Social Media, News Consumption, and Polarization: Evidence from a Field Experiment

Ro'ee Levy[*]

June 23, 2020

Abstract

Does the consumption of ideologically congruent news on social media exacerbate polarization? I estimate the effects of social media news exposure by conducting a large field experiment randomly offering participants subscriptions to conservative or liberal news outlets on Facebook. I collect data on the causal chain of media effects: subscriptions to outlets, exposure to news on Facebook, visits to online news sites, and sharing of posts, as well as changes in political opinions and attitudes. Four main findings emerge. First, random variation in exposure to news on social media substantially affects the slant of news sites individuals visit. Second, exposure to counter-attitudinal news decreases negative attitudes toward the opposing political party. Third, in contrast to the effect on attitudes, I find no evidence that the political leaning of news outlets affects political opinions. Fourth, Facebook's algorithm is less likely to supply individuals with posts from counter-attitudinal outlets, conditional on individuals subscribing to them. Together, the results suggest that social media algorithms may limit exposure to counter-attitudinal news and thus increase polarization.

JEL Codes: D72, L82, L86, O33

[*] Yale University, roee.levy@yale.edu. I am deeply grateful to Ebonya Washington, Joseph Shapiro, and Dean Karlan for their guidance and support throughout this project. I thank Eran Amsalem, Ian Ball, Dirk Bergemann, Leonardo Bursztyn, Alex Coppock, Oren Danieli, Eduardo Fraga, Matthew Gentzkow, Matthew Graham, Alexey Makarin, Martin Mattsson, Brendan J. Nyhan, Rohini Pande, David Rand, Oren Sarig, David Schönholzer, Katherine Wagner, and Mor Zoran for helpful comments and advice. I also thank seminar participants at Berkeley, Booth, Cornell, Microsoft Research, Northwestern, NYU, UCLA Anderson, and Yale for their valuable comments. Financial support from the ISPS Field Experiment's Initiative, the Tobin Center for Economic Policy, the Yale Department of Economics, the Yale Program in Applied Economics and Policy, the Yale School of Management, and David Rand is greatly appreciated. All errors are my own. The experiment is registered at the AEA RCT registry (ID 0002713).

In 2019, more than 70% of American adults consumed news on social media, compared to fewer than one in eight Americans in 2008.[1] Based on Pew surveys, Facebook is the dominant social media platform for news consumption, and "among millennials, Facebook is far and away the most common source for news about government and politics" (Pew, 2014). As social media becomes a major news source, there are growing concerns that individuals are exposed to more pro-attitudinal news, defined as news matching their ideology, and as a result, polarization increases (Sunstein, 2017).

In this paper, I test whether these concerns are warranted. I analyze the effects of exposure to pro- and counter-attitudinal news outlets by conducting a large online field experiment randomizing exposure to news on Facebook, and by collecting survey, browsing, and social media data.

To motivate the experiment, I first provide descriptive statistics on online news consumption. I show that social media, and specifically Facebook, tends to expose people to more segregated, pro-attitudinal, and extreme news, compared to other news sites visited.

I recruited American Facebook users to the experiment using Facebook ads. After completing a baseline survey, participants were randomly assigned to a liberal treatment, a conservative treatment, or a control group. Participants in the liberal and conservative treatments were asked to subscribe to up to four liberal or conservative outlets on Facebook, respectively (e.g., MSNBC or Fox News), by clicking a "Like Page" button embedded in the survey.[2] Remarkably, in each treatment, approximately half the participants complied by subscribing to at least one outlet. When individuals subscribe to an outlet, posts shared by the outlet may subsequently appear in their Facebook feed. A post usually contains the story's headline and often includes a link to the full news story on the outlet's website.

I designed the experiment to have high external validity. A nudge offering subscriptions to outlets is very common on social media and participants could have subscribed to any of these outlets, at no cost, without the intervention. Besides the offer, the experiment did not directly intervene in any behavior. The news supplied to participants was the actual news provided by leading media outlets during the study period. Facebook's algorithm determined which of the posts shared by the outlets appeared in the participants' Facebook feeds. Finally, participants decided whether to skip, read, or share posts. As a result, the effects of the treatments are almost identical to the experience of millions of Americans who subscribe to news outlets on Facebook.

I estimate the effect of the intervention on exposure to news in the Facebook feed, news sites visited, news shared, political opinions, and affective polarization, defined as negative attitudes toward the opposing political party. Affective polarization is a primary outcome of interest since this measure of polarization has been increasing (Iyengar and Krupenkin, 2018), and there are

[1] 2008 figure is based on the Pew Research Center 2008 Biennial Media Consumption Survey. The 2019 figure is based on the the Pew Research Center American Trends Panel Wave 51, July 2019.

[2] To simplify terminology, throughout the paper I will describe the action of "liking" a page of a news organization as subscribing to an outlet on Facebook.

concerns over its implications for governance, accountability of elected officials, and even labor markets (Iyengar et al., 2019).

To measure subscriptions to outlets on Facebook and posts shared, I asked participants to log in to the survey using their Facebook account. To measure exposure to news in the Facebook feed and visits to news sites, I developed a Google Chrome extension to collect this data for a subset of participants who were offered the extension and installed it (the extension was only offered to participants who took the survey on a computer using Google Chrome). To estimate the effect on opinions and attitudes, I invited participants to an endline survey approximately two months after the intervention. My sample is composed of 37,494 participants who completed the baseline survey. 34,592 of those participants provided access to their posts for at least two weeks, 1,835 installed the extension for at least two weeks, and 17,635 took the endline survey.

This paper has four main findings. First, exposure to news on social media substantially affects online news consumption. Following an increased exposure to posts from the randomly offered outlets, participants visited the news sites of the outlets, even when the outlets did not match their ideology. Visiting the websites had an economically and statistically significant effect on the mean slant of participants' online news consumption. The difference between the intention-to-treat (ITT) effect of the liberal and conservative treatments on the slant of all news sites visited is 14% of the difference in the slant of sites visited by liberals and conservatives in the control group.

Various economic theories explain why individuals optimally choose to consume news that matches their ideology.[3] However, I find that news consumption strongly responds to an exogenous shock to the feed, meaning that individuals often consume news incidentally, and do not re-optimize their browsing behavior to keep the slant of the news sites they visit constant. The results imply that social media algorithms can substantially alter news consumption habits, and that while social media is associated with pro-attitudinal news, individuals are willing to engage with counter-attitudinal news when it is made more accessible on social media.

My second finding is that exposure to counter-attitudinal news *decreases* affective polarization, compared to pro-attitudinal news. I construct an affective polarization index measuring attitudes toward political parties. The index includes questions such as how participants feel toward their own party and the opposing party (i.e., a "feeling thermometer"). The ITT and treatment-on-treated (TOT) effects of the counter-attitudinal treatment on the index, compared to the pro-attitudinal treatment, are -0.03 and -0.06 standard deviations, respectively.[4] The TOT effect should be interpreted as the effect on individuals who subscribe to new outlets when nudged to subscribe. Comparing each treatment to the control group suggests that the effect on polarization is driven

[3] This could occur since outlets sharing the consumer's ideology convey more useful information (Chan and Suen, 2008), provide direct utility (Mullainathan and Shleifer, 2005), or are perceived to be of higher quality (Gentzkow and Shapiro, 2006). See Gentzkow et al. (2015) for a review.

[4] A pro-attitudinal treatment is defined as a liberal treatment assigned to a liberal participant or a conservative treatment assigned to a conservative participant, and a counter-attitudinal treatment is defined as a liberal treatment assigned to a conservative participant or a conservative treatment assigned to a liberal participant.

by the counter-attitudinal treatment but this result should be interpreted cautiously since participants in the control group were more likely to complete the endline survey (there is no differential attrition between the two treatment arms).

I compare the results to existing benchmarks by focusing on the feeling thermometer questions which have been asked in many previous surveys. The experiment's ITT and TOT effects decreased the difference between participants' feelings toward their own party and the opposing party by 0.58 and 0.96 degrees on a 0-100 scale, respectively. For comparison, according to the American National Election Survey (ANES), this measure of affective polarization increased by 3.83-10.52 degrees between 1996 and 2016.[5]

Third, in contrast to the effect on attitudes, I do not find evidence that the slant of news outlets affects political opinions. The effect of the liberal and conservative treatments on a political opinions index, focusing on issues and political figures covered during the study period, such as the Mueller investigation, is economically small, precisely estimated, and not statistically significant.

The paper's fourth finding is that Facebook's algorithm may limit exposure to counter-attitudinal news. I show that participants in the counter-attitudinal treatment were exposed to substantially fewer posts on Facebook from the outlets they subscribed to in the intervention, compared to participants in the pro-attitudinal treatment.

Combined, the results paint a complicated picture. On the one hand, Facebook's algorithm seems to filter counter-attitudinal news, probably since it attempts to personalize news based on the user's behavior and perceived interests. While it is not possible to estimate the effect of specific posts filtered by the algorithm, I show that exposure to counter-attitudinal news decreases affective polarization. This suggests that social media algorithms may be increasing polarization. On the other hand, this paper also shows that individuals are willing to engage with counter-attitudinal news, and social media platforms provide a setting where a subtle nudge can substantially diversify news consumption and consequently decrease polarization.

Literature

This paper contributes to the literature on social media and news consumption. In his seminal book "The Filter Bubble," Eli Pariser warned that the "era of personalization is here" (Pariser, 2011). However, recent reviews of the literature concluded that "we lack convincing evidence of algorithmic filter bubble in politics" (Guess et al., 2018). Papers in this literature typically estimate segregation in online news based on cross-sectional analysis of browsing behavior.[6] These papers usually lack social media data and cannot measure segregation *within* one's social media

[5] The increase in polarization depends on the weights and the respondents included in the sample. When using only the ANES face-to-face sample for consistency, as used by Boxell et al. (2018), the increase is 3.83. When including also the web sample in 2016, as used by Iyengar et al. (2019), the increase is 10.52. The ANES top codes the thermometer at 97 degrees. The results stay almost exactly the same when I top code the results in the same way.

[6] See Flaxman et al. (2016), Gentzkow and Shapiro (2011), Guess (2018), and Peterson et al. (2019).

feed. One exception is a paper by Bakshy et al. (2015), arguing that exposure to counter-attitudinal news is mostly limited by individual choices and not by algorithmic ranking. The paper analyzes rich Facebook data but does not exploit exogenous variation.[7] I advance the literature by generating experimental variation in subscriptions to outlets and collecting data on exposure to posts from those outlets. This allows me to decompose the mechanisms limiting exposure to counter-attitudinal news and demonstrate the existence of a filter bubble, i.e., that Facebook's algorithm is more likely to expose individuals to news matching their ideology, conditional on subscription.

My findings contribute to the literature on social media and polarization by generating variation in the main mechanism through which social media is suspected to increase polarization: the distance between individuals' ideology and the slant of the news they consume. Related papers show that the Internet and Facebook may increase polarization (Allcott et al., 2020; Lelkes et al., 2015), but based on demographics, they are probably not a primary driver in the rise of polarization (Boxell et al., 2018).[8] Since these papers focus on the reduced-form effect of social media, they do not identify the causal effect of pro- or counter-attitudinal news. Indeed, a recent review of the literature argued that "it is far from clear ... that partisan news actually causes affective polarization" (Iyengar et al., 2019). To the best of my knowledge, this paper provides the first experimental evidence that counter-attitudinal news decreases affective polarization. It advances the literature by showing *how* social media can affect polarization and by providing evidence that nudges diversifying news exposure on social media can be effective.

My study also contributes to a well-established literature on media persuasion by randomly assigning subscriptions to news outlets. Both survey experiments (e.g., Coppock et al. 2018) and papers with quasi-experimental designs (e.g., DellaVigna and Kaplan 2007) find that individuals are persuaded by the news they consume.[9] In many contexts, the "gold standard" for measuring causal effects is field experiments, as they combine the strong identification of lab experiments with high external validity. However, field experiments estimating media effects are not common. One notable exception is a study randomizing subscriptions to the Washington Post and Washington Times, which does not find an effect on opinions but is limited by a relatively small sample size (Gerber et al., 2009). This paper studies a different setting, social media, and shows how the unique features of this setting, specifically the algorithm, affect the supply of news. Focusing on social media also allows me to analyze engagement with news and quantify the effect of news exposure.

Methodologically, this paper contributes to a growing literature conducting online media-related experiments (Allcott et al., 2020; Chen and Yang, 2019; Jo, 2018; Mosquera et al., 2019) by demonstrating how an experiment can exploit social media's existing infrastructure to gradually dis-

[7] In addition, Bakshy et al. (2015) focus on posts shared by individuals' social networks, while I focus on posts shared by outlets individuals subscribe to, which are associated with greater segregation.

[8] Other studies estimating the effect of social media on political behavior include Bursztyn et al. (2019), Enikolopov et al. (2019), and Müller and Schwarz (2019). See Zhuravskaya et al. (2020) for a recent review.

[9] Other studies estimating media effects on political opinions and behavior include Chiang and Knight (2011), Durante et al. (2019), Gentzkow et al. (2011), and Okuyama (2019). See Strömberg (2015) for a review.

tribute news to participants in a natural setting. Along with Bail et al. (2018), who randomize exposure to content from liberal and conservative bots on Twitter, this paper is one of the first papers generating variation in social media feeds (Zhuravskaya et al., 2020). In contrast to most online experiments, participants were not asked to consume any content or continue complying with the treatment over time, and they did not receive notifications reminding them of the intervention besides the invitations to the endline survey. The natural, unobtrusive intervention means that it is unlikely that experimenter effects drive the study's result. To precisely detect the small effects that are expected as a result of a subtle intervention, I collect a sample size that is an order of magnitude larger than most other related experiments.

1 Background: Facebook

This study focuses on Facebook since it is the dominant social network, used by seven out of ten American adults. Most of these users visit Facebook several times a day,[10] and the platform accounts for 45% of all time spent on social media (Williamson, 2018). Despite its prominence, Facebook has been understudied, especially compared to Twitter (Guess et al., 2018).

The most important Facebook feature is the news feed, where users scroll through a list of posts curated by Facebook's algorithm. Posts in the feed are typically shared by the user's Facebook friends, shared by Facebook pages the user subscribes to ("likes"), or are sponsored posts (advertisements shared by pages to promote content). The posts may include text, video, pictures, and links.

Facebook is a very popular source for news consumption. Approximately 52% of Americans get news on Facebook, more than the share of Americans getting news on all other social media platforms combined.[11] While this study focuses on the US, understanding the effect of Facebook has global implications. A 2018 survey among Internet users aged 16-64 in 43 countries estimated that 79% of users outside China use Facebook monthly (GlobalWebIndex, 2018). A survey by Reuters Institute found that in 37 out of 38 middle and high-income countries surveyed, more than 20% of the population consumed news through Facebook weekly (Reuters Institute, 2019). Facebook probably directly affects the news exposure of more individuals than any other company.[12]

With Facebook's growing influence, it has faced several controversies in recent years, including an effort by the Russian-based Internet Research Agency to influence the elections, the spread of fake news during the 2016 US election cycle, and Cambridge Analytica's attempt to assist campaigns with personally targeted ads. The concerns over each of these scandals were based on the assumption that individuals are easily persuaded by political information on social media.

[10] Facebook usage is based on the Pew Research Center January 2019 Core Trends Survey.

[11] Calculation based on the Pew Research Center American Trends Panel Wave 37.

[12] A recent paper analyzing data from the Reuters Institute report found that Facebook "reaches the widest international audience of any media organization in our sample" (Kennedy and Prat, 2019).

2 Design and Data

This section summarizes the experimental design, data, and empirical strategy. The design of the experiment, along with the subsamples analyzed are also presented in Figure 1 and Table 1.

