language (M = 0.10, SD = 0.13) than the science accounts (M = 0.04, SD = 0.07), $t(4720.08) = 41.65$, $p < .001$, $d = 0.56$.

Recruitment. Participants were recruited solely via Twitter/X advertisements that targeted individuals who were following partisan accounts similar to the ones we wanted people to unfollow using Twitter/X's "follower look-alike" targeted advertisement settings. 7,339 Twitter/X users took a pre-recruitment survey where they supplied their Twitter/X handles. We invited 2,446 individuals who 1) followed at least one hyperpartisan account or 2) shared at least one low-quality news URL (as rated by NewsGuard) on their timeline prior to the treatment to participate in the study. 1,317 individuals began the study, and we included in our analyses the 1,133 participants (M_{age} = 45.17, 640 M, 450 F, 29 non-binary/other, 803 Democrat, 287 Republican) who completed both the pre- and post-treatment surveys. We found no evidence of differential attrition across experimental conditions (see ***Supplementary Appendix S1*****)**.

Procedure. Participants were randomly assigned to one of three conditions. In the two unfollow conditions (conditions one and two), participants were again asked to unfollow any accounts from the list of hyperpartisan accounts that they were currently following. In the follow/unfollow condition (condition one), participants were additionally asked to follow all of the science accounts they were not currently following. In the control condition (condition three), participants were asked to unfollow the same eight irrelevant "placebo" accounts as before. To assess the long-term effects of the intervention, we sent out follow-up surveys one month, six months, and 11 months after the treatment was completed. Twitter/X data was collected from participants once a week starting one week before the treatment and up to 12 weeks after the treatment (view ***Supplementary Appendix S1*** for the extended procedure).

Compliance. Compliance with the experimental treatments was considerably higher than in *Experiment 1*, likely because of the targeted recruitment strategy, the higher incentives for compliance, and the more extensive list of accounts to unfollow (see ***Figure 1***). During the treatment month, participants in the unfollow conditions followed 9.11 (95% CI = [7.89, 10.33]) fewer hyperpartisan accounts on average compared to the control condition, $t(478.07) = 14.69$, $p < .001$, $d = 1.06$. The distribution of hyperpartisan accounts unfollowed varied widely, with one participant unfollowing 57 accounts (see ***Supplementary Appendix S1*** for histograms of the distributions of accounts followed/unfollowed). During the treatment month, participants in the follow/unfollow condition followed 15.18 (95% CI = [14.75, 15.61]) more science accounts on average than the participants in the control condition, $t(355.36) = 69.46$, $p < .001$, $d = 5.37$. This means that participants in the two unfollow conditions unfollowed more than nine accounts on average, and participants in the follow/unfollow condition followed more than 15 accounts on average.