2.1 Experimental Design

I recruited American adults to the experiment in February-March 2018 using Facebook ads.[13] Individuals who clicked the ads were directed to the survey landing page, where they reviewed the consent form and could begin the survey by logging in using their Facebook account. After logging in to the survey, and before treatment assignment, four *potential* liberal outlets and four *potential* conservative outlets were defined for each participant. The potential outlets were set such that they did not include outlets the participant already subscribed to on Facebook, to ensure only new outlets would be offered to participants. Toward the end of the survey, participants were randomly assigned to a liberal treatment, a conservative treatment or a control group, with the randomization blocked by participants' self-reported baseline ideology.[14] Participants in the conservative treatment were offered to subscribe to their four potential conservative outlets and participants in the liberal treatment were offered to subscribe to their four potential liberal outlets. Participants in the control group were not offered any outlets.

I nudged participants to subscribe to the outlets by explaining that subscribing could expose them to new perspectives. Participants were not required to subscribe to any outlet and did not receive monetary compensation for subscribing. The intervention did not provide exclusive access to these outlets, and any individual can subscribe to these outlets on Facebook at no cost and with minimum effort, regardless of the intervention. Since participants were logged into their Facebook account when taking the survey, the offer to subscribe to outlets was integrated within the survey, and the only action required by participants was to click the standard Like Page button.[15] Facebook users often encounter this button, for example when Facebook suggests pages they may be interested in or when outlets purchase ads promoting their page.

After participants subscribed to an outlet by "liking" its Facebook page, posts from the outlet appeared in their feeds, among many other posts, according to Facebook's algorithm. Participants decided whether to read a post, click a link, share a post or unsubscribe from an outlet, just like the decisions they make regarding other posts appearing in their feed. Due to the simple common intervention, the organic nature of any subsequent effect, and the fact that participants were not

[13] 978,628 people saw the ads and 87,648 people clicked the link in the ads. The ads are discussed in Appendix A.3.1.

[14] Respondents were asked where they position themselves ideologically on a 7-point ideological scale from very liberal to very conservative, with an option of "I haven't thought about it much." Each block is composed of three sequential participants who chose the same answer among the eight ideological scale options. The first participant in a block was randomly assigned to one of the three treatment groups, the second participant was randomly assigned to one of the two remaining groups, and the third participant was assigned to the remaining group.

[15] The button was generated using Facebook's Page Plugin. Appendix Figure A.1 provides an example survey page with the intervention.

reminded of the intervention, experimenter effects are unlikely to play a large role in explaining the effects, at least compared to similar studies.[16] Because individuals can subscribe to outlets on Facebook at no cost and no monetary incentives were provided, the intervention is scalable.

2.2 The Setting: Media Outlets and the News Environment

The primary liberal outlets offered in the experiment are Huffington Post, MSNBC, The New York Times, and Slate. The primary conservative outlets are Fox News, The National Review, The Wall Street Journal, and The Washington Times. The news outlets were chosen to ensure participants are offered a diverse set of popular outlets (Fox News and the New York Times are two of the three most popular news pages on Facebook) with a clear ideological slant. If a participant already subscribed to a primary liberal outlet or a primary conservative outlet, the outlet was replaced with an alternative liberal or conservative outlet, respectively.[17] Appendix Table A.1 displays the full list of outlets offered.

Figure 2 shows that the most prominent men and women mentioned in posts shared by the primary outlets are political figures. Unsurprisingly, President Trump is the dominant figure mentioned. Important political stories covered during the study period can be observed in the figure: Trump's alleged affair with Stormy Daniels, Robert Mueller's investigation, and the negotiation with North Korea's leader, Kim Jong Un. The figure also demonstrates that liberal outlets focused on scandals related to the presidency and mentioned Michael Cohen, Stormy Daniels, Scott Pruitt, and Vladimir Putin, much more often than conservative outlets.

2.3 Data Collection and Subsamples

2.3.1 External Data

Outlets I measure the slant of news at the outlet level, the common method used in the literature. I determine an outlet's slant according to a dataset by Bakshy et al. (2015) defining the slant of 500 news domains based on the self-reported ideology of Facebook users sharing articles from the domains. Using this definition, a completely liberal outlet has a slant of approximately negative one, a middle-of-the-road outlet has a slant of approximately zero, and a completely conservative outlet has a slant of approximately one. The dataset correlates well with other measures of slant (e.g., Gentzkow and Shapiro, 2010). I refer to outlets in this dataset as *leading news outlets*. I determine the Facebook pages of leading outlets by searching for pages with names similar to each outlet's domain and manually checking the pages. Facebook pages were found for 371 outlets.

[16] Participants were asked at the end of the survey what they think is the purpose of the study. Appendix C.1 shows that participants understood the study was about media and politics and that there do not appear to be dramatic differences between the answers of participants in the pro- and counter-attitudinal treatments.

[17] Approximately 55% of participants did not subscribe in baseline to any of the primary conservative and liberal outlets. The effects on political beliefs are robust to including only these participants.

Comscore Browsing Data To provide descriptive statistics on news consumption outside the experimental sample, I analyze the 2017 and 2018 Comscore WRDS Web Behavior Database Panel. Each observation in the dataset is a domain visited by an individual along with the referral domain. I merge this dataset with the list of leading news outlets (Bakshy et al., 2015). The combined 2017 and 2018 datasets include 94,342 individuals who visited at least one news site.[18] I classify the channels through which visitors reached websites as social, search, or direct visits. Facebook is by far the dominant referral source in the social category.

For more details on the outlet and Comscore datasets, see Appendices A.1 and A.2, respectively.

2.3.2 Experiment Data

The analysis of the experiment relies on three datasets: self-reported survey data, Facebook data, and browser data. To the best of my knowledge, this is the first study combining experimental variation with social media and browsing data.

Survey Data The endline survey measures self-reported political opinions, affective polarization, and changes in news consumption habits. 17,635 participants took the endline survey and constitute the *endline survey subsample*.

Facebook Data on Pages Liked and Posts Shared Participants logged in to the survey using their Facebook account, through a Facebook app created for the project. They were asked to provide separate permissions to access the pages they subscribe to and posts they share. Providing permissions was voluntary, they could be revoked at any time, and were revoked automatically approximately two months after participants logged in to a survey. I observe all posts shared or pages liked until permissions are revoked. Since baseline subscriptions were required to define the potential outlets for each participant, only participants who provided permissions to access their subscriptions are included in the baseline sample.[19]

Data on posts shared in baseline is used to estimate the effect of the intervention on political behavior. I exclude posts sharing photos, albums, music, and events. The remaining posts typically include text with a link or an embedded video. Since posts shared are observable to the participant's social network or the general public, sharing posts can have a direct cost to the reputation of the participant. Approximately 92% of baseline participants provided access to the posts they shared for at least two full weeks following the intervention constituting the *access posts subsample*.

[18] Each observation includes a unique machine (computer) id, which I assume represents an individual. While Comscore attempts to identify unique individuals, it is still possible that multiple individuals use the same machine.

[19] Providing permission was not required to complete the survey or to be eligible for any rewards. The vast majority of participants who completed the survey provided these permissions. Participants who revoked permissions post-treatment are still included in the baseline sample.

Extension Data on Browser Behavior and the Facebook Feed Participants who completed the baseline survey using Google Chrome on a computer were asked to install a browser extension collecting Facebook feed data and news-related browsing behavior, in exchange for a small reward.[20] The offer was made toward the end of the survey, but before the intervention, to ensure take-up is not affected by the intervention. 2,262 of the 8,084 participants who were offered the extension, installed it. I focus on 1,835 participants who kept the extension installed for at least two weeks and constitute the *extension subsample*.

The Facebook feed data is used to analyze news exposure by estimating how often participants were exposed to posts from outlets on Facebook. I observe the posts that participants saw when they used their computer mouse to scroll their feed. I do not observe whether a post is a sponsored advertisement, but identify suspected ads as posts participants were exposed to from pages they did not subscribe to, posts appearing repeatedly and posts participants were exposed to over a long period of time. I attribute a post to a news outlet if it was created by the outlet's Facebook page or contains a link to the outlet's domain.[21] While the variation generated by the experiment is in subscriptions to the outlets' Facebook pages, I include in the analysis news articles shared by the participants' friends, to accurately capture total exposure to news outlets on Facebook.

The browsing behavior data is used to estimate the effect on the news sites participants visited. The extension can greatly reduce measurement error, compared to self-reported estimates of news consumption, especially since individuals' self-reported media habits may be more polarized than their actual media habits (Guess et al., 2017).

The extension data was only collected when participants used a computer while being signed into their Chrome account. In practice, individuals often use Facebook and browse news sites on a mobile device or at work, where they may use a different browser. Therefore, the estimates for the number of posts participants were exposed to in their feed and the number of sites they visited are lower bounds.[22]

Additional details on the survey, Facebook, and extension data can be found in Appendices A.3, A.4, and A.5, respectively.

Subsamples The datasets define three separate subsamples. To maximize power, when analyzing the effects on opinions and attitudes, I focus on the *endline survey subsample*. When analyzing media outcomes, I focus on the *extension subsample* and the *access posts subsample* (or their overlap).

[20] In exchange for installing the extension, participants could choose between receiving a $5 gift card, participating in a lottery with a $200 gift card, or receiving a copy of the study results.

[21] To match URLs with news outlets, I first convert over ten million URLs to their final endpoint, following redirects. This is required since many links on Facebook are based on URL-shortening services such as tinyurl.com.

[22] In the baseline survey, participants were asked how many links to articles about government and politics they clicked on Facebook in the past 24 hours using a computer and on a mobile phone. Among participants in the extension subsample who provided a numerical answer under 1,000, approximately 72% of news links were clicked on a computer, so it is likely that most, but not all data is collected for these participants.

Appendix Table A.2 presents descriptive statistics on the subsamples and shows that the extension subsample is more liberal and older, as would be expected when excluding participants who took the survey on a smartphone. The share of compliers is greater in the extension subsample, which assists in detecting treatment effects despite the smaller sample size.

2.4 Outcomes

2.4.1 Media

I measure subscriptions to outlets on Facebook, exposure to news in the Facebook feed, news sites visited, and posts shared, using the following quantitative outcome measures.

First, I estimate the direct effect of the experiment according to the number of times participants engaged with the *potential outlets* (the four liberal outlets and the four conservative outlets defined for each participant). For example, I measure the number of posts participants observed from their potential liberal and conservative outlets in their feed. Second, I measure the mean slant of all *leading news outlets* participants engaged with. Third, to measure the effects of the pro- and counter-attitudinal treatments on total news consumption, I define a *congruence scale*, calculated as the mean slant of news consumed, multiplied by (-1) for liberal participants. This scale has a higher value when individuals consume more extreme content matching their ideology. Fourth, I estimate the *share of counter-attitudinal news,* defined as the share of news from counter-attitudinal outlets among all news from pro- and counter-attitudinal outlets.

2.4.2 Opinions and Attitudes

I analyze the effects of news exposure on two primary outcomes: political opinions and affective polarization. For both outcomes, an index is composed by taking an average of all the valid non-missing index components and then standardized by subtracting the control group mean and dividing by the control group's standard deviation.

The political opinions index is composed of twenty survey questions focusing on domestic political issues and political figures covered in the news during the study period, such as new tariffs, the March For Our Lives Movement, and the investigation regarding Russian interference in the elections.[23] Each outcome variable is defined such that a higher value is associated with a more conservative opinion and then standardized.

I construct an affective polarization index composed of five outcomes. First, I use the feeling thermometer questions (*feeling thermometer*). Second, participants are asked how well the following statement describes them on a scale from 1 to 5: "I find it difficult to see things from

[23] The full list of questions is presented in Appendix Figure A.9.

Democrats/Republicans point of view" (*difficult perspective).* Third, participants are asked a similar question on the following statement: "I think it is important to consider the perspective of Democrats/Republicans" (*consider perspective*). Both statements are based on a political empathy index by Reit et al. (2017). Fourth, participants are asked if they think the Democratic and Republican parties have a lot (3), some (2), a few (1), or almost no good ideas (0) (*party ideas*). For each of the four previous measures, I calculate the difference between attitudes toward the participant's party and attitudes toward the other party, a typical measure of affective polarization. Fifth, to measure social-distance, participants are asked if they would feel very upset (2), somewhat upset (1), or not upset at all (0) if they had a son or daughter who married someone from the opposing party, either a Democrat or Republican (*marry opposing party*).[24] Each outcome variable is defined such that a higher value is associated with more polarization and then standardized.

2.5 Empirical Strategy

When estimating the effect of the intervention on engagement with the liberal and conservative outlets, the slant of news participants engaged with, and their political opinions, I compare the liberal and conservative treatments. When measuring the effect on polarization or engagement with pro- and counter-attitudinal outlets, it no longer makes sense to use these treatments (a conservative treatment is not expected to make participants more or less polarized than a liberal treatment), and therefore I focus on the pro-attitudinal and counter-attitudinal treatments. The strategy broadly follows the study's pre-analysis plan, discussed in Appendix B.2.

Liberal and Conservative Treatments I estimate the following ITT regression:

$$Y_i = \beta_1 T_i^L + \beta_2 T_i^C + \alpha X_i + \varepsilon_i \tag{1}$$

where $T_i^L, T_i^C \in \{0, 1\}$ is whether participant i is assigned to the liberal or conservative treatment, respectively. As defined in the pre-analysis plan, when estimating the effect on political opinions, I focus on the difference between the liberal and conservative treatments, by testing whether $\beta_1 < \beta_2$ (i.e., the conservative treatment made participants more conservative, compared to the effect of the liberal treatment). To increase power, when estimating the effect on political opinions, I

[24] Participants stating in the endline survey that they are Republicans or Democrats were asked how they would feel if they had a son or daughter who married a Democrat or Republican, respectively. Participants who did not explicitly identify with either party were asked about one of the parties randomly. I asked participants about the opposing party since I was concerned that respondents would find it odd to state how upset they would be if they had a son or daughter who married someone from their own party. However, conditioning the question on an endline variable could potentially bias the result. For example, if some baseline Democrats or Republicans were affected by the counter-attitudinal treatment, and as a result, no longer identified with their party, they were less likely to be asked how they feel about the opposing party in endline and the average participant asked about the opposing party would be slightly less moderate in this treatment arm. I include this measure in the affective polarization index since it is the only social-distance measure in the index, it is included in the pre-analysis plan, and the bias is expected to go against the direction of my findings. Appendix Table A.12 shows that the results are robust to excluding this measure from the index.

control for the following set of covariates, X: self-reported ideology, party affiliation, approval of President Trump, ideological leaning, age, age squared, gender, and baseline questions measuring political opinions that are similar to questions used in the endline survey. When estimating the effect on media outcomes, I only control for baseline outcomes, when they exist.[25] All regressions use robust standard errors unless noted otherwise. Appendix B.3 describes the control variables.

Pro-Attitudinal and Counter-Attitudinal Treatments I estimate the following ITT regression:

$$Y_i = \beta_1 T_i^A + \beta_2 T_i^P + \alpha X_i + \varepsilon_i \tag{2}$$

where $T^A \in \{0,1\}$ is whether the participant was assigned to the counter-attitudinal treatment, defined as a liberal treatment assigned to a conservative participant or a conservative treatment assigned to a liberal participant. $T^P \in \{0,1\}$ is whether the participant was assigned to the pro-attitudinal treatment, defined as a liberal treatment assigned to a liberal participant or a conservative treatment assigned to a conservative participant. X is the same set of control variables used when analyzing the effect on political opinions, with baseline measures of political opinions replaced with baseline measures of affective polarization. $\beta_1 < \beta_2$ tests whether individuals become more polarized when assigned to pro-attitudinal news, compared to counter-attitudinal news.

I determine whether participants are liberal or conservative (their ideological leaning) according to the following hierarchy: the party the participant identifies with or leans toward, her self reported ideology, and if the ideological leaning still cannot be determined, the candidate the participant preferred in the 2016 elections. I use this definition since it allows me to determine the ideological leaning of the vast majority of participants in the sample.[26]

2.6 Balance and Attrition

Table 2 presents descriptive statistics for participants in the baseline sample for the liberal treatment, conservative treatment, and control group, and shows the sample is balanced. Appendix Table A.3 presents a balance table according to whether the treatment matched the participant's ideology (pro- or counter-attitudinal), and shows that the sample is balanced along the redefined treatment arms as well. The sample size in this table is slightly smaller because it excludes participants for whom an ideological leaning cannot be defined.