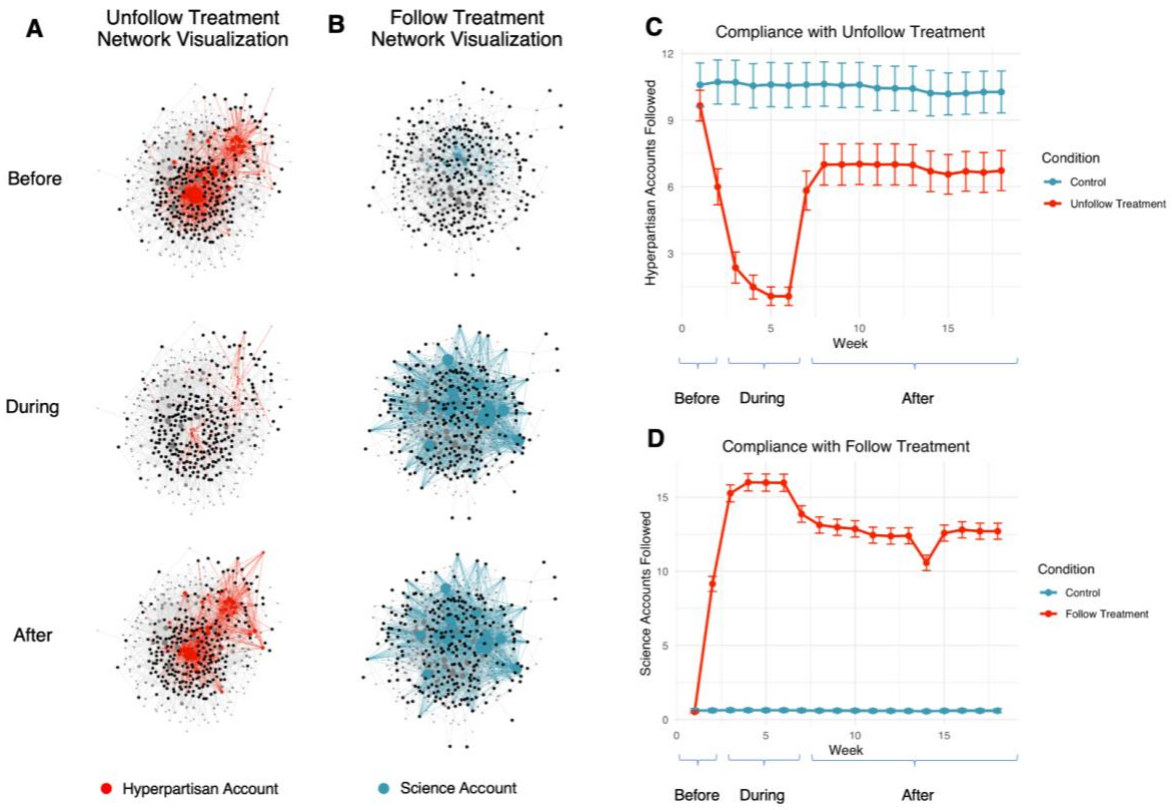

Figure 1. Panel A shows a network visualization of *Experiment 2* participants (black dots) assigned to the unfollow conditions, the hyperpartisan accounts they followed (red dots), and a random subset of the other accounts they followed (gray dots) one week before the treatment, during the final week of the treatment, and ten weeks after the treatment. **Panel B** shows a network visualization of *Experiment 2* participants in the follow condition (black dots), the science accounts they followed (blue dots), and a random subset of other accounts they followed (grey dots) one week before the treatment, during the final week of the treatment, and ten weeks after the treatment. **Panel C** shows the mean number of hyperpartisan accounts followed two weeks before, four weeks during, and up to twelve weeks after the treatment in the unfollow and control conditions. **Panel D** shows the mean number of science accounts followed two weeks before, four weeks during, and up to twelve weeks after the treatment in the follow and control conditions. On average, during the treatment month, participants in the unfollow conditions were following 9.11 fewer hyperpartisan accounts than those in the control, and participants in the follow condition were following 15.18 more science accounts than those in the control. Only 42% of participants chose to re-follow at least one of the hyperpartisan accounts when we explicitly gave them the opportunity to do so at the end of the treatment month, and only 26% chose to unfollow at least one science account. Error bars in panels **C** and **D** represent 95% confidence intervals.

Pre-Registered Main Outcomes. Following our pre-registration, we collapsed the two unfollow conditions and compared them to the control condition for our main analysis (see additional analyses broken down by condition in ***Supplementary Appendix S10***). Replicating *Experiment 1*, the unfollow conditions improved feelings toward the out-party over the past month as compared to the control condition, $\beta = 0.14$, 95% CI = [0.05, 0.24], $p = 0.003$ (see ***Figure 3***). More specifically, recent feelings toward the out-party increased by 17.6% in the experimental condition and decreased by 5.9% in the control condition, representing a net 23.5% increase in out-party attitudes in the experimental condition relative to control. The treatment had a stronger impact on out-party animosity for conservatives as compared to liberals (see ***Supplementary***

Appendix S11 for moderation analyses), contrasting with prior work that shows smaller effects of many interventions for conservatives[54,55].

As in *Experiment 1*, feelings toward the in-party over the past month once again remained unchanged, β = 0.01, 95% CI = [-0.07, 0.10], p = 0.751, and the polarization index (feelings toward the out-party minus feelings toward the in-party) was non-significant but directionally similar, β = -0.07, 95% CI = [-0.15, 0.010], p = 0.089. Replicating *Experiment 1*, participants in the unfollow conditions reported a more positive Twitter/X feed, β = 0.13, 95% CI = [0.03, 0.22], p = 0.013. The unfollow conditions had no significant impact on well-being (p = 0.752).

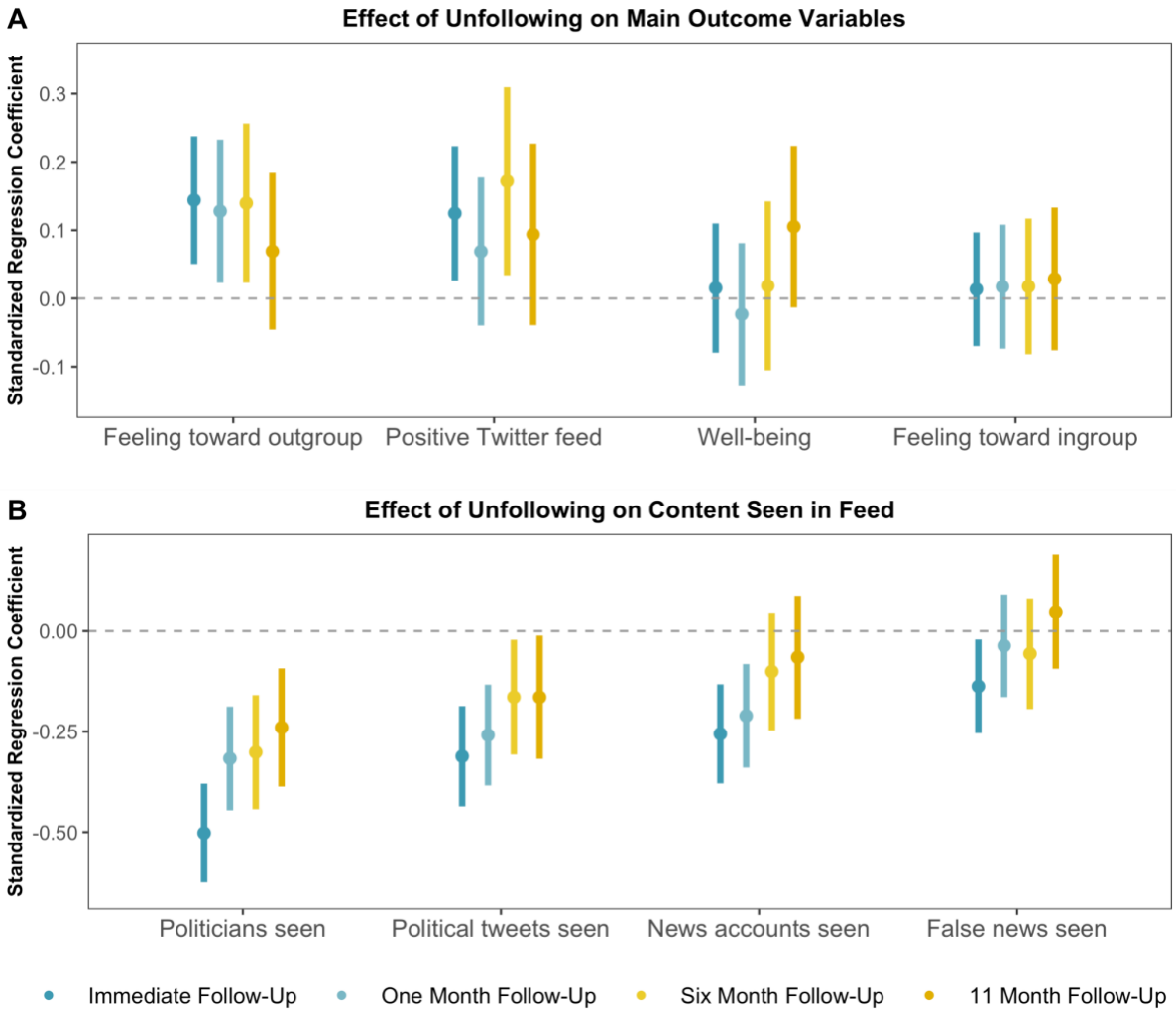

Figure 2. Unfollowing hyperpartisan accounts on social media durably reduced out-party animosity, with effects persisting up to six months after the treatment was completed. It also reduced the amount of political content people reported seeing in their feed up to eleven months after the treatment was completed, or one year after the experiment began. **Panel A** shows intent-to-treat effects of the two unfollow treatments on the main pre-registered outcome variables at the immediate follow-up, the one-month follow-up, the six-month follow-up, and the 11-month follow-up. **Panel B** shows intent-to-treat effects for variables that measured content seen in one's feed at the immediate follow-up and all subsequent follow-ups. Error bars represent 95% confidence intervals.

Follow treatment. The follow/unfollow condition improved well-being compared to the unfollow only treatment, β = 0.15, 95% CI = [0.04, 0.26], p = 0.009, but did not affect any of the

other main dependent variables (all *ps* > 0.349). This suggests that the well-being effect in *Experiment 1* may have been driven by following science accounts instead of unfollowing hyperpartisan accounts.

Enduring effects. We ran exploratory analysis to test the enduring effects of this intervention. The unfollow treatment still had a lasting effect on out-party animosity when measured one month after the treatment was over, β = 0.13, 95% CI = [0.02, 0.23], *p* = 0.017 (see ***Figure 2***). The other main outcome variables remained unchanged after one month (all *ps* > 0.110). The unfollow treatment continued to have a significant effect on out-party animosity when measured six months after the treatment ended, β = 0.14, 95% CI = [0.02, 0.26], *p* = 0.019. The unfollow treatment's impact on out-party animosity was no longer significant when measured at the 11-month follow-up (*p* = 0.238). Exploratory analysis found that these long-term effects were largely driven by the subset of participants who did not refollow hyperpartisan accounts (see ***Supplementary Appendix S10*** for additional analysis of this subset). Exploratory analysis further found that, for the subset of participants who did not refollow hyperpartisan accounts (as compared to the control condition), the reduction in out-party animosity was still significant at the 11-month follow-up, β = 0.13, 95% CI = [0.00, 0.27], *p* = 0.049 (see ***Supplementary Appendix S10***).

Effects on information diet. People who unfollowed hyperpartisan accounts reported seeing less false news (β = -0.14, 95% CI = [-0.25, -0.02], *p* = 0.021), less content about politics (β = -0.31, 95% CI = [-0.44, -0.19], *p* < 0.001), less news (β = -0.26, 95% CI = [-0.38, -0.13], *p* < 0.001), and fewer posts from politicians (β = -0.50, 95% CI = [-0.63, -0.38], *p* < 0.001) (see ***Figure 2*** and ***Supplementary Appendix S10***). Eleven months after the treatment ended (and one year after the experiment began), participants in the unfollow treatment still reported seeing less political content (β = -0.16, 95% CI = [-0.32, -0.01], *p* = 0.036) and fewer politicians (β = -0.24, 95% CI = [-0.39, -0.09], *p* < 0.001). To provide a more detailed picture of which aspects of people's feeds changed, ***Figure 3*** shows specific items of the positive perceptions of one's feed index and the well-being index (see ***Supplementary Appendix S10*** for full regression results).