Similar to other opt-in panels, the sample is not nationally representative. Participants tend to be more liberal than the US population and, as expected, more participants say that they get most of their news on social media (18%), compared to the national population (13%). The share of female

[25] Partial baseline data exists for posts shared and news sites visited, but not for posts observed on Facebook.

[26] Approximately 3% of participants do not self-identify as liberals or conservatives, did not identify with the Republican or Democratic party, and did not vote for Trump or Clinton. They are excluded from the analysis when analyzing the effect of the pro- and counter-attitudinal treatments. The effect on affective polarization is also robust to including only participants who identify with or lean toward the Democratic or Republican party.

participants and the average age is similar to the US population. Self-reported exposure to news on Facebook in line with one's views is similar to US Facebook users. Overall, the sample seems at least as representative as samples of Mechanical Turk users (Berinsky et al., 2012).[27]

Tables 2 and Appendix Table A.3 also test for differential attrition among the three subsamples. The access posts and extension subsamples have low attrition rates compared to baseline takeup (as shown in Table 1) and very small differences in attrition by treatment arm, and therefore their results are unlikely to be affected by attrition.[28] However, more participants completed the endline survey in the control group (48%), compared to the liberal (45%) and conservative (45%) treatment arms. The differential attrition mostly stems from participants in the conservative and liberal treatments not completing the final screen of the baseline survey after they encountered the intervention.[29]

Appendix Tables A.4 and A.5 present balance tables for the endline survey subsample and show that despite the attrition, the two treatment arms and control group are similar on observables. Participants in the pro-attitudinal treatment who completed the endline survey are *not* substantially more polarized in baseline than participants in the counter-attitudinal treatment. Moreover, there is no differential attrition between the conservative and liberal treatments and no differential attrition between the pro- and counter-attitudinal treatments. When estimating the effect on the primary endline survey outcomes, I compare the two treatment arms to each other to mitigate concerns over differential attrition. Still, it is possible that attrition could affect the results.

2.7 Compliance

Throughout the analysis, I focus on ITT estimates. To measure the effect of complying with the treatment, defined as subscribing to at least one offered outlet, I also analyze TOT estimators by regressing the dependent variable on compliance and instrumenting compliance with the random treatment assignment.[30] Since the intervention only offers new outlets to participants, defiers do not exist in this experiment.[31] Because compliance is defined as liking an outlet when it was

[27] One advantage of the sample is that Facebook users are not experienced, semi-professional survey takers. Participants were asked in the endline survey how many additional surveys they completed in the past month, the median answer is 1 and the mean answer is 7. For comparison, a 2014 study found that the median Mechanical Turk worker reported participating in 20 academic studies in the *week* before the question was asked (Rand et al., 2014).

[28] There is a very small, but statistically significant difference between the conservative treatment and the other groups in the number of participants who provided permissions to access their posts for two weeks following the intervention (the *Access Post, Two Weeks* variable). However, this minimal difference seems to be random, since it already existed before the intervention, as can be seen by the variable *Access Post, Pre-Treat.* There is no differential attrition in providing access to posts for at least two weeks among all participants who provided access before the intervention.

[29] These participants did not complete the survey either due to a technical issue that affected a small share of participants or since they preferred not to complete the survey after the intervention. As a result, they were less likely to provide their email address, and therefore, it was more challenging to recruit them to the endline survey.

[30] Compliance is measured using Facebook data. Participants were also asked in the baseline survey how many pages they subscribed to. For 88% of participants, the self-reported number equals the number measured using Facebook data, suggesting that data was collected properly and that generally, participants answered questions truthfully.

[31] Defying the experiment would mean unsubscribing from an offered outlet, but participants are only offered outlets they are not already subscribed to. There are rare cases where I only observe a partial list of outlets in baseline and as

offered, always-takers do not exist either.[32] If compliers are more likely to engage with the outlets and be affected by them (perhaps because they are more interested in the content or open to new opinions), the TOT is expected to be larger than the ATE.

In the entire baseline sample, 59% of participants who were offered pro-attitudinal outlets complied with the pro-attitudinal treatment and subscribed to at least one outlet, compared to 48% of participants offered counter-attitudinal outlets. Table 3 shows that participants were more likely to subscribe to outlets they are familiar with, to outlets with a perceived ideology similar to their own ideology, and to outlets they perceive as more moderate. Appendix Table A.6 presents descriptive statistics on the compliers by treatment arm and shows that liberals, women, and participants who subscribe to more outlets on Facebook were more likely to comply with both treatments. To test whether participants open to new ideas comply more often with the treatments, I use two questions from a brief measure of the big five personality domains, self-reported certainty in political opinions, and exposure to counter-attitudinal news in baseline. Based on these measures, participants complying with the counter-attitudinal treatment are slightly more open than non-compliers, but the differences are not large.

This section deals with immediate compliance with the intervention, which is especially useful when interpreting the TOT effects. However, the experiment is designed to allow participants to opt-out of information at any stage in the process. They can always unsubscribe from the offered outlets or ignore posts from the outlets appearing in their feed. Therefore, the effects found will probably be driven by participants who decide to consume the content offered when it becomes accessible. This feature increases the external validity of the results because these participants are often the policy-relevant population, as they are more likely to engage with the offered outlets in other circumstances as well.

3 Descriptive Analysis: Segregation in Online News Consumption

With the rise of social media, it is important to understand whether it is associated with different news consumption patterns. In this section, I present descriptive statistics on segregation in social media and online news. I use four main measures in the analysis.

First, *isolation* is the difference between the share of conservatives consuming news from the news sites that conservatives visit and the share of conservatives consuming news from the sites that

a result, a participant could have theoretically been offered an outlet she already subscribed to and "unliked" the page instead of "liking" it. However, I estimate that I observed a partial list of outlets for less than 1% of participants and I do not have evidence that participants unsubscribed from outlets as a result of the intervention.

[32] In a handful of cases participants subscribed to their potential outlets, even though the outlets were not offered, possibly since the survey included questions about these outlets. However, these cases are extremely rare and therefore, I am not defining them as compliance for simplicity. When focusing on the two weeks following the intervention instead of immediate compliance, an always-taker would be defined as a participant who would subscribe to a potential outlet in that period, regardless of the intervention. In the control group, only 0.2% and 0.5% of participants subscribed to a potential conservative or liberal outlet, respectively, in the two weeks following the intervention.

liberals visit. A higher value means that conservatives disproportionately visit news sites visited by other conservatives. To make the measure comparable to estimates by Gentzkow and Shapiro (2011), I aggregate all visits at the daily level and use the adjusted leave-out estimator of isolation (Gentzkow and Shapiro, 2011). Second, *segregation* is defined as the square root of the expected square distance between the slant of news sites visited by participants in the sample. To keep the measure in the unit interval, the slant of outlets is normalized to range from zero to one. Third, the *absolute value of slant* is the mean absolute value of user-level news consumption slant, where an outlet's slant is defined by Bakshy et al. (2015) and ranges from negative one to one. All the measures are formally defined in Appendix B.1.

In addition to these measures, I calculate the share of pro- and counter-attitudinal news for each medium. Similarly to Bakshy et al. (2015), I divide news sites into five quintiles: very liberal, liberal, moderate, conservative, and very conservative. An extreme pro-attitudinal outlet is defined as a very conservative outlet visited by a conservative or a very liberal outlet visited by a liberal.

To determine whether a site is pro-attitudinal and to calculate the isolation and congruence measures, I first define whether participants are liberal or conservative. For the extension data, I use the participants' ideological leaning as defined in Section 2.5. For Comscore data, I define conservatives as individuals living in zip codes with an above-median share of donations to Republican candidates in the 2016 and 2018 election cycles, based on FEC data.

3.1 Segregation in Online News

Table 4a uses 2018 Comscore data to show that news consumed through social media is more segregated and extreme than news consumed through all other channels. Rows (7)-(8) of Table 4b complement the analysis and show that news consumed through Facebook is also more segregated among control group participants in the extension subsample.[33]

The difference between segregation across news consumption channels could stem from differences in the individuals using these channels. Appendix Table A.7a presents the results for 8,882 individuals in the Comscore sample who visited multiple news sites through Facebook and multiple news sites through other means. As all individuals in this group consume news through both sources, the comparison better isolates the effect of the medium. While the share of news sites visited through Facebook is much greater among these individuals (26%), sites visited through Facebook remain substantially more segregated.

Figure 3a presents the distribution of the mean slant of news consumption for these individuals and shows that news sites visited through Facebook are more extreme. Through Facebook, 57% of individuals consume news that is on average more conservative than the Wall Street Journal or

[33] To compare the extension and Comscore samples, Appendix Table A.7c reanalyzes the extension data with ideology defined according to participants' zip codes. While the segregation measure is similar in the samples, participants visit more extreme sites in the extension subsample. The table also demonstrates that isolation and congruence are underestimated when using zip code as a proxy for ideology.

more liberal than the Washington Post, and among all other news sites visited, 39% of individuals consume such partisan news.[34] Figure 3b shows a clear correlation between the consumers' ideology and the slant of their news consumption. More importantly, the slope for news consumed through Facebook is steeper than the slope for news consumed through other means, indicating that sites visited through Facebook tend to better match the consumers' ideology.[35]

Has social media led to increased segregation generally in online news consumption? In the extension sample, the overall segregation level for all online news is 0.20, which is similar to a value of 0.25 found by Peterson et al. (2019) using 2016 data from the Wakoopa toolbar, and larger than a value of 0.11 found by Flaxman et al. (2016) using 2013 Bing toolbar data.[36] To compare isolation levels to previous estimates, I use visit-level measures of isolation, which give more weight to individuals who visit more news sites. The isolation index for browsing behavior in the extension sample is 0.22 (row 6 in Appendix Table A.8b), similar to a value of 0.21-0.24 calculated by Peterson et al. (2019) and larger than a value of 0.07-0.08 calculated by Gentzkow and Shapiro (2011).[37] Finally, Appendix Table A.7b compares segregation over time using Comscore data and does not find substantial changes in segregation between 2007-2008 and 2017-2018.

The analysis does not lead to conclusive results. Segregation online may have increased, but it probably did not change dramatically, perhaps due to the extent of social media usage. While Facebook is one of the top two most important traffic sources (along with Google), social media still accounts for a limited amount of traffic. For an average user in the Comscore sample, 4% of news sites were visited through Facebook. In the extension subsample, which only includes Facebook users, the figure is 14%. These estimates may underestimate Facebook usage since they rely on browsing activity on computers, while Facebook may be more popular on mobile.[38]

3.2 Segregation Within Facebook

Why does news consumed through Facebook tend to be more extreme and segregated? Two mechanisms that could increase segregation are homophily in social networks (an "echo chamber" effect) and the abundance of accessible, free media options allowing consumers to personalize their news feed. Rows (9) and (10) of Table 4b compare the segregation of news sites visited

[34] Washington Post and Wall Street Journal are in the 36th and 63th percentile of the Bakshy et al. (2015) dataset. When using the 25th and 75th percentile, which are similar to Boston Globe and Fox News, 19% of individuals consume on average partisan news when visiting news sites through Facebook and 5% consume such news outside Facebook.

[35] The figure also suggests that estimating segregation by comparing the news consumption of Republicans and Democrats, as is common in the literature, might mask important heterogeneity within Republicans and Democrats.

[36] To make the results comparable to Peterson et al. (2019), I also define the slant of outlets based on the participant's self-reported ideology. Using this definition, the segregation estimate is 0.23.

[37] While I attempt to make the samples as comparable as possible, each study still analyzes the data slightly differently. For example, Flaxman et al. (2016) limit their sample to individuals who regularly read online news, they determine the slant of the outlet according to the estimated share of Republicans among the outlet's readers and they estimate segregation via a hierarchical Bayesian model.

[38] For comparison, Parse.ly (2018) tracks pages viewed in thousands of sites and estimates that 16% of traffic related to Donald Trump in April-May 2018 is from social media and that Facebook is the largest external referral source for traffic in the law, government and politics category.

through links shared by Facebook friends and Facebook pages.[39] While both mechanisms are associated with increased segregation, links shared by Facebook pages are associated with greater segregation in all the measures. For example, the isolation index is 0.15 when participants visit news sites not through Facebook, 0.19 when they visit sites through friends, and 0.44 when they visit sites shared by Facebook pages. Therefore, it is important to study the forces determining which pages appear in the social media feed and the effect of posts shared by these pages.

Table 4b also provides a comparison of segregation in outlets individuals subscribe to on Facebook (row 1), posts they see in their feed (row 2), news sites they visit (row 8), and posts they share (row 11). The table shows that segregation is highest among subscriptions; that websites visited through Facebook have similar segregation to the Facebook feed; and that posts individuals share are slightly more segregated than the feed.[40]

To conclude, in a 2019 survey, 83% of Americans stated that one-sided news is a very big or moderately big problem on social media.[41] This section provides evidence that this concern is warranted, as it shows that Facebook is indeed more segregated and extreme than other online news. The next section estimates the causal effects of exposure to more and less segregated news using the random variations generated by the experiment.

4 Findings: Demand for News on Social Media

4.1 Individuals Are Willing to Engage with Counter-Attitudinal News

Figure 4 displays the effects of the pro- and counter-attitudinal treatments on engagement with the potential pro- and counter-attitudinal outlets, respectively. To keep the results comparable across media outcomes, the figure is calculated for the participants who both installed the browser extension and provided permissions to access their posts for at least two weeks. Each row in the figure is estimated by regressing engagement with the four potential pro- or four counter-attitudinal outlets in the two weeks following the intervention on the pro- or counter-attitudinal treatment. The control group is the reference group.[42]

The first panel of Figure 4 shows that the counter-attitudinal treatment increased the number of subscriptions to counter-attitudinal outlets by 1.42, compared to the control group. The effect is significant as the entire confidence interval is greater than zero. The increase is similar to the number of outlets participants immediately subscribed to in the intervention (1.51, not shown in the figure) since few participants unsubscribed from these outlets within two weeks.

[39] Both posts shared by friends and posts shared by pages could be affected by Facebook's algorithm.

[40] The subset of posts shared by friends that participants click on (row 9) is more segregated than the posts from friends that participants are exposed to in their feed (row 3). This complements the conclusion of Bakshy et al. (2015) who focus on posts shared by friends.

[41] Pew Research Center American Trends Panel Wave 51, July 2019.

[42] I use linear regressions for ease of interpretation. Since the dependent variables are count data, Appendix Table A.9 shows that the effects on exposure, browsing, and sharing posts are mostly robust to running Poisson regressions.

Exposure to Posts on Facebook The second panel of the figure shows that in the two weeks following the intervention, participants in the pro- and counter-attitudinal treatments were exposed to 64 and 31 additional posts from the potential pro- and counter-attitudinal outlets, respectively. For comparison, control group participants were exposed to 266 posts from leading news outlets, and 2,335 posts in total, in the two weeks following the intervention, suggesting that the intervention affected news exposure but did not take over the participants' feeds.

The effect on exposure is driven mostly by organic posts published by pages and not by sponsored posts or posts shared by friends, meaning that participants were exposed to the content directly, without commentary from their social network (see Appendix Figure A.2). To test whether participants noticed the posts, they were asked in the endline survey how often they saw news from various outlets in their Facebook feed in the past week. Appendix Figure A.3 shows that participants reported seeing more news from the outlets they were offered and that participants in the counter-attitudinal treatment were more likely to say that opinions they see in their feed are often not aligned with their views. This implies that the effect on the feed was noticeable for at least two months, and confirms that the treatment affected the subsample of participants who completed the endline survey and not only on participants who installed the extension.

News Sites Visited The third panel of Figure 4 shows that the counter-attitudinal treatment increased total visits to the websites of the counter-attitudinal outlets by 79%, an ITT effect of 1.34 visits over a baseline of 1.70 visits in the two weeks following the intervention. The pro-attitudinal treatment increased the number of visits to the websites of pro-attitudinal outlets by 21%, an ITT effect of 2.72 visits over a baseline of 13.23.