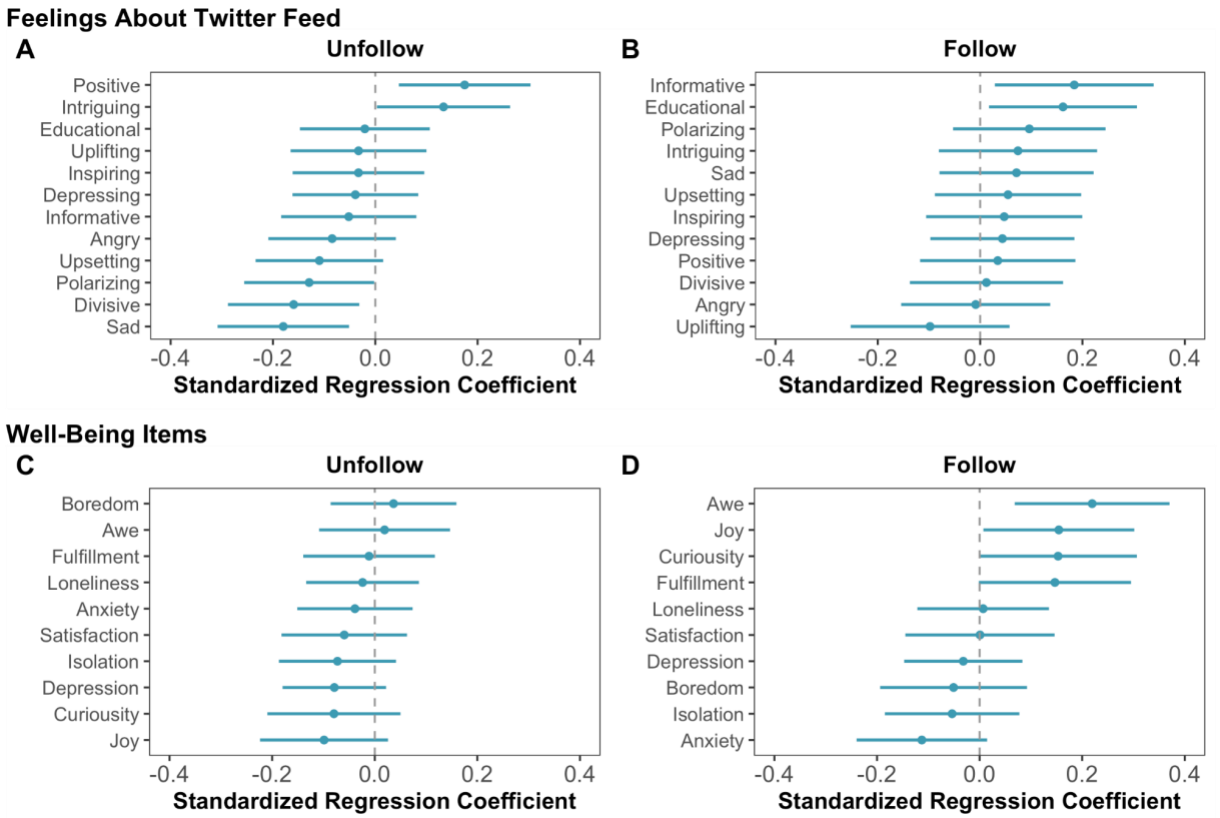

Figure 3. Effects of follow and unfollow conditions on feelings toward one's feed immediately post-treatment. **(Panel A)** People in the unfollow conditions found their feeds to be more positive, more intriguing, less divisive, and less sad compared to the control condition. **(Panel B)** People in the follow/unfollow condition (compared to the unfollow only condition) perceived their feeds to be significantly more informative and more educational. **(Panel C)** While the unfollow only treatment did not impact well-being, **(Panel D)** people in the follow/unfollow condition (compared to the unfollow only treatment) reported feeling increased well-being, which was primarily driven by increased feelings of awe, joy, curiosity, and fulfillment. See ***Supplementary Appendix S10*** for full results.

Behavioral Effects. To analyze the change in quality of news shared on Twitter/X, we used difference-in-difference models that tested for an interaction between condition and time (one month before vs. one-month after), with standard errors clustered at the user-level, following past analyses of Twitter news sharing behavior[56]. The pre-registered outcome variables of interest were the NewsGuard ratings of news URLs (e.g., nytimes.com) liked (or favorited) and reposted (or retweeted) by participants, as well as the NewsGuard ratings of news accounts (e.g., @NYTimes) liked and posted by participants. The unfollow treatment did not significantly improve the quality of news *links* posted or liked by participants during the month of the treatment (*ps* > 0.087). However, it did improve the quality of news *accounts* that participants reposted, $\beta = 0.45$, 95% CI = [0.09, 0.82], $p = 0.016$, and liked, $\beta = 0.90$, 95% CI = [0.40, 1.39], $p < 0.001$ (see ***Figure 4*** and ***Supplementary Appendix S10***). The different results for links versus accounts may be explained by the fact that we specifically had participants unfollow accounts that were rated as low-quality by NewsGuard. In other words, the news account variable may have been more closely related to (and thus more impacted by) the experimental treatment.

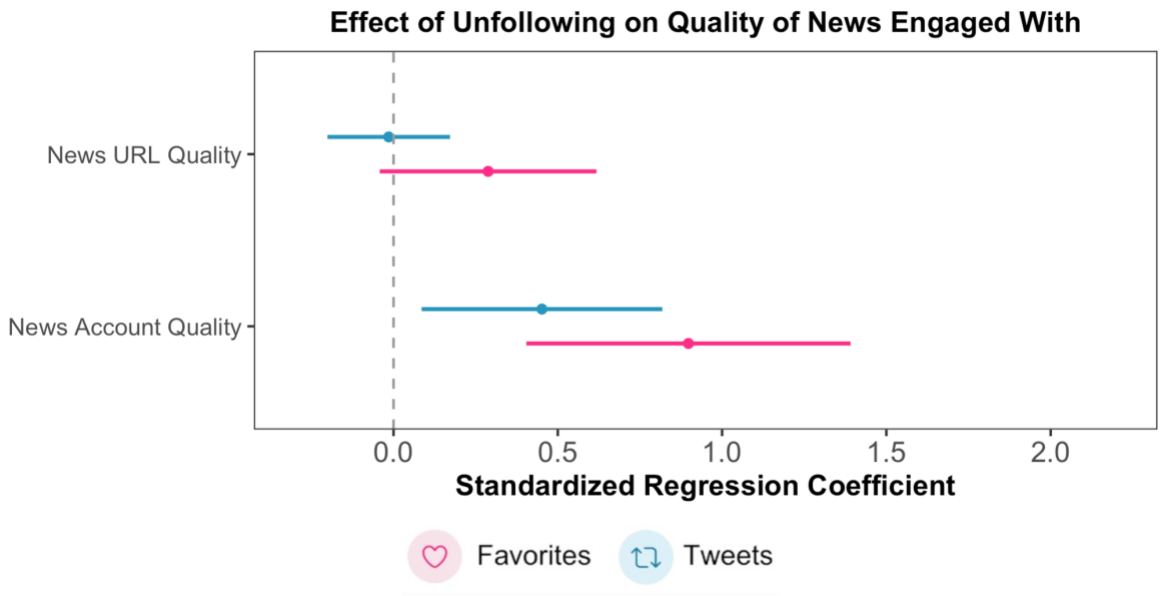

Figure 4. The unfollow treatment improved the quality of news accounts reposted and liked on Twitter/X (as determined by NewsGuard) during the month of the treatment compared to the month before. The unfollow treatment did not, however, improve the quality of news URLs posted or liked. Error bars represent 95% confidence intervals.

Long-Term Changes to Twitter/X. This experiment occurred during a period (March 2023-March 2024) in which Twitter/X underwent many design changes following Elon's Musk's purchase of the platform on October 28, 2022. For instance, in April 2023, Twitter/X removed "verified" checkmarks, which confirm the identifies of prominent people or organizations, and laid off 80% of its staff. Twitter/X was officially re-named to X in July 2023. Many speculated that these changes may have led to an increase in hostile and false content on the platform[57], and some analyses have shown politically contentious accounts[58] and posts containing hate speech[59] were amplified after Elon Musk's purchase of the platform.