Appendix Figure A.4 separately estimates the effects of the intervention on the number of visits to the outlets' websites through a link appearing in the Facebook feed and on visits not directly associated with Facebook. While there is a strong and significant effect on visits through Facebook, there also seems to be an effect on other visits, albeit the latter result is not precisely estimated. It is possible that once participants read an article on the outlets' websites, they followed links to other articles as well. Alternatively, when participants became more familiar with the new outlets, they may have started visiting those outlets even without a Facebook referral. Appendix Figure A.5 shows that participants were more likely to click posts ranked higher in the feed. This could occur both because participants are more curious when they just start scrolling their feed and because Facebook's algorithm ranks posts according to expected interest. Interestingly, conditional on the order of posts, participants were as likely to visit a link from an outlet they subscribed to as a result of the intervention, compared to other news outlets.

Sharing Behavior The fourth panel of Figure 4 shows that participants not only consumed news from counter-attitudinal outlets when they appeared in their feeds, they also shared the posts. To increase power, in Appendix Figure A.6, I analyze this effect using the entire access posts subsample and show that both treatments had a significant effect on the number of posts shared by these participants. The fact that participants chose to share the posts suggests that they considered the posts important, and implies that participants expanded the treatments to their social network.

Complementing previous studies focusing on Twitter (Halberstam and Knight, 2016), participants were much more likely to share pro-attitudinal posts. However, the relative effect on sharing counter-attitudinal posts compared to the control group (an increase of 105%) is stronger than the relative effect of the pro-attitudinal treatment (53%). Participants may have shared posts while commenting negatively on their content. The second panel of Appendix Figure A.6 focuses on posts that were shared with no commentary by the participants and shows that even among these posts, the counter-attitudinal treatment had a significant effect on the number of posts shared.

4.2 The Social Media Feed Strongly Affects Online News Consumption

The previous section demonstrated that individuals engage with the potential outlets when they appear in their feed, suggesting that news is often consumed incidentally when it becomes more accessible. This raises the question of whether individuals adjust the rest of their news consumption such that the slant of their news diet will not change. For example, individuals randomly offered the New York Times may start consuming more articles from the outlet's website, but consequently consume less news from the Boston Globe, which offers a similar perspective. To test whether the treatment affected the mean slant of all news participants engaged with, I focus on the conservative and liberal treatments since there are clear predictions on how these treatments would affect the slant.

Exposure to Posts on Facebook The first panel in Figure 5 shows that when participants were randomly offered liberal or conservative outlets, their feed became substantially more liberal or conservative, respectively. The combined ITT effect of the liberal and conservative treatments equals 36% of the gap between the slant of the feed of liberals and conservatives in the control group. The corresponding TOT effect is 47%.[43] The change in slant provides a strong first stage, which is useful when analyzing the effect on political beliefs. It also allows me to test whether a change in the social media feed affects the slant of news sites visited or whether participants maintain a constant slant. The latter would suggest that participants re-optimize the sites they visit following an exogenous shock to their feed.

News Sites Visited I find that individuals do *not* re-optimize the slant of their news consumption. The second panel of Figure 5 shows that the treatments had a strong and significant effect on the slant of news sites visited by the participants. The combined effects of the liberal and conservative treatments equals 14%-19% (ITT-TOT) of the difference in the slant of news sites visited by

[43] These figures probably overestimate the effect on the feed because the slant is estimated based on a list of close to 500 leading news outlets and not all posts appearing in the feed.

conservatives and liberals in the control group. Based on the Comscore panel, the TOT effect of the liberal treatment would have shifted the online news diet of an individual in Pennsylvania, a swing state, to a diet similar to an individual in New York, a blue state, and the TOT effect of the conservative treatment would have led to a news diet similar to an individual in South Carolina, a red state.[44] Appendix Table A.10 shows that the effect is robust across various subsamples (e.g., when excluding participants who did not complete the endline survey).

By combining the exposure and browsing data, I find that when the compliers' news feed became one standard deviation more conservative, the slant of the news sites they visit became 0.31 standard deviations more conservative. The effect on the slant of the subset of news sites visited through Facebook is 0.72 standard deviations (both effects are significant at the 1% level). These estimates are calculated by instrumenting the slant of the posts observed in the Facebook feed with the treatment assignment. The regressions rely on the exclusion restriction that the treatments only affects the slant of sites visited through the slant of the Facebook feed. While the intervention is only expected to have an effect through the Facebook feed, the treatments could affect the feed in many ways. I am condensing the feed, a complicated object, to a scalar, the mean slant of news an individual was exposed to. This scalar is strongly affected by the treatment assignment and has intuitive economic meaning, but other changes in the feed, not captured in this measure, could affect the news sites visited. Since these calculations rely on stronger assumptions than the ITT and TOT estimates, they should be interpreted cautiously.

To test for spillovers across news outlets, I recalculate the effect of the treatments on the mean slant of all leading outlets, excluding the eight potential outlets defined for each individual. Appendix Figure A.7 shows that the mean slant of news consumption is not strongly affected by the treatments when the potential outlets are excluded, implying that the experiment did not have large crowd-in or crowd-out effects.

Persistence Is is possible that participants were initially curious about the new outlets they were offered but quickly stopped engaging with them. Figure 6 shows that the effect of the liberal treatment on news slant, compared to the conservative treatment, declines over the first six weeks after the intervention but mostly remains positive and significant. Appendix Figure A.8 repeats this analysis for the first twelve weeks after the intervention. While these results should be interpreted more cautiously since a substantial number of participants did not keep the extension installed or provide permissions to access posts over this longer time period, they suggest that the effects of the experiment declined but remained significant for at least twelve weeks.

The long-term effects also alleviate concerns that experimenter effects are driving the results in this section, as it is unlikely that participants remembered which posts appeared in their feed as

[44] For each individual in Comscore's 2017 and 2018 panels, the websites visited are matched with the leading news outlets to determine the individual's mean news consumption slant. Individuals who visited only one news site are excluded. The slant is then calculated at the state level for all panel members in the state. The example focuses on states where there is a larger sample of Comscore panelists.

a result of the intervention two months after the baseline survey, assumed that the experimenter expected them to persistently visit these websites, were constantly conscious that some of their browsing behavior could be observed, and were willing to spend time visiting news sites only to leave an impression on the experimenter. Furthermore, a survey question in the endline survey suggests that most participants did not remember which outlets they subscribed to and therefore their behavior or answers are unlikely to have been driven by experimenter effects.[45]

4.3 Discussion

This section shows that people are willing to substantially change their news consumption and engage with counter-attitudinal news on social media, as a result of a subtle nudge. In Appendix C.2, I analyze the content of posts participants engaged with based on the words appearing in the posts and the sections of the articles the posts linked to (e.g., Politics, Business, or Arts). I find that a large share of content tends to be political, even when the outlets the participants engaged with were counter-attitudinal.

How do these results coincide with the previous section, which shows that news consumed through social media, tends to be pro-attitudinal? If news is consumed incidentally on social media, and the Facebook feed tends to be pro-attitudinal, individuals are more likely to visit pro-attitudinal websites through social media but they will start visiting counter-attitudinal websites when they appear in their feed. Passive news consumption can also explain why Chen and Yang (2019) find that providing access to uncensored Internet does not lead to consumption of censored foreign news. As long as consumers are passive, providing access to new outlets may not be sufficient to affect news consumption since consumers will continue visiting their default outlets appearing in their bookmarks, search results, or social media feeds. My intervention may have affected news consumption because it increased the salience of specific outlets and decreased the search costs required to visit them by showing them on Facebook often.[46]

This conclusion raises concerns regarding the power of social media companies in shaping news consumption habits. The effect of the social media feed on news consumption implies that any change to the feed, stemming from new subscriptions or a change in the algorithm, can drastically change one's news diet. Attempts to change the feed by suggesting new content happen all the time. They can stem from companies attempting to maximize profits by increasing user

[45] Participants were asked "In a previous survey, we may have asked if you are interested in 'liking' Facebook news pages. Did you like a page in the previous survey?" Only 40% of participants in the treatment arms stated that they remembered whether they liked a page and which pages they liked. Unfortunately, many participants did not understand this question and assumed it refers to a previous question in the endline survey. Therefore, I interpret this question as providing qualitative evidence that many participants did not remember which outlets they subscribe to and not for empirical analysis. The misunderstanding probably leads to an overestimation of the number of participants who remember which pages they liked as some respondents may have remembered the previous question in the endline survey but not the outlets offered in the baseline survey. Furthermore, even among the minority of participants who understood the question and stated that they remember which pages they liked, some did not state the correct outlets.

[46] An additional difference between the studies is the setting. My intervention took place in an uncensored media environment where participants are often already familiar with the outlets offered.

engagement or originate from entities attempting to maximize political goals, such as political candidates purchasing ads or even foreign agents promoting Facebook pages in order to influence the American electorate.[47]

5 Findings: Opinions and Attitudes

5.1 Social Media News Exposure Does Not Strongly Affect Political Opinions

The top panel of Figure 7 shows that the treatments did not affect the political opinions index. While the point estimate has the expected sign, the effect is minimal (0.005 standard deviations), precisely estimated, and not statistically significant. The upper bound for the combined liberal and conservative treatment effects, based on a 95% confidence interval, is only 0.8% of the difference in political opinions between liberals and conservatives in the control group. Appendix Figure A.9 shows that the effect on each component of the political opinions index is small, and I cannot reject a null effect for any of the components.

Why did the treatments not affect political opinions even though they dramatically affected the Facebook feed of participants? Previous studies found a null effect that masked substantial heterogeneity (Baysan, 2019). Perhaps some participants were persuaded by the outlets they consumed, while for others, there was a backlash effect and opinions moved in the opposite direction of their treatment assignment. Appendix Figure A.10 estimates the effect of the interaction of ideology and treatment arm on the political opinions index and finds no evidence for a backlash effect. A second option is that social media provides a growing share of news, but is still not a dominant news source, compared to television. This could explain why the results of this study differ from studies on Fox News (DellaVigna and Kaplan, 2007; Martin and Yurukoglu, 2017). Interestingly, I do not find evidence for heterogeneity based on whether participants reported getting most of their news on social media (see Appendix C.3). It is also possible that the null effects are explained by the fact that the intervention lasted for two months. However, the intervention length was long enough to affect attitudes, as discussed in the next section.

The results differ from a recent study by Bail et al. (2018), who expose individuals to political content on Twitter and find evidence for a backlash effect. Differences in the experiments' design can explain the differing results. Bail et al. (2018) expose individuals to a bot retweeting counter-attitudinal *views*. Individuals plausibly become more upset when they are exposed to opposing opinion leaders, compared to counter-attitudinal news outlets. Bail et al. (2018) also provided monetary incentives to continuously follow the bots, asked participants to disable Twitter's timeline algorithm to ensure they viewed the tweets, and included weekly surveys to verify compliance. In my setting, participants were randomly offered outlets but could decide whether

[47] For example, many ads purchased by Russian organizations in their attempt to influence the 2016 election promoted Facebook pages. Congress has published the ads and they can be found here: https://intelligence.house.gov/social-media-content/social-media-advertisements.htm

to comply with the treatment and engage with the content. Therefore, compliers with each treatment arm are different by design and this could affect the results. Social scientists have criticized the generalizability of forced exposure media experiments since the effects found may be concentrated among individuals who would not consume the content outside the experimental setting (Bennett and Iyengar, 2008; Hovland, 1959). For example, conservatives who get upset when visiting msnbc.com are less likely to consume content from MSNBC in my setting but may consume similar content in the Bail et al. setting, and this type of consumption could drive the backlash effect.

5.2 Exposure to Counter-Attitudinal News Decreases Affective Polarization

The bottom panel of Figure 7 shows that the counter-attitudinal treatment modestly decreased the affective polarization index compared to the pro-attitudinal treatment. The ITT and TOT effects are 0.03 and 0.06 standard deviations, respectively. This suggests that the concerns over more segregated news consumption are not misguided. When estimating the effect on each component of the index separately in Appendix Figure A.11, the effect is largest for the difficulty in seeing things from each party's point of view measure.

Appendix Tables A.11, A.12, and A.13 show that the result is robust to excluding covariates, dropping each of the five components of the affective polarization measures from the index one at a time, and excluding participants who already subscribed to at least one of the primary outlets before the intervention. Appendix Table A.14 shows that an effect is detected when focusing on the subsample of participants who completed the endline survey and installed the extension. The effect is stronger among this group, which also had higher compliance rates. Appendix C.4 shows that the effect is similar when the regressions are reweighted to match populations means in ideology, party affiliation, gender, age, and the baseline feeling thermometer measure. Appendix C.5 estimates heterogeneous effects using causal forests and shows that the predicted effect in the entire baseline sample is very similar to the effect among the endline survey subsample.

Comparing each treatment separately to the control group shows that most of the difference between the pro- and counter-attitudinal treatments stems from the counter-attitudinal treatment, perhaps because the relative effect of this treatment on engagement with the outlets was larger compared to baseline. In all specifications, the effect of the counter-attitudinal treatment is negative, statistically significant, and stronger than the effect of the pro-attitudinal treatment. However, this comparison suffers from differential attrition, due to lower attrition in the control group. Therefore, in Appendix Table A.11, I also calculate Lee bounds for the effects of each treatment (Lee, 2009). Due to the relatively small treatment effect, the bounds include a null effect. As an additional robustness test, I exclude control group participants who were recruited using the last email or ad inviting them to the endline survey (Behaghel et al., 2015). Without these participants, I compare the 46% of participants in each treatment arm who were "easiest" to recruit and attrition is similar across treatments. The results using this method are almost identical to the main specification.

I do not find evidence for substantial heterogeneity across most covariates I test for, including age, ideological leaning, baseline interest in news, and baseline exposure to counter-attitudinal news (Appendix C.3). One exception is that the treatment seemed to have a stronger effect on participants who were less polarized in baseline according to the feeling thermometer question. However, this effect is significant only at the 10% level and more research is required on heterogeneity.

In the rest of this section, I interpret the magnitudes of the effect using three approaches. First, I compare the effect of the intervention to benchmarks in the control group and outside the experiment. Second, I use the browser data to estimate the effect of a change in exposure to pro- and counter-attitudinal news on affective polarization. Third, I conduct two back-of-the-envelope calculations to estimate how affective polarization would have changed if Facebook had a more balanced feed. All the results are based on the effect of the offered outlets over two months and could be different with longer exposure or if different outlets were offered.

The ITT and TOT effects of the counter-attitudinal treatment decrease the difference between the feeling toward the participant's party and the opposing party by 0.58 and 0.96 degrees (on a 0-100 scale), respectively. For comparison, in the past 20 years, the feeling thermometer measure increased by 3.83-10.52 degrees. An additional point of comparison is a recent experiment by Allcott et al. (2020) who found that disconnecting from Facebook decreases the feeling thermometer measure by 2.09 degrees. Hence, one way to interpret these results is that almost half of the depolarizing effect of disconnecting from Facebook can be achieved by replacing 1-4 subscriptions to pro-attitudinal outlets with subscriptions to counter-attitudinal outlets.[48]

To estimate the effect of exposure to pro- or counter-attitudinal news on polarization, I focus on participants who installed the browser extension and completed the endline survey (i.e., the overlap between the extension and the endline subsamples). I use two summary measures for exposure to pro- and counter-attitudinal news: the share of counter-attitudinal news in the Facebook feed and the feed's congruence scale. I calculate these statistics based on all posts observed between the baseline and endline survey, for participants who observed at least two pro- or counter-attitudinal posts. I estimate the effect of each measure on affective polarization, and instrument the measure with the treatment assignment. Similar to the discussion in Section 4.2, the IV regressions rely on the exclusion restriction that the treatment only has an effect on affective polarization through its effect on the measure analyzed.

I find that an increase of one standard deviation in the share of exposure to counter-attitudinal news decreases affective polarization by 0.13 standard deviations and an increase of one standard

[48] This interpretation ignores the small differences between the settings of the studies and the samples. I estimate an effect over two months in the spring of 2018, while Allcott et al. (2020) conduct the study over one month in the fall of 2018. Furthermore, while both samples were recruited using Facebook ads, the sample compositions could still differ, for example, since Allcott et al. (2020) screen respondents who report using Facebook for less than 15 minutes per day or who are not willing to deactivate Facebook for 24 hours.

deviation in the congruence scale has a similar effect.[49] One challenge in studying affective polarization based on non-experimental survey data (e.g., Garrett et al., 2014) is determining whether the correlation between news exposure and affective polarization is due to selection, i.e., individuals with more negative views of the opposing party select into more pro-attitudinal news exposure, or a causal effect, i.e., pro-attitudinal news makes people more polarized. Appendix Table A.15 shows that the effects of news exposure on affective polarization are approximately 26%-34% of the coefficients obtained using a cross-sectional regression among the control group, suggesting that the correlation is both due to a causal effect and selection.

I use the effect of the Facebook feed to estimate how affective polarization would have changed if individuals were exposed to more balanced news on Facebook. I find that if the feed had an equal share of pro- and counter-attitudinal news, the difference between the feelings toward one's party and the opposing party would decrease by 3.94 degrees. For this calculation, I estimate the effect of increasing the share of exposure to counter-attitudinal news by 33 percentage points, the difference between exposure in the control group and an exposure of 50%. The estimation does not rely on out-of-sample predictions as the share of counter-attitudinal news was greater than 50% for many participants in the counter-attitudinal treatment. Using a similar exercise, I find that if the congruence of the Facebook feed equaled zero, polarization would decrease by 3.40 degrees.

Perhaps a balanced news feed is not a realistic counterfactual because most individuals do not consume balanced news, regardless of social media. Therefore, in a second back-of-the-envelope calculation I estimate how affective polarization would change if individuals were exposed in their Facebook feed to the same share of counter-attitudinal outlets, or the same congruence scale, as they encounter when visiting news sites not through Facebook. I find that the feeling thermometer outcome would decrease by 0.25-0.62 degrees. These calculations should be interpreted carefully since they do not take into account general equilibrium effects.[50] Nevertheless, they suggest that the Facebook feed may slightly amplify polarization.

5.2.1 Interpretation

Why did the treatments affect attitudes toward political parties but not political opinions? One possibility is that participants learned new facts about the world and these facts swayed their attitudes. Based on eight pre-registered survey questions, I test whether a change in participants' knowledge could explain the effect on polarization. In Appendix C.6, I do not find evidence for strong effects on knowledge.

[49] The effects are significant at the 10% level as the sample size is smaller when focusing on participants who both installed the extension and completed the endline survey.

[50] For example, it is likely that if Facebook drastically changed its feed, individuals would use other social networks instead. Some of this effect may be captured in the calculations since participants in the counter-attitudinal treatment used Facebook less often (as discussed in Section 6). However, with network effects, the decrease in Facebook use could be greater. The calculations also ignore the indirect effect of Facebook on news sites visited.

There is evidence that Americans believe that members of the opposing are more likely to hold extreme views than they actually do (Yudkin et al., 2019), and therefore, attitudes may have changed because participants learned the opposing party is not as extreme as they thought.[51] I do not find evidence that the pro- and counter-attitudinal treatments had a significant effect on the distance between participants' baseline ideology and the perceived ideology of each party (Appendix Figure A.12).

Another option is that exposure to pro- and counter-attitudinal news affects attitudes due to increased negative coverage (Levendusky, 2013). This explanation predicts that pro-attitudinal outlets would increase negative attitudes toward the opposing party and counter-attitudinal outlets would affect consumers' attitudes toward their own party. This prediction is inconsistent with the data. I measure separately the effect of each treatment on attitudes toward each party and show in Appendix Table A.16 that exposure to counter-attitudinal outlets affecting attitudes toward the opposing party is driving the results.

An alternative explanation, consistent with the data, is that participants exposed to counter-attitudinal news learned to rationalize the opinions of the opposing party. Intuitively, participants may have learned some of the opposing party's arguments and thus understood better why the other party supports certain positions. This led to more positive attitudes but did not change political opinions as long as participants did not find these arguments particularly important. In Appendix D, I formalize this discussion using a model where political opinions are a weighted average of multiple beliefs and parties place different weights on beliefs.

There could be other explanations for the change in affective polarization.[52] The literature on affective polarization is new and more research is needed to pinpoint the precise mechanisms explaining how affective polarization evolves.

6 Findings: Exposure to Pro-Attitudinal News on Social Media

The previous section shows that exposure to pro-attitudinal news affects partisan hostility, therefore it is important to understand what influences the news individuals are exposed to on social media. This section decomposes the gap in exposure to posts shared by the pro- and counter-attitudinal outlets offered in the experiment into three main forces: participants are less likely to subscribe to counter-attitudinal news outlets; Facebook's algorithm supplies fewer posts from counter-attitudinal outlets, conditional on participants subscribing to them; and participants use

[51] This theory is consistent with a study by Orr and Huber (2018) who find that negative attitudes toward individuals from the opposing party decrease when information is provided about their policy position.

[52] The counter-attitudinal treatment may have mitigated tribalism, which could have decreased affective polarization (Mason, 2015). Indeed, field experiments have found that strengthening partisan behavior can affect political behavior and beliefs (Gerber et al., 2010). I use party affiliation as a proxy for tribalism and find in Appendix Figure A.12 that the treatments did not significantly affect this proxy. However, the point estimate of the effect on Democratic party affiliation has the predicted sign, and I cannot reject a small effect on affiliation with the Democratic party.

Facebook less often when offered counter-attitudinal outlets. The decomposition exercise is based on the following framework:

$$E_{ij} = S_{ij}A_{ij}U_i$$

where E_{ij}, exposure, is the number of posts individual i was exposed to from outlet j. Exposure is a product of whether individual i subscribed to outlet j (S_{ij}), the share of posts shared by the outlet among all posts the individual observed (A_{ij}), and the total number of posts individual i observed on Facebook (U_i). I decompose the gap in exposure using the following formula:

$$\Delta E = \underbrace{S_\Delta A_C U_C}_{\text{Subscriptions}} + \underbrace{S_C A_\Delta U_C}_{\text{Algorithm}} + \underbrace{S_C A_C U_\Delta}_{\text{Usage}} + \underbrace{S_\Delta A_\Delta U_C + S_\Delta A_C U_\Delta + S_C A_\Delta U_\Delta + S_\Delta A_\Delta U_\Delta}_{\text{Combinations}} \tag{3}$$

where for each variable, $_C$ denotes the value for the counter-attitudinal treatment and $_\Delta$ denotes the difference between the pro- and counter-attitudinal treatments. *Subscriptions* is the additional counter-attitudinal posts participants assigned to the counter-attitudinal treatment would have been exposed to if they would have subscribed to the same number of outlets as participants assigned to the pro-attitudinal treatment. *Algorithm* is the additional posts these participants would have been exposed to if Facebook's algorithm would have supplied them with the same share of posts from counter-attitudinal outlets, as the share supplied when subscribing to pro-attitudinal outlets. *Usage* is the additional posts these participants would have been exposed to if they would have used Facebook as much as participants assigned to the pro-attitudinal treatment.

S_C and U_C are the mean number of new subscriptions and the total number of posts participants were exposed to, respectively, in the counter-attitudinal treatment. I estimate S_Δ and U_Δ by regressing the number of subscriptions and total exposure on whether participants were assigned to the pro- or counter-attitudinal treatment. To estimate A_Δ and A_C, I pool the two groups of potential outlets for each participant such that each observation is a participant and either the group of pro-attitudinal outlets or the group of counter-attitudinal outlets. I then regress the share of posts observed by a participant that was shared directly by each group of outlets (among all posts the participant was exposed to) on the full interaction of the number of new outlets the participant subscribed to and whether the group of outlets is pro-attitudinal. Since subscriptions are endogenous, they are instrumented with whether the group of outlets was randomly offered to the participant. The calculations are discussed in detail in Appendix C.7 along with alternative estimations.

Figure 8 shows that the strongest force associated with increased exposure to pro-attitudinal news is the algorithm. This demonstrates that even when individuals are willing to subscribe to outlets with a different point of view, Facebook's algorithm is less likely to show them content from those outlets (a phenomenon often described as a filter bubble). I also find evidence that participants prefer to subscribe to pro-attitudinal news outlets and that participants decrease their Facebook usage after they are offered to subscribe to counter-attitudinal outlets. The last effect is only significant at the 10% level and should be interpreted more cautiously. Still, it could explain why personalization is leading to segregation online. When consumers are exposed to more counter-attitudinal news, they may decrease their Facebook usage, and therefore, platforms have an incentive to filter counter-attitudinal news to maximize engagement.[53]

This section does *not* suggest that Facebook's algorithm intentionally increases segregation by ranking posts according to whether they match the user's beliefs, or that the interaction of the slant of an outlet and ideology of a user has a causal effect on a post's ranking. Platforms rank posts based on many signals that can be correlated with whether an outlet is counter-attitudinal, including the consumer's past engagement with the outlet, her social network, and possibly other pages she subscribes to. In other words, the effect of the algorithm also captures the behavior and perceived interests of the user. Indeed, Appendix C.7.2 shows that the effect of the algorithm slightly increases over time, suggesting that engagement with content plays a role in the ranking of posts.

Personalization of news exposure is still an important departure from how news was supplied in the past. Until recently, the engagement of an individual with news, e.g., the articles she reads in the newspaper or the cable channels she chose to watch, did not affect her supply of news. Interestingly, even though I find that the algorithm seems to be filtering counter-attitudinal posts, Section 3 shows that the posts control group participants are exposed to in their feed are not more pro-attitudinal than the outlets they subscribe to on Facebook. One possible explanation for the differing results is that participants in the control group are not randomly offered pro- and counter-attitudinal outlets. They probably subscribe to outlets as a response to non-random nudges. If Facebook users typically receive nudges to subscribe to pro-attitudinal outlets, then users will often subscribe to these outlets and only users who are specifically interested in opposing content will subscribe to counter-attitudinal outlets. As a result, the algorithm may filter less counter-attitudinal content.[54] The comparison to the control group descriptive statistics not only demonstrates why an experiment is necessary but also has policy implications. Adjusting the algorithm to offer more balanced news outlets, conditional on subscription, would not make a big difference if individuals only subscribe to pro-attitudinal outlets. Therefore, to increase diversity in news exposure, nudges encouraging participants to subscribe to diverse outlets may also be required.

While I focus on Facebook, this section's conclusions likely apply to other platforms personalizing content as well. For example, since 2016, Twitter has been ranking tweets according to how

[53] This result raises the question of whether the algorithm also personalizes content within an outlet, by showing conservatives relatively conservative posts shared by an outlet and liberals relatively liberal posts shared by the same outlet. In Appendix C.7.3, I find no evidence for within-outlet personalization.

[54] While I cannot observe offers to subscribe to outlets, the control group participants subscribing to pro- and counter-attitudinal outlets are substantially different from each other. Among the 20 most popular liberal and conservatives pages, there is a difference of 0.32 standard deviations in the absolute value of ideology of participants subscribing to at least one pro- and counter-attitudinal outlet (the difference between compliers in the pro and counter-attitudinal treatments is 0.05 standard deviations). Moreover, subscriptions to counter-attitudinal outlets occur several months later than subscriptions to pro-attitudinal outlets, and posts from more recent subscriptions are probably more likely to appear in the feed. The experiment assures that all subscriptions occur at the same time, due to a random offer.

interesting and engaging they would be for a specific user, and as a result, may increase exposure to pro-attitudinal news.[55] Furthermore, major news outlets have also started to personalize their websites and the articles they suggest to their customers.[56]

7 Conclusions

Consumption of news through social media is increasing, but the effect of social media on public opinion remains controversial. I show that news consumption on social media is an important phenomenon because consumers are exposed to different news on social media, individuals incidentally consume news when it becomes accessible in their social media feed and news consumption on social media affects attitudes.

This paper suggests that a more nuanced view is needed regarding the effect of media on public opinion. On the one hand, I show that exposure to pro-attitudinal news increases affective polarization compared to counter-attitudinal news. This result provides a mechanism complementing other important studies finding that social media can increase polarization and raises concern since affective polarization may decrease trust in governance and the accountability of elected officials. On the other hand, it seems that individuals are not easily persuaded by the political leaning of their news exposure. The results of the experiment are in line with the long term increase in affective polarization, without an equivalent change in political opinions (Mason, 2015). This suggests that a more segregated news environment may partially explain the increase in affective polarization over the past several decades.[57]

Methodologically, this paper has several limitations. First, I only observe online news consumption. While I show that the intervention did not seem to have substantial spillovers across online outlets, to precisely measure total news consumption, future studies would need to collect consumption data from other mediums, such as television, as well. Furthermore, I collect data on browsing behavior and the Facebook feed on a computer, but a growing share of news is consumed through smartphones. Second, while I argue that due to the organic nature of the intervention, it is unlikely that experimenter effects play a major role in this study, I cannot rule out that the perceived expectations of the experimenter affected the results. Third, the endline survey suffers from high attrition. I use several methods to alleviate this concern, but attrition could still affect the survey outcomes. Fourth, moderate outlets represent a large share of online news consumption, but they are visited less often through social media. This study does not generate random variation in exposure to moderate outlets and therefore, cannot speak to their effects.

[55] Factors taken into account when ranking tweets include the tweet's author and the user's past relationship with the author. Therefore, it is plausible that tweets from pro-attitudinal accounts will receive a higher ranking.

[56] In 2017, the New York Times announced that it will tailor its homepage to the interests of individual readers. The New York Times. A 'Community' of One: The Times Gets Tailored. March 18, 2017.

[57] For example, cable news is more segregated than broadcast news, and the Internet is more segregated than local newspapers (Gentzkow and Shapiro, 2011).

Fifth, while the experiment has high external validity when it comes to analyzing partisan outlets on Facebook in 2018, the result may not hold for other periods. For example, Trump's presidency is exceptional in the stability of the president's approval ratings. If other opinions were relatively stable throughout the period as well, the null effect on political opinions could be explained by the period when the survey took place.

This study has important policy implications. I demonstrate that Facebook's algorithm limits exposure to counter-attitudinal news. Automated personalization of news content may have stronger impacts in the future, due to growth in online news consumption and advances in machine learning algorithms customizing news exposure. However, I also find that individuals are willing to engage with counter attitudinal news. Therefore, even though social media platforms are associated with pro-attitudinal content, they can expose individuals to more perspectives. Suggestions include making algorithms more transparent,[58] nudging users to diversify their feed, and modifying algorithms to encourage serendipitous encounters (Pariser, 2011; Sunstein, 2017). The experiment described in this paper essentially measures the effect of one such intervention and shows that a simple nudge can be effective.[59]

While social media algorithms may increase affective polarization through their effect on news consumption, platforms also have the potential to mitigate these effects.

[58] In a 2018 survey in 18 European and English speaking countries, only 29% of respondents knew that algorithms predicting user interest determine which stories appear on Facebook (Reuters Institute, 2018).

[59] Social media platforms have recently started rolling out features that could potentially diversify users' feeds. In 2017, Facebook implemented a feature that shows users articles related to a post in their feed from additional outlets. In 2018, Twitter announced that it will allow users to follow topics in addition to specific accounts.

References

Allcott, H., L. Braghieri, S. Eichmeyer, and M. Gentzkow (2020). “The Welfare Effects of Social Media”. *American Economic Review* 110(3), 629–676.

Bail, C., L. Argyle, T. Brown, J. Bumpus, H. Chen, M. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky (2018). “Exposure to Opposing Views can Increase Political Polarization: Evidence from a Large-Scale Field Experiment on Social Media”. *Proceedings of the National Academy of Sciences of the United States of America* 115(37), 9216–9221.

Bakshy, E., S. Messing, and L. A. Adamic (2015). “Exposure to Ideologically Diverse News and Opinion on Facebook”. *Science* 348(6239), 1130–1132.

Baysan, C. (2019). “The Polarizing Effects of Persuasive Communication: Experimental Evidence from Turkey”.