We conducted exploratory analysis to see how participants' online experiences and behavior have changed during this period. A mixed-effects model with random intercepts for participants revealed a main effect of time (pre-treatment to 11 months post-treatment), indicating that all participants, irrespective of experimental condition, reported perceiving their Twitter/X feeds as less positive over the year-long study period, $\beta = -0.34$, 95% CI = [-0.45, -0.24], $p < 0.001$. Furthermore, over the course of the year-long study, participants also reported seeing less reliable content in their feeds, $\beta = -0.47$, 95% CI = [-0.60, -0.33], $p < 0.001$, and reported using Twitter/X much less frequently, $\beta = -0.54$, 95% CI = [-0.68, -0.41], $p < 0.001$ (see ***Figure 5*** and ***Supplementary Appendix S10*** for all models).

Importantly, the unfollow treatment (as compared to the control) did not lead to a significant decrease in self-reported Twitter/X engagement at any time point (all *ps* > 0.239), nor did it lead to a decrease in the number of likes or Twitter/X posts (all *ps* > 0.490) (see ***Supplementary Appendix S10*** for all regression models), assuaging concerns that the unfollow treatment may

have led to a decrease in engagement. Instead, declines in engagement were found across all participants over the course of the year-long experiment—regardless of experimental condition.

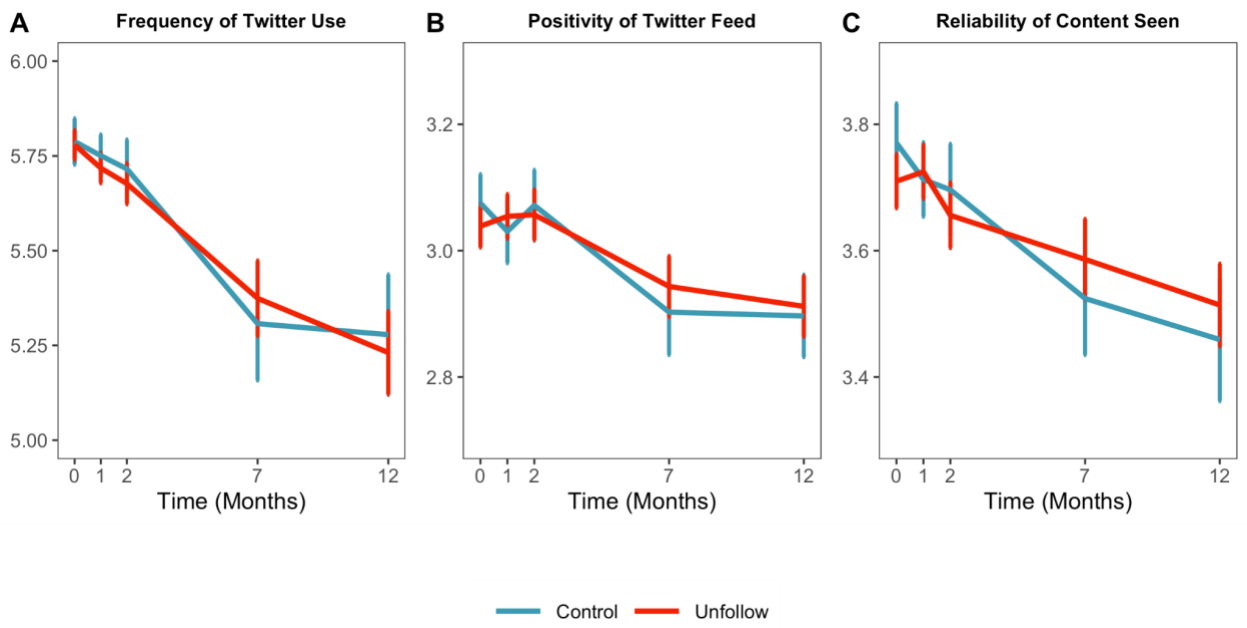

Figure 5. Regardless of experimental condition, there were several long-term changes that occurred from when this experiment began (March 2023) to when this experiment was completed (March 2024). Specifically, people used Twitter/X less frequently (**Panel A**), viewed their Twitter/X feeds as less positive (**Panel B**), and viewed the content in the feeds as less reliable (**Panel C**). This study was conducted during a period of notable design changes made to Twitter/X shortly after Elon Musk's purchase of the platform. The unfollow condition did not reduce self-reported engagement with Twitter/X at any time point, nor did it affect the number of likes or posts recorded by participants during the treatment month. All y-axes represent five-point Likert scales. Specific scale wording is reported in ***Supplementary Appendix S8***. Error bars represent 95% confidence intervals.

To see whether Twitter/X behavior changed before versus after Elon Musk's purchase of the platform, we also conducted exploratory analysis to investigate whether the quality of news people posted and liked on Twitter/X declined after Elon Musk's purchase. We ran regression models with a binary variable indicating whether a tweet was posted or liked before or after before or after October 28, 2022—the day Elon Musk purchased Twitter/X. As in our prior Twitter/X analysis, standard errors were clustered at the user-level. We used all tweets and favorites that we were able to collect from participants using the Twitter/X API up until the free API was shut down in July 2023[60]. The average NewsGuard quality score of news links (β = -0.21, [-0.35, -0.07], p = 0.004) and news accounts (β = -0.39, 95% CI = [-0.60, -0.17], p < 0.001) shared on Twitter/X was significantly lower after, as opposed to before, Elon's Musk's purchase of the platform. We found similar effects for the NewsGuard quality score of URLs liked (β = -0.22, 95% CI = [-0.39, -0.06], p = 0.009), and found a non-significant effect trending in the same direction for news accounts favorited (β = -0.31, 95% CI = [-0.77, 0.15], p = 0.182). While we cannot causally attribute any of these trends to Elon Musk's platform changes, they support speculations about the increase in misinformation spread after his purchase[57].