Behaghel, L., B. Crépon, M. Gurgand, and T. L. Barbanchon (2015). “Please Call Again: Correcting Nonresponse Bias in Treatment Effect Models”. *Review of Economics and Statistics* 97(5), 1070–1080.

Bennett, W. L. and S. Iyengar (2008). “A New Era of Minimal Effects? The Changing Foundations of Political Communication”. *Journal of Communication* 58(4), 707–731.

Berinsky, A. J., G. A. Huber, and G. S. Lenz (2012). “Evaluating Online Labor Markets for Experimental Research: Amazon.com’s Mechanical Turk”. *Political Analysis* 20(3), 351–368.

Boxell, L., M. Gentzkow, and J. M. Shapiro (2018). “Greater Internet Use is Not Associated with Faster Growth in Political Polarization among US Demographic Groups”. *Proceedings of the National Academy of Sciences of the United States of America* 115(3), 10612–10617.

Bursztyn, L., G. Egorov, R. Enikolopov, and M. Petrova (2019). “Social Media and Xenophobia: Evidence from Russia”. *Working Paper*.

Chan, J. and W. Suen (2008). “A Spatial Theory of News Consumption and Electoral Competition”. *Review of Economic Studies* 75(3), 699–728.

Chen, Y. and D. Y. Yang (2019). “The Impact of Media Censorship: 1984 Or Brave New World?” *American Economic Review* 109(6), 2294–2332.

Chiang, C. F. and B. Knight (2011). “Media Bias and Influence: Evidence from Newspaper Endorsements”. *Review of Economic Studies* 78(3), 795–820.

Coppock, A., E. Ekins, and D. Kirby (2018). “The Long-lasting Effects of Newspaper Op-Eds on Public Opinion”. *Quarterly Journal of Political Science* 13(1), 59–87.

DellaVigna, S. and E. Kaplan (2007). “The Fox News Effect: Media Bias and Voting”. *The Quarterly Journal of Economics* 122(3), 1187–1234.

Durante, R., P. Pinotti, and A. Tesei (2019). “The Political Legacy of Entertainment TV”. *American Economic Review* 109(7), 2497–2530.

Enikolopov, R., A. Makarin, and M. Petrova (2019). “Social Media and Protest Participation: Evidence from Russia”.

Flaxman, S. R., G. Sharad, and J. M. Rao (2016). “Filter Bubbles, Echo Chambers, and Online News Consumption”. *Public Opinion Quarterly* 80, 298–320.

Garrett, R. K., S. D. Gvirsman, B. K. Johnson, Y. Tsfati, R. Neo, and A. Dal (2014). “Implications of Pro- and Counterattitudinal Information Exposure for Affective Polarization”. *Human Communication Research* 40(3), 309–332.

Gentzkow, M. and J. M. Shapiro (2006). “Media Bias and Reputation”. *Journal of Political Economy* 114(2), 280–316.

– (2010). “What Drives Media Slant? Evidence From U.S. Daily Newspapers”. *Econometrica* 78(1), 35–71.

– (2011). “Ideological Segregation Online and Offline”. *Quarterly Journal of Economics* 126(4), 1799–1839.

Gentzkow, M., J. M. Shapiro, and M. Sinkinson (2011). “The Effect of Newspaper Entry and Exit on Electoral Politics”. *American Economic Review* 101, 2980–3018.

Gentzkow, M., J. M. Shapiro, and D. F. Stone (2015). “Media Bias in the Marketplace: Theory”. *Handbook of Media Economics, 1B*. Vol. 1. Elsevier B.V., 623–645.

Gerber, A. S., G. A. Huber, and E. Washington (2010). “Party Affiliation, Partisanship, and Political Beliefs: A Field Experiment”. *American Political Science Review* 104(4), 720–744.

Gerber, A. S., D. Karlan, and D. Bergan (2009). “Does the Media Matter? A Field Experiment Measuring the Effect of Newspapers on Voting Behavior and Political Opinions”. *American Economic Journal: Applied Economics* 1(2), 35–52.

GlobalWebIndex (2018). *Flagship Report 2018*.

Guess, A. (2018). “(Almost) Everything in Moderation: New Evidence on Americans’ Online Media Diets”.

Guess, A., B. Nyhan, B. Lyons, and J. Reifler (2018). *Avoiding the Echo Chamber about Echo Chambers*. Knight Foundation.

Guess, A., B. Nyhan, and J. Reifler (2017). *"You're Fake News" Findings from the Poynter Media Trust Survey*. The Poynter Ethics Summit.

Halberstam, Y. and B. Knight (2016). “Homophily, Group Size, and the Diffusion of Political Information in Social Networks: Evidence from Twitter”. *Journal of Public Economics* 143, 73–88.

Hovland, C. I. (1959). “Reconciling Conflicting Results Derived from Experimental and Survey Studies of Attitude Change”. *American Psychologist* 14(1), 8–17.

Iyengar, S. and M. Krupenkin (2018). “The Strengthening of Partisan Affect”. *Political Psychology* 39, 201–218.

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.

Jo, D. (2018). “Better the Devil You Know: An Online Field Experiment on News Consumption”.

Kennedy, P. J. and A. Prat (2019). “Where Do People Get Their News”. *Economics Policy Journal* 5-27.

Lee, D. S. (2009). “Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects”. *Review of Economic Studies* 76(3), 1071–1102.

Lelkes, Y., G. Sood, and S. Iyengar (2015). "The Hostile Audience: The Effect of Access to Broadband Internet on Partisan Affect". *American Journal of Political Science* 61(1), 5–20.

Levendusky, M. (2013). "Partisan Media Exposure and Attitudes Toward the Opposition". *Political Communication* 30(4), 565–581.

Martin, G. J. and A. Yurukoglu (2017). "Bias in Cable News: Persuasion and Polarization". *American Economic Review* 107(9), 2565–2599.

Mason, L. (2015). ""I Disrespectfully Agree": The Differential Effects of Partisan Sorting on Social and Issue Polarization". *American Journal of Political Science* 59(1), 128–145.

Mosquera, R., M. Odunowo, and T. Mcnamara (2019). "The Economic Effects of Facebook". *Experimental Economics*, 1–28.

Mullainathan, S. and A. Shleifer (2005). "The Market for News". *American Economic Review* 95(4), 1031–1053.

Müller, K. and C. Schwarz (2019). "From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment".

Okuyama, Y. (2019). "Toward Better Informed Decision-Making: the Impacts of a Mass Media Campaign on Women's Outcomes in Occupied Japan".

Orr, L. V. and G. A. Huber (2018). "The Policy Basis of Measured Partisan Animosity in the United States".

Pariser, E. (2011). *The Filter Bubble*. The Penguin Press.

Parse.ly (2018). "The Authority Report: 2018 Traffic Sources by Content Categories and Topics".

Peterson, E., G. Shared, and S. Iyengar (2019). "Partisan Selective Exposure in Online News Consumption: Evidence from the 2016 Presidential Campaign". *Political Science Research and Methods*, 1–17.

Pew (2014). *Political Polarization and Media Habits*. Pew Research Center.

Rand, D. G., A. Peysakhovich, G. T. Kraft-Todd, G. E. Newman, O. Wurzbacher, M. A. Nowak, and J. D. Greene (2014). "Social Heuristics Shape Intuitive Cooperation". *Nature Communications* 5, 1–12.

Reit, E., R. Willer, and J. Zaki (2017). "Causes and Consequences of Political Empathy". *Work in Progress, Standford University*.

Reuters Institute (2018). *Digital News Report 2018*. University of Oxford.

– (2019). *Digital News Report 2019*. University of Oxford.

Strömberg, D. (2015). "Media and Politics". *Annual Review of Economics* 7(1), 173–205.

Sunstein, C. (2017). *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

Williamson, D. A. (2018). *US Time Spent with Social Media 2019*. eMarketer.

Yudkin, D., S. Hawkins, and T. Dixon (2019). *The Perception Gap: How False Impressions are Pulling Americans Apart*. More In Common.

Zhuravskaya, E., M. Petrova, and R. Enikolopov (2020). "Political Effects of the Internet and Social Media". *Annual Review of Economics*.

Figure 1: Experimental Design

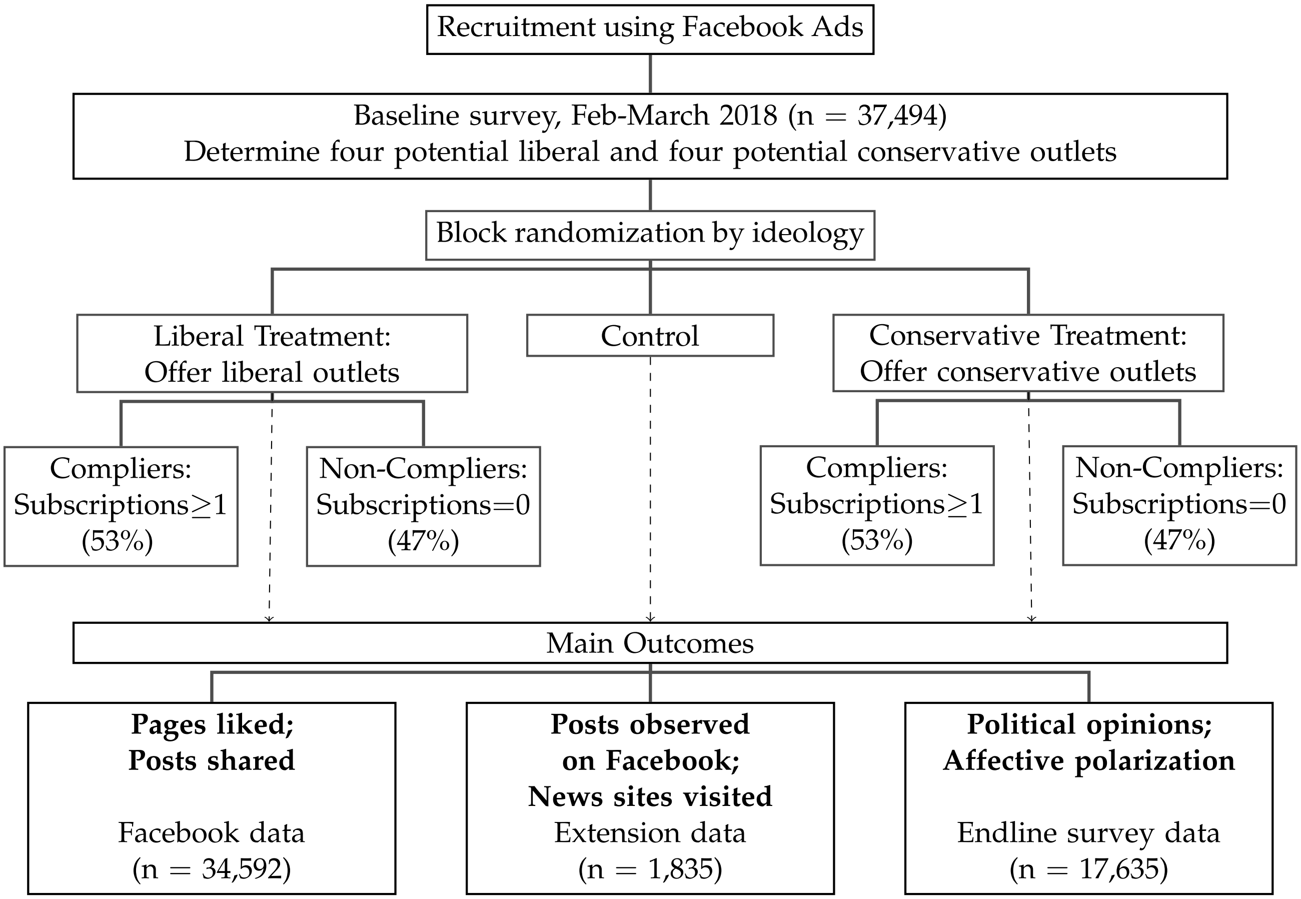
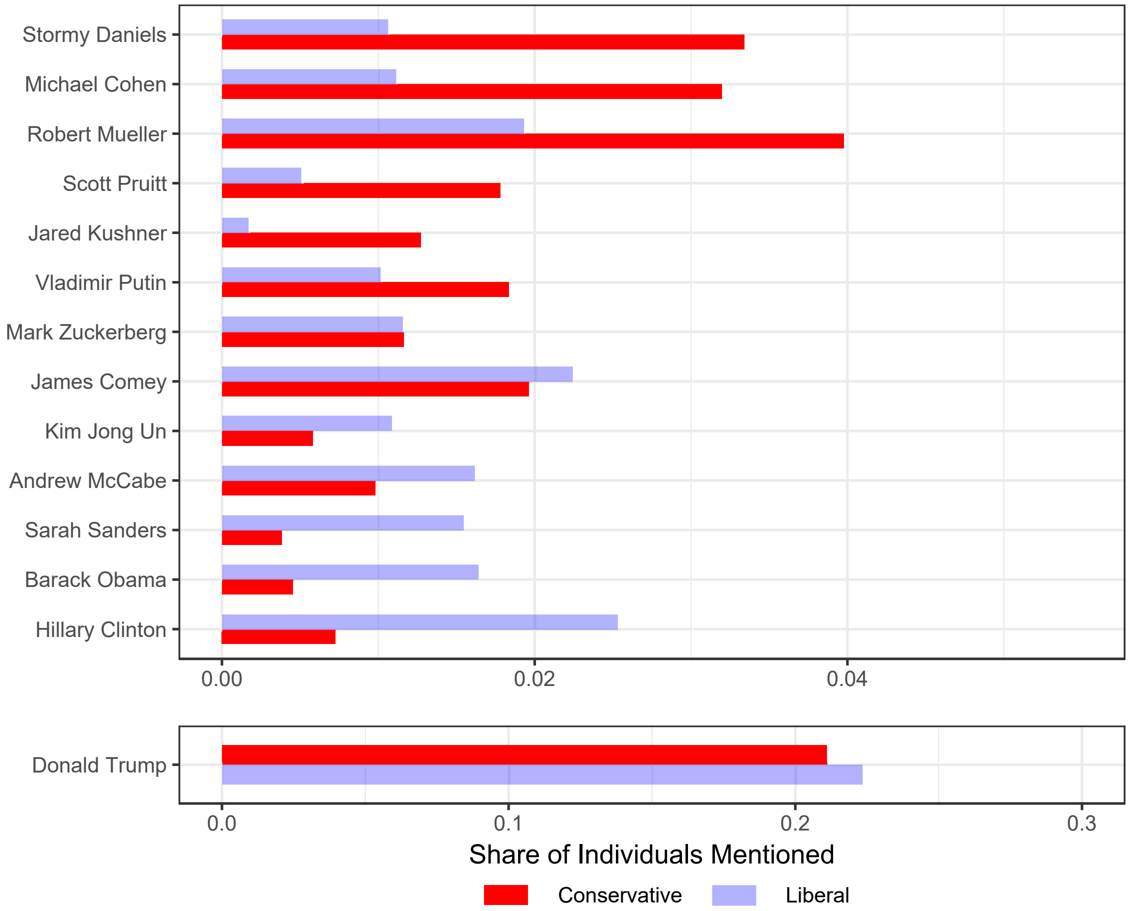
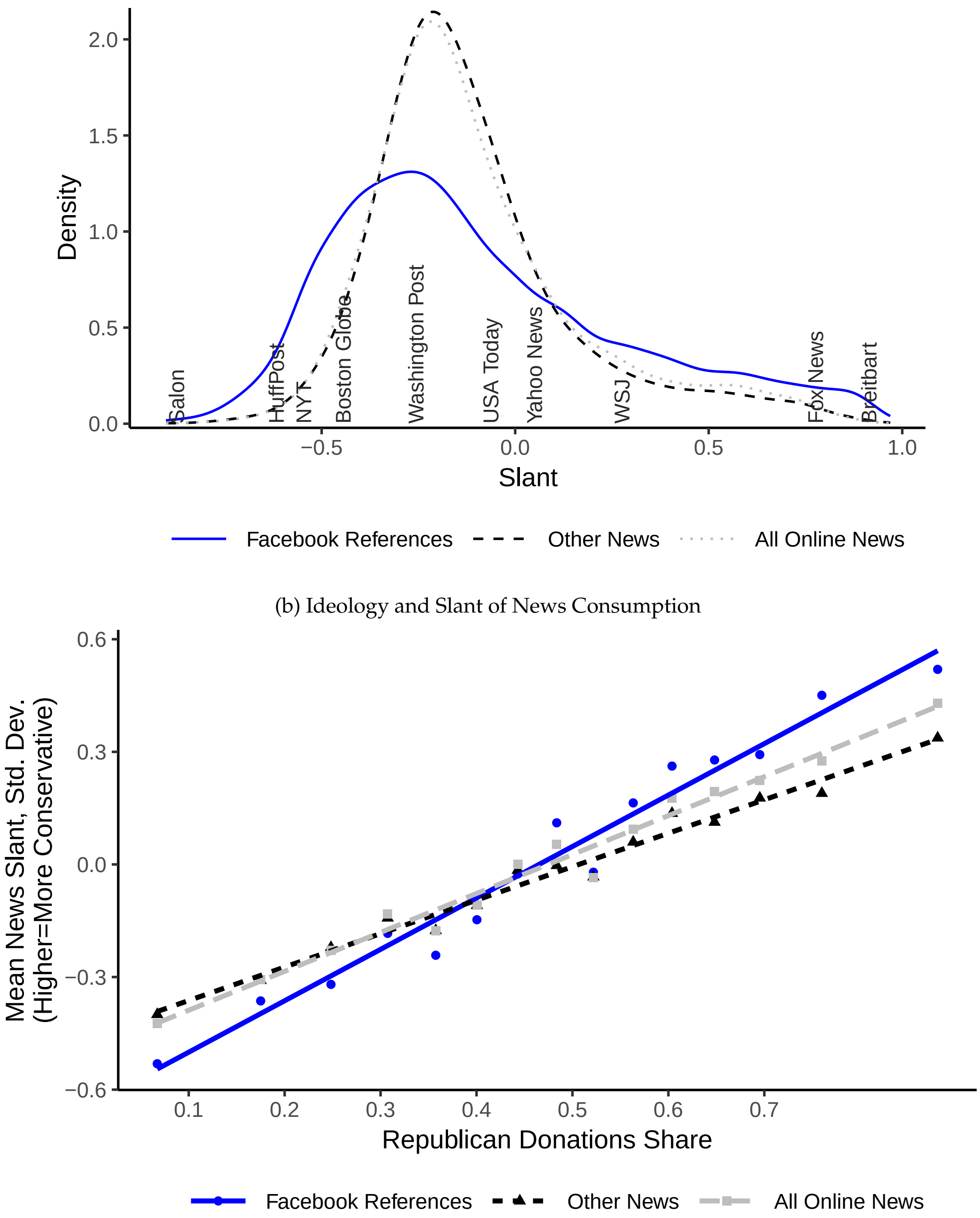

Recruitment using Facebook Ads
Baseline survey, Feb-March 2018 (n = 37,494)
Determine four potential liberal and four potential conservative outlets
Block randomization by ideology
Liberal Treatment: Offer liberal outlets
Control
Conservative Treatment: Offer conservative outlets
Compliers: Subscriptions≥1 (53%)
Non-Compliers: Subscriptions=0 (47%)
Compliers: Subscriptions≥1 (53%)
Non-Compliers: Subscriptions=0 (47%)
Main Outcomes
Pages liked; Posts shared
Facebook data (n = 34,592)
Posts observed on Facebook; News sites visited
Extension data (n = 1,835)
Political opinions; Affective polarization
Endline survey data (n = 17,635)

Figure 2: Figures Discussed in the News During the Study Period, All Posts Published by the Primary Outlets



This figure shows the prominent men and women mentioned in posts shared by the primary outlets between February 28 and April 25, 2018, the median dates the baseline survey and endline survey were taken. Approximately 33% of posts with text mentioned a name. The x-axis is the share of times an individual was mentioned in a post by one of the four primary conservative outlets (top bars) and by one of the four primary liberal outlets (bottom bars), of all mentions of individuals. To fit all the figures on the same scale, the x-axis is broken for Donald Trump, who is by far the most dominant person mentioned. The figures were identified using the Spacy Natural Language Processing algorithm and post-processing names (e.g., removing possessive 's). Names that appear in only one outlet are excluded. If only a last name is mentioned, it is associated with the dominant first and last name combination when such a combination exists. To simplify the graph, the names 'Trump' and 'Donald Trump' are determined to be the same individual, even though 'Trump' could refer to other members of President Trump's family.

Figure 3: News Consumption in the Comscore Panel

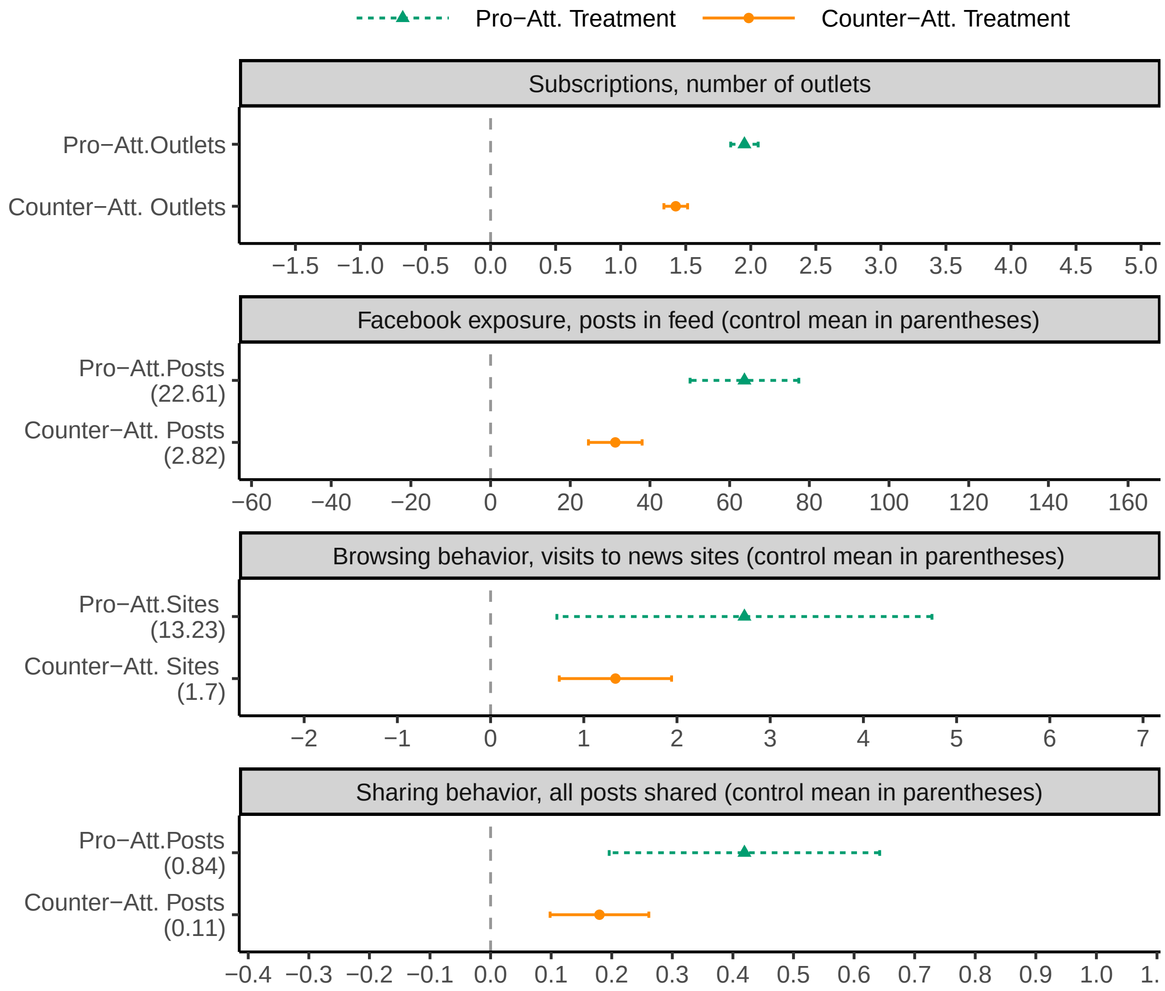


(a) Distribution of Mean News Slant

(b) Ideology and Slant of News Consumption

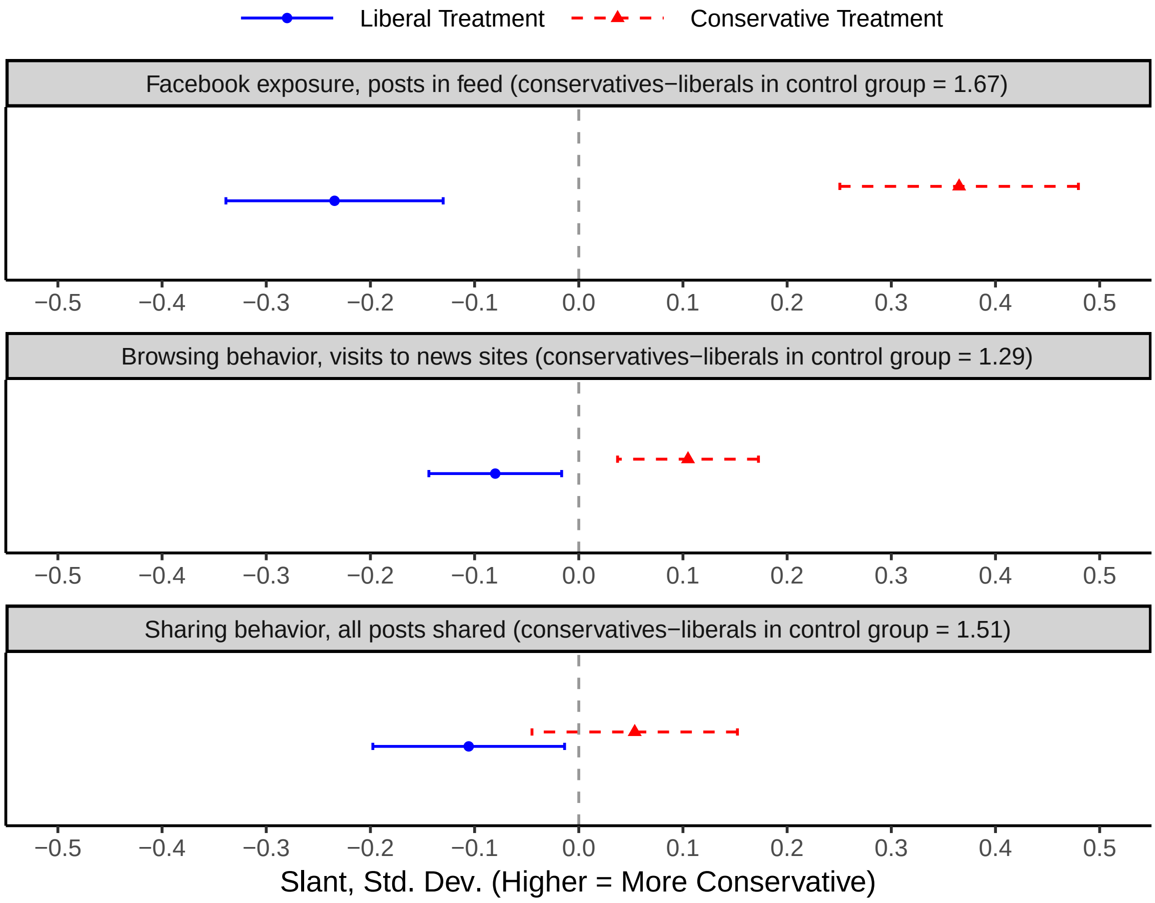
This first figure shows the distribution of the mean slant of news sites visited by individuals in the 2017 and 2018 Comscore Web Behavior Database Panel (smoothing bandwidth = 0.05). Major news outlets are added to the x-axis for reference. The slant of each domain is based on Bakshy et al. (2015). A visit is referred from Facebook if the referring domain is "facebook.com." This second figure presents a binned scatter plot with the share of donations in a zip code based on the 2016 and 2018 election cycles FEC donation data on the x-axis and mean slant on the y-axis. The sample includes all individuals who visited news sites multiple times through Facebook and through other means.

Figure 4: Effects of the Pro- and Counter-Attitudinal Treatments on Subscriptions, News Exposure, News Sites Visited and Sharing Behavior, Two Weeks Following the Intervention



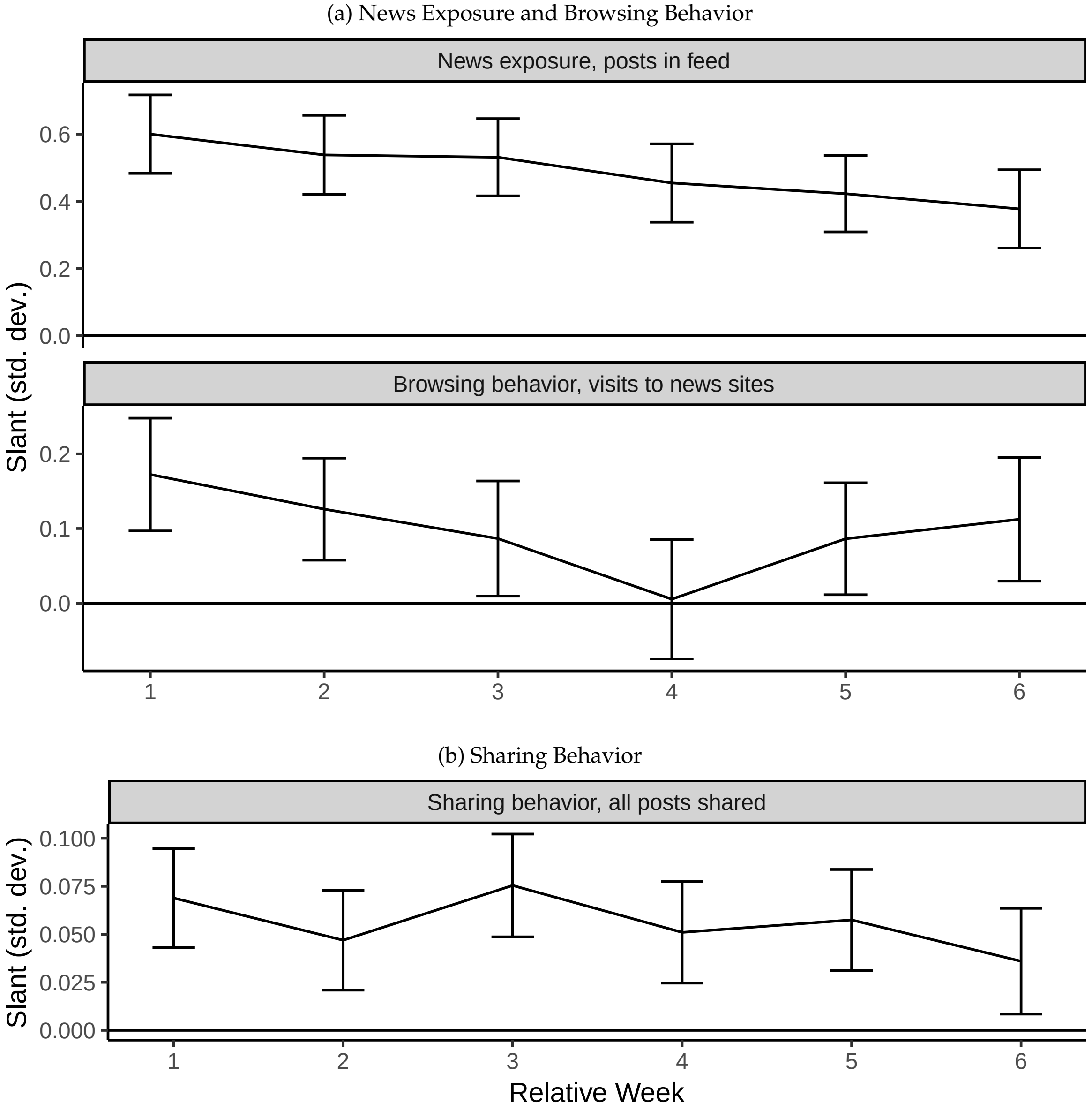
This figure shows the effect of the treatments on engagement with the participants' potential outlets in the two weeks following the intervention. The dependent variable is engagement with either the four potential pro-attitudinal outlets or the four potential counter-attitudinal outlets and the independent variable is the treatment. The outcomes are the number of outlets participants subscribed to, posts from the outlets that appeared in their Facebook feed, visits to the outlets' websites, and posts shared from the outlets by the participants. For example, in the third panel, the triangle and dashed line represent the point estimate and the confidence interval of the effect of the pro-attitudinal treatment on visits to the websites of the potential pro-attitudinal outlets, compared to the control group. The regressions control for the outcome measure in baseline if it exists. The sample includes 1,648 participants with a liberal or conservative ideological leaning who installed the extension and provided permissions to access their posts for at least two weeks. Error bars reflect 90 percent confidence intervals.