Discussion

In a correlational study and two pre-registered field experiments, we found that following hyperpartisan social media influencers is associated with out-party animosity, and that unfollowing these influencers reduces out-party animosity. The effect of unfollowing persisted at least six months after the experiment was completed—and nearly one year after the experiment was completed for the majority of participants who chose to not re-follow the hyperpartisan accounts. This work illustrates the long-term, causal impact of exposure to hyperpartisan content on attitudes and behaviors in a real-world setting.

Unfollowing also improved the quality of news accounts people interacted with, led people to feel more satisfied with their Twitter/X feeds, and reduced the amount of political content people reported seeing in their feeds a full year later. Following influencers who shared content about science increased well-being, primarily by increasing feelings such as awe and curiosity. Unfollowing hyperpartisan accounts did not reduce overall Twitter/X usage (both measured via self-report and Twitter/X behavior), even though self-reported engagement with Twitter/X declined steeply over the course of this experiment—which happened shortly after Elon Musk's purchase of the platform.

We also observed that the perceived reliability of news participants saw on Twitter/X and the perceived positivity of participants' feeds decreased over the course of the year-long experiment. These changes were also reflected in behavior, with participants engaging with lower quality news after Elon Musk's purchase of the platform. While we cannot make causal claims, this exploratory analysis supports speculations that the spread of misinformation and harmful content increased after Elon Musk's purchase of the platform[57].

The long-term effects of unfollowing stand in stark contrast to the typical fleeting effects of many "light-touch" interventions in the social and behavioral sciences[61,62]. Even though unfollowing only took a few minutes, it was a structural intervention that caused lasting changes to participants' daily information diet for at least a year[62–64]. Supplementary analysis (***Supplementary Appendix S11***) indicated that participants used Twitter/X for more than an hour each day on average, and used Twitter/X more than other social media platforms and media sources. Targeting Twitter/X "power-users" who follow more political elites than the average user[65] and use Twitter/X more frequently may have contributed to the enduring effects of this intervention.

In contrast with prior analysis suggesting that much of online polarization is driven by self-selection[37–39], this study finds one's daily social media information diet has an enduring causal impact on partisan animosity and news sharing behavior. Moving beyond experiments that estimate the average effect of social media usage among the general population[11–13], this experiment reveals a specific process (e.g., following hyperpartisan influencers) by which social media might contribute to partisan animosity. These experiments suggest that a small handful of extreme influencers may have a powerful impact on social media users.

Future work should further explore the psychological processes underlying the effects of this intervention. Prior work has shown that exaggerated out-party "meta-perceptions," or inflated beliefs about how much one's out-group dislikes one's in-group, can contribute to partisan animosity[33,66]. Exposure to polarizing content about one's out-group on social media may cause

individuals to perceive their out-group as more extreme and polarized than it actually is[18,32], which, in turn, can increase partisan animosity. Consistent with this account, mediation analysis found that the treatment's effect on partisan animosity was significantly mediated by out-party meta-perceptions (see ***Supplementary Appendix S10***).

This work may shed light on why prior research[21] has found that breaking open one's echo chamber by following cross-partisan elites backfires: political elites on social media frequently use divisive rhetoric[23,25], which may make people think their opposing party is more extreme than it actually is. Instead of merely breaking open one's online echo chamber, this research suggests a better anecdote to polarization may be to avoid exposure to hyperpartisan social media voices altogether and seek out more constructive political voices instead.

In addition to this theoretical contribution, this work also has practical applications for interventions[43,45]. Social media platforms could promote healthier information diets through, for instance, altering their algorithms and changing other design features. Social media platforms such as Facebook already boost trustworthy (and downrank untrustworthy) news during election times[67], meaning that these algorithmic solutions are feasible. These solutions would also not be censorship or interfere with individual autonomy, as the majority people already report wanting social media to be less divisive[24,30]. Thus, these solutions would align people's social media experience with their stated preferences. If platforms choose not to implement algorithmic solutions, individuals can also curate their information diets without any cooperation from social media companies.

It could be argued that unfollowing political elites or news accounts—even hyperpartisan ones—could lead to negative consequences, such as reduced political knowledge or engagement. Assuaging these concerns, we found null effects of the intervention on news knowledge, political attitudes, social media engagement, and a number of other variables (see ***Supplementary Appendix S11***). Other work, however, has demonstrated the beneficial effects of exposure to news on social media[68], and even noted the potential upsides of online outrage for spurring political participation and activism[69]. Thus, there might be tradeoffs between reducing exposure to low-quality, toxic, and hyperpartisan content and increasing exposure to high-quality news and important political information.

Conclusions

While many studies have highlighted the potential benefits of eliminating social media entirely[11–13], this approach is all-or-nothing: like a sledgehammer, it smashes both the harmful and beneficial aspects of one's online experience. In contrast, unfollowing hyperpartisan accounts is a more precise approach: like a scalpel, it surgically removes the negative aspects of social media while preserving the positive aspects. We hope this research will lead to a more nuanced discussion about the specific components of social media that cause harm and the ways that social media can be improved.

References

1. Van Bavel, J. J., Rathje, S., Harris, E., Robertson, C. & Sternisko, A. How social media shapes polarization. *Trends in Cognitive Sciences* (2021).
2. Kubin, E. & von Sikorski, C. The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association* **45**, 188–206 (2021).
3. Boxell, L., Gentzkow, M. & Shapiro, J. M. Greater Internet use is not associated with faster growth in political polarization among US demographic groups. *Proceedings of the National Academy of Sciences* **114**, 10612–10617 (2017).
4. Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. & Lazer, D. Fake news on Twitter during the 2016 US presidential election. *Science* **363**, 374–378 (2019).
5. Baribi-Bartov, S., Swire-Thompson, B. & Grinberg, N. Supersharers of fake news on Twitter. *Science* **384**, 979–982 (2024).
6. Kumar, D., Hancock, J., Thomas, K. & Durumeric, Z. Understanding the Behaviors of Toxic Accounts on Reddit. in *Proceedings of the ACM Web Conference 2023* 2797–2807 (ACM, Austin TX USA, 2023). doi:10.1145/3543507.3583522.
7. Number of worldwide social network users 2028. *Statista* https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.
8. Harris, E., Rathje, S., Robertson, C. & Van Bavel, J. J. The SPIR Model of Social Media and Polarization: Exploring the Role of Selection, Platform Design, Incentives, and Real-World Context. *International Journal of Communication* (In Press).
9. Haidt, J. Why the Past 10 Years of American Life Have Been Uniquely Stupid. *The Atlantic* **11**, (2022).
10. Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C. R. & Hertwig, R. How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behaviour* 1–8 (2020) doi:10.1038/s41562-020-0889-7.
11. Asimovic, N., Nagler, J., Bonneau, R. & Tucker, J. A. Testing the effects of Facebook usage in an ethnically polarized setting. *Proceedings of the National Academy of Sciences* **118**, (2021).
12. Allcott, H., Braghieri, L., Eichmeyer, S. & Gentzkow, M. The welfare effects of social media. *American Economic Review* **110**, 629–76 (2020).
13. Allcott, H. *et al.* The effects of Facebook and Instagram on the 2020 election: A deactivation experiment. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2321584121 (2024).
14. Sunstein, C. R. *# Republic: Divided Democracy in the Age of Social Media*. (Princeton University Press, 2018).
15. Rathje, S., He, J. K., Roozenbeek, J., Van Bavel, J. J. & van der Linden, S. Social media behavior is associated with vaccine hesitancy. *PNAS Nexus* (2022).
16. Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W. & Starnini, M. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* **118**, (2021).
17. Nyhan, B. *et al.* Like-minded sources on Facebook are prevalent but not polarizing. *Nature* 1–8 (2023).
18. Bail, C. *Breaking the Social Media Prism*. (Princeton University Press, 2021).
19. Hobolt, S. B., Lawall, K. & Tilley, J. The polarizing effect of partisan echo chambers. *American Political Science Review* 1–16 (2023).
20. Törnberg, P. How digital media drive affective polarization through partisan sorting. *Proc.*

Natl. Acad. Sci. U.S.A. **119**, e2207159119 (2022).
21. Bail, C. A. *et al.* Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* **115**, 9216–9221 (2018).
22. Yu, X., Wojcieszak, M. & Casas, A. Partisanship on Social Media: In-Party Love Among American Politicians, Greater Engagement with Out-Party Hate Among Ordinary Users. *Polit Behav* (2023) doi:10.1007/s11109-022-09850-x.
23. Rathje, S., Van Bavel, J. J. & van der Linden, S. Out-group animosity drives engagement on social media. *Proc Natl Acad Sci USA* **118**, e2024292118 (2021).
24. Heltzel, G. & Laurin, K. Why Twitter sometimes rewards what most people disapprove of: The case of cross-party political relations. (2024).
25. Frimer, J. A. *et al.* Incivility is rising among American politicians on Twitter. *Social Psychological and Personality Science* 19485506221083811 (2022).
26. Brady, W. J., Gantman, A. P. & Van Bavel, J. J. Attentional capture helps explain why moral and emotional content go viral. *Journal of Experimental Psychology: General* **149**, 746 (2020).
27. Antypas, D., Preece, A. & Camacho-Collados, J. Negativity spreads faster: A large-scale multilingual twitter analysis on the role of sentiment in political communication. *Online Social Networks and Media* **33**, 100242 (2023).
28. Robertson, C. E. *et al.* Negativity drives online news consumption. *Nature Human Behaviour* 1–11 (2023).
29. Van Bavel, J. J., Robertson, C. E., Del Rosario, K., Rasmussen, J. & Rathje, S. Social Media and Morality. *Annu. Rev. Psychol.* **75**, 311–340 (2024).
30. Rathje, S., Robertson, C., Brady, W. J. & Van Bavel, J. J. People think that social media platforms do (but should not) amplify divisive content. *Perspectives on Psychological Science* (2023).
31. Milli, S., Carroll, M., Pandey, S., Wang, Y. & Dragan, A. D. Twitter's Algorithm: Amplifying Anger, Animosity, and Affective Polarization. *arXiv preprint arXiv:2305.16941* (2023).
32. Robertson, C., del Rosario, K. & Van Bavel, J. J. Inside the Funhouse Mirror Factory: How Social Media Distorts Perceptions of Norms. *Current Opinion in Psychology* (2024).
33. Moore-Berg, S. L., Ankori-Karlinsky, L.-O., Hameiri, B. & Bruneau, E. Exaggerated meta-perceptions predict intergroup hostility between American political partisans. *Proceedings of the National Academy of Sciences* (2020).
34. Diresta, R. *Invisible Rulers: The People Who Turn Lies into Reality*. (2023).
35. Pennycook, G. & Rand, D. G. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* **116**, 2521–2526 (2019).
36. Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J. & Stein, B. A Stylometric Inquiry into Hyperpartisan and Fake News. Preprint at http://arxiv.org/abs/1702.05638 (2017).
37. Waller, I. & Anderson, A. Quantifying social organization and political polarization in online platforms. *Nature* **600**, 264–268 (2021).
38. Robertson, R. E. *et al.* Users choose to engage with more partisan news than they are exposed to on Google Search. *Nature* 1–7 (2023).
39. González-Bailón, S. *et al.* Asymmetric ideological segregation in exposure to political news on Facebook. *Science* **381**, 392–398 (2023).
40. Guess, A. M. *et al.* Reshares on social media amplify political news but do not detectably