Figure 5: Effect of the Treatments on News Slant



This figure shows the effect of the liberal and conservative treatments on the mean slant, in standard deviations, of all news individuals engaged with. In each panel, the dependent variable is the mean slant of outlets and the independent variable is the treatment. The regressions control for the outcome in baseline, if it exists. The sample includes participants who installed the extension and provided permissions to access their posts for at least two weeks following the intervention. Error bars reflect 90 percent confidence intervals.

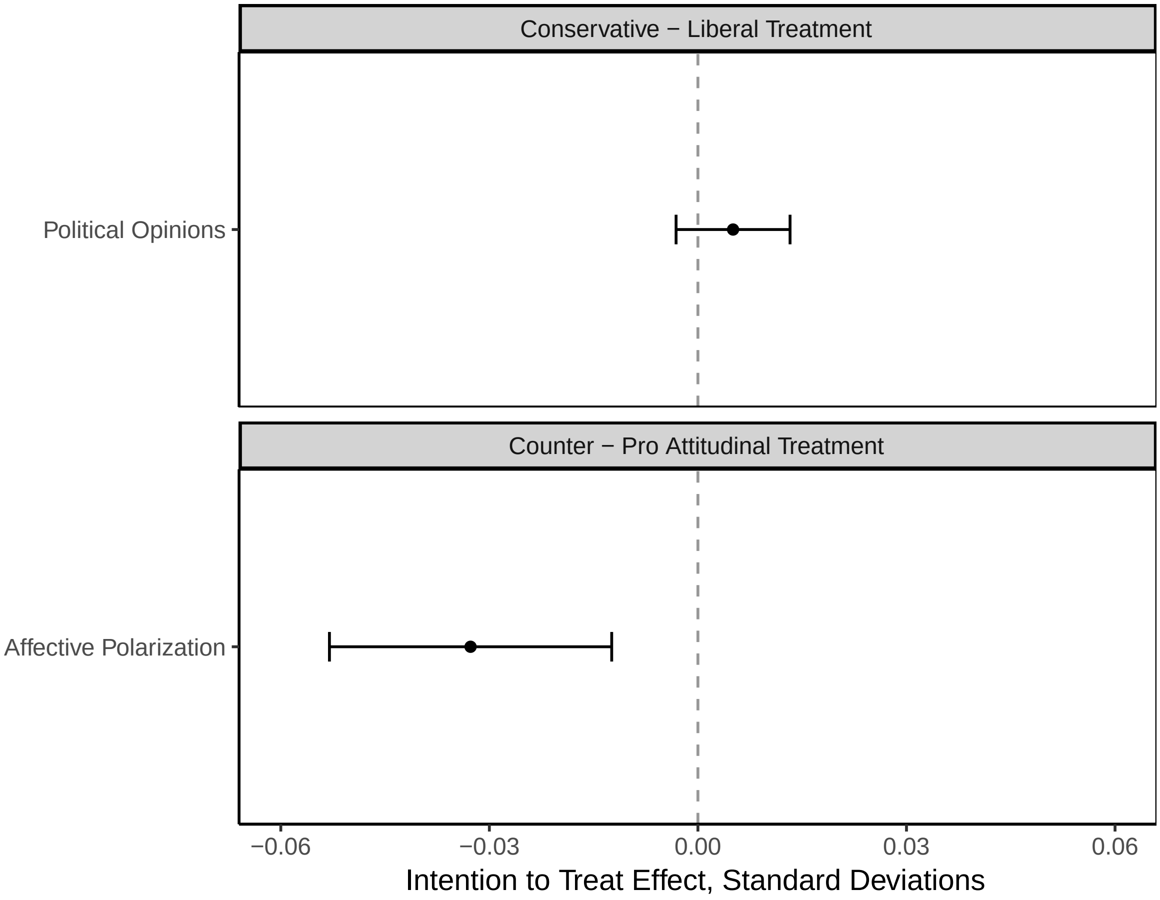
Figure 6: Effects of the Conservative Treatment on Mean Slant by Week, Compared to the Liberal Treatment


(a) News Exposure and Browsing Behavior

(b) Sharing Behavior

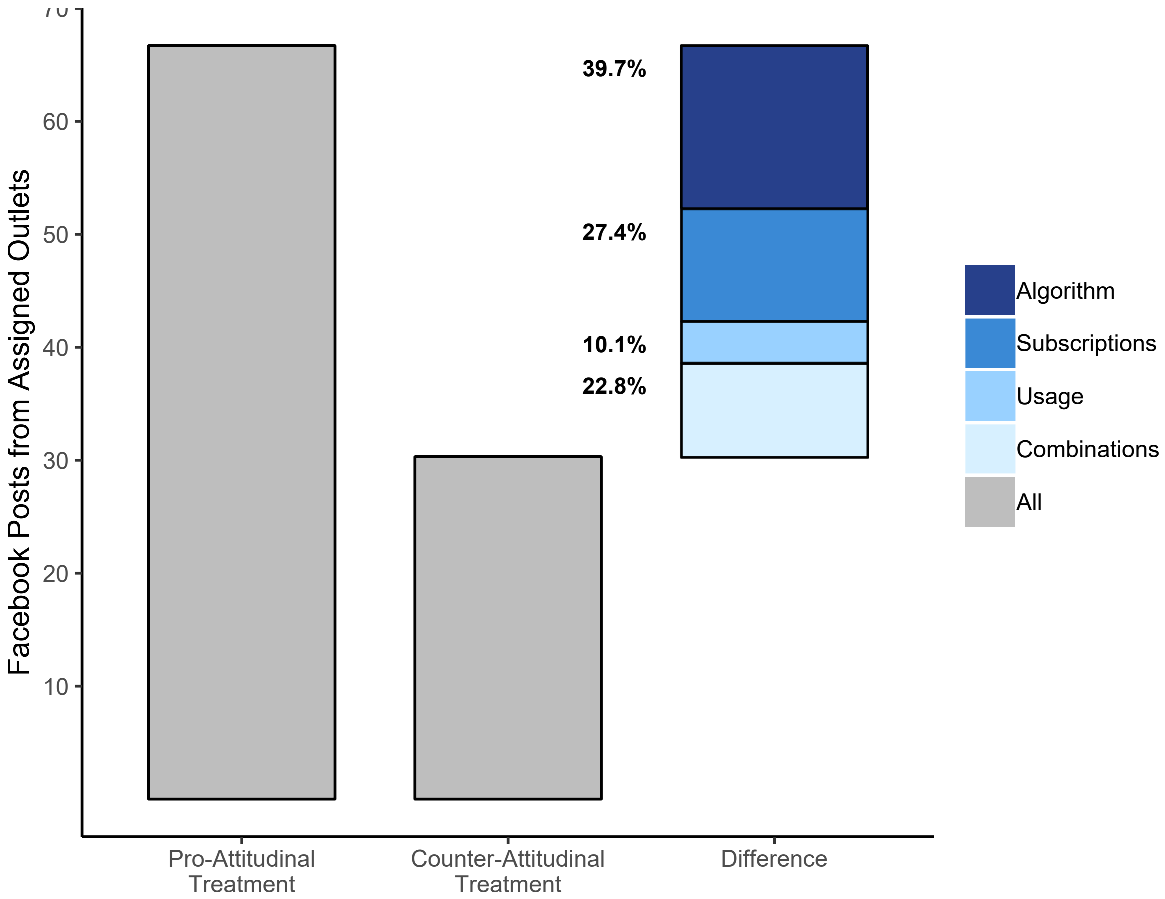
These figures show the difference between the effect of the liberal and conservative treatments on the mean slant of news engagement over time. Each panel presents a series of regressions, where the dependent variable is the slant of outlets in a specific week. The regressions control for the outcome in baseline when it exists. In the x-axis, relative week 1 is a full week immediately following the intervention. In sub-figure (a), the data is based on 1,596 participants who kept the extension installed for at least six weeks following the intervention. In sub-figure (b), the data is based on 29,131 participants who provided access to posts they shared for at least six weeks. Error bars reflect 90 percent confidence intervals.

Figure 7: Effect of the Treatments on Political Opinions and Polarization



This figure shows the effect of the treatments on the primary endline survey outcomes. The first panel shows the effect of the conservative treatment on the political opinions index, compared to the liberal treatment. A higher value is associated with a more conservative outcome. The second panel shows the effect of the counter-attitudinal treatment on the affective polarization index, compared to the pro-attitudinal treatment. A higher value is associated with a more polarized outcome. The indices are described in Section 2.4.2 and the regressions specifications are detailed in Section 2.5. The panels are based on 17,635 participants who took the endline survey. Error bars reflect 90 percent confidence intervals.

Figure 8: Decomposing the Gap Between Exposure to Posts from the Offered Pro-Attitudinal and Counter-Attitudinal Outlets



This figure decomposes the gap between the number of posts participants were exposed to from the offered pro-attitudinal and counter-attitudinal outlets. The y-axis is the number of posts seen per day and the x-axis is the treatment arm. *Algorithm* describes the gap explained by Facebook's tendency to show participants a greater share of posts from pro-attitudinal outlets (among all posts in the feed) conditional on subscriptions. *Subscriptions* describes the gap explained by participants' tendency to subscribe to more offered outlets in the pro-attitudinal treatment. *Usage* describes the gap explained by participants' tendency to view fewer posts on Facebook (use Facebook less often) in the counter-attitudinal treatment. *Combinations* describe interactions between these expressions. For example, a participant may have not subscribed to an outlet since it is counter-attitudinal, and she may have not viewed posts from the outlets even if she would have subscribed. Data is based on 1,059 participants in the pro- and counter-attitudinal treatments for which posts in the Facebook feed could be observed in the two weeks following the intervention and at least one post is observed. The calculations appear in Appendix C.7.

Table 1: Samples, Data Sources and Outcomes

Sample / Subsample	**Data Sources**	**Number of Participants and Retention**	**Main Outcomes Measured**
Baseline sample	Baseline survey; Facebook data on participants' subscriptions to outlets	37,494 (all participants)	Subscriptions to outlets in the intervention (compliance)
Access posts subsample	Facebook data for participants who provided permissions to access their posts and subscriptions for at least two weeks	34,592 (94% of participants who provided permissions in baseline)	Subscription to outlets over time; sharing behavior
Extension subsample	Browser data from participants who installed the chrome extension for at least two weeks	1,835 (81% of participants who installed the extension in baseline)	Posts observed in the Facebook feed (exposure); news sites visited (browsing behavior)
Endline survey subsample	Endline survey, approximately two months after baseline	17,635 (47% of participants who completed the baseline survey)	Political opinions; affective polarization

This table describes the main sample and the subsamples analyzed along with the data sources, the number of participants, and the main outcomes analyzed. The subsamples and data are described in Section 2.3. The outcomes are described in Section 2.4.

Table 2: Balance Table, Liberal and Conservative Treatments

	Mean			Difference		
Variable	Sample N=37,494	US	FB Users	Control - Lib.	Control - Cons.	Cons. - Lib.
Baseline Survey						
Ideology (-3, 3)	-0.61	0.17		0.01	0.01	0.00
Democrat	0.38	0.35	0.30	0.01	0.00	0.01
Republican	0.17	0.28	0.21	-0.01	0.00	-0.01
Independent	0.37	0.32	0.35	-0.00	-0.00	-0.00
Vote Support Clinton	0.53			-0.00	-0.00	-0.00
Vote Support Trump	0.26			0.00	-0.00	0.01
Feeling Therm., Rep.	29.07	43.06		0.11	0.25	-0.13
Feeling Therm., Dem.	46.99	48.70		0.40	0.46	-0.06
Difficult Pers., Rep. (1, 5)	3.13			0.02	0.00	0.02
Difficult Pers., Dem. (1, 5)	2.39			-0.00	0.01	-0.01
Facebook Echo Chamber	1.18		1.12	-0.00	-0.00	0.00
Follows News	3.35	2.42		0.01	0.01	-0.00
Most News Social Media	0.18	0.13		-0.00	0.00	-0.00
Device						
Took Survey Mobile	0.67			-0.01*	-0.00	-0.01*
Facebook						
Female	0.52	0.52	0.55	-0.01	-0.00	-0.00
Age	47.69	47.30	42.86	0.22	-0.13	0.35
Total Subscriptions	474			5.15	9.04	-3.89
News Outlets Slant (-1, 1)	-0.18			0.00	0.00	0.00
Access Posts, Pre-Treat.	0.98			0.00	0.01***	-0.00**
Attrition						
Took Followup Survey	0.47			0.03***	0.03***	-0.00
Access Posts, 2 Weeks	0.92			0.00	0.01**	-0.01**
Extension Install, 2 Weeks	0.05			0.00	-0.00	0.00
F-Test				1.21	0.89	1.05
P-Value				[0.21]	[0.64]	[0.39]

This table presents descriptive statistics, along with the difference between participants assigned to each treatment arm. *Vote Support* is the share of participants who voted for the candidate or did not vote and preferred the candidate. *Difficult Pers.* is whether participants find it difficult to see things from Democrats' / Republicans' point of view. *Facebook Echo Chamber* is whether the opinions participants see about government and politics on Facebook are in line with their views always or nearly all the time (3), most of the time (2), some of the time (1), or not too often (0). *Follows News* is whether participants follow government and politics always (4), most of the time (3), about half the time (2), some of the time (1), or never (0). *Total Subscriptions* is the number of Facebook pages participants subscribed to in baseline. *News Outlets Subscriptions* is subscriptions to pages of leading news outlets. *News Outlets Slant* is the slant of news outlets subscriptions. F-tests are calculated by regressing the treatment on the pre-treatment variables, with missing values replaced with a constant and an indicator for a missing value. Data sources for the US and Facebook population are specified in Appendix Section C.4.1. *p<0.1 **p<0.05 ***p<0.01.

Table 3: Compliance with the Treatments

	(1)	(2)	(3)	(4)
Cons. Treat., Cons. Ideology	0.497***	0.513***		
	(0.008)	(0.008)		
Lib. Treat., Cons. Ideology	0.333***	0.349***		
	(0.007)	(0.008)		
Cons. Treat., Lib. Ideology	0.555***	0.541***		
	(0.006)	(0.006)		
Lib. Treat., Lib. Ideology	0.636***	0.623***		
	(0.005)	(0.006)		
Know Slant			0.262***	0.230***
			(0.005)	(0.006)
Outlet Ideology, Abs. Value (Std. Dev.)			−0.072***	−0.047***
			(0.002)	(0.003)
Ideological Distance (Std. Dev.)			−0.083***	−0.083***
			(0.002)	(0.002)
Controls		X		X
Observations	36,728	36,728	97,937	97,937

This table estimates the association between participants' characteristics and compliance with each treatment arm. In columns (1)-(2), the dependent variable is whether the participant subscribed to at least one offered outlet and the independent variable is the interaction of participant's ideological leaning and her treatment assignment. In columns