affect beliefs or opinions. *Science* **381**, 404–408 (2023).
41. Guess, A. M. *et al.* How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science* **381**, 398–404 (2023).
42. Hartman, R. *et al.* Interventions to reduce partisan animosity. *Nature human behaviour* **6**, 1194–1205 (2022).
43. Rathje, S. & van der Linden, S. Shifting online incentive structures to reduce polarization and the spread of misinformation. in *Research Handbook on Nudges and Society* 91–108 (Edward Elgar Publishing, 2023).
44. Van Der Linden, S. & Kyrychenko, Y. A broader view of misinformation reveals potential for intervention. *Science* **384**, 959–960 (2024).
45. van der Linden, S. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nat Med* 1–8 (2022) doi:10.1038/s41591-022-01713-6.
46. Rathje, S., Roozenbeek, J., Van Bavel, J. J. & van der Linden, S. Accuracy and social motivations shape judgements of (mis) information. *Nature human behaviour* 1–12 (2023).
47. Guess, A., Nagler, J. & Tucker, J. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science advances* **5**, eaau4586 (2019).
48. Budak, C., Nyhan, B., Rothschild, D. M., Thorson, E. & Watts, D. J. Misunderstanding the harms of online misinformation. *Nature* **630**, 45–53 (2024).
49. Allen, J., Watts, D. J. & Rand, D. G. Quantifying the impact of misinformation and vaccine-skeptical content on Facebook. *Science* **384**, eadk3451 (2024).
50. Mosleh, M. & Rand, D. G. Measuring exposure to misinformation from political elites on Twitter. *nature communications* **13**, 7144 (2022).
51. Lapowski, I. Newsguard wants to fight fake news with humans, not algorithms. *Wired, August* **23**, (2018).
52. Lin, H. *et al.* High level of correspondence across different news domain quality rating sets. *PNAS nexus* **2**, pgad286 (2023).
53. Avalle, M. *et al.* Persistent interaction patterns across social media platforms and over time. *Nature* **628**, 582–589 (2024).
54. Rathje, S. Letter to the editors of Psychological Science: Meta-analysis reveals that accuracy nudges have little to no effect for US conservatives: Regarding Pennycook et al. (2020). *Psychological Science* (2022).
55. Martel, C. *et al.* On the Efficacy of Accuracy Prompts Across Partisan Lines: An Adversarial Collaboration. *Psychol Sci* **35**, 435–450 (2024).
56. Bor, A., Osmundsen, M., Rasmussen, S. H. R., Bechmann, A. & Petersen, M. B. ‘ Fact-checking’ videos reduce belief in misinformation and improve the quality of news shared on Twitter. (2020).
57. Myers, S. L., Thompson, S. A. & Hsu, T. The Consequences of Elon Musk’s Ownership of X. *The New York Times* (2023).
58. Barrie, C. Did the Musk takeover boost contentious actors on Twitter? *Harvard Kennedy School Misinformation Review* **4**, 1–19 (2023).
59. Hickey, D. *et al.* Auditing Elon Musk’s impact on hate speech and bots. in *Proceedings of the international AAAI conference on web and social media* vol. 17 1133–1137 (2023).
60. Rathje, S. To tackle social-media harms, mandate data access for researchers. *Nature* **633**, 36–36 (2024).
61. Lai, C. K. *et al.* A comparative investigation of 17 interventions to reduce implicit racial preferences. *Journal of Experimental Psychology: General* **143**, 1765–1785 (2014).

62. Paluck, E. L., Porat, R., Clark, C. S. & Green, D. P. Prejudice reduction: Progress and challenges. *Annual Review of Psychology* **72**, (2020).
63. Guess, A. M., Barberá, P., Munzert, S. & Yang, J. The consequences of online partisan media. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2013464118 (2021).
64. Broockman, D. & Kalla, J. Consuming cross-cutting media causes learning and moderates attitudes: A field experiment with Fox News viewers. *Journal of Politics* (2024).
65. Wojcieszak, M., Casas, A., Yu, X., Nagler, J. & Tucker, J. A. Most users do not follow political elites on Twitter; those who do show overwhelming preferences for ideological congruity. *Sci. Adv.* **8**, eabn9418 (2022).
66. Ruggeri, K. *et al.* The general fault in our fault lines. *Nature Human Behaviour* **5**, 1369–1380 (2021).
67. Bagchi, C. *et al.* Social media algorithms can curb misinformation, but do they? Preprint at http://arxiv.org/abs/2409.18393 (2024).
68. Altay, S., Hoes, E. & Wojcieszak, M. News on Social Media Boosts Knowledge, Belief Accuracy, and Trust: A Field Experiment on Instagram and WhatsApp.
69. Spring, V. L., Cameron, C. D. & Cikara, M. The upside of outrage. *Trends in Cognitive Sciences* **22**, 1067–1069 (2018).

Funding:

Russell Sage Foundation (SR, JVB)
Tanner Verhey/Google Trust and Safety (JR, SVL, SR)
Center for the Science of Moral Understanding (SR)
Templeton World Charity Foundation; TWCF-2023-31570 (SR, JVB, doi.org/10.54224/31570)
AE Foundation (SR, JVB)
National Science Foundation, Grant #2334148 (SR, JVB)
National Science Foundation SBE Postdoctoral Research Fellowship, Grant #2404649
Gates Cambridge Scholarship; Grant #OPP1144 (SR)
Heterodox Academy (SR, JVB)

Author contributions:

Conceptualization: SR, JVB
Methodology: SR, CP, TH, SVL, JVB
Investigation: SR, CP, TH, SVL, JVB
Visualization: SR, JKH, CP
Funding acquisition: SR, JR, KG, SVL, JVB
Project administration: SR
Supervision: JVB, KG, SVL, JVB
Writing – original draft: SR
Writing – review & editing: SR, CP, TH, JKH, JR, KG, SVL, JVB

Competing interests: Authors declare no competing interests.

Data and materials availability: De-identified data, code, replication materials, and pre-registrations, are available on our OSF: https://osf.io/e3jfk.

Acknowledgements: We thank Tanner Verhey for feedback on this manuscript, Dan-Mircea Mirea for help with statistical analysis and R code, and Yara Kyrychenko for help with network analysis. We also are thankful for comments from the Social Identity and Morality Lab, the Social Decision-Making Lab, and the Polarization and Social Change Lab.

Correspondence and requests for materials should be addressed to srathje@alumni.stanford.edu and jay.vanbavel@nyu.edu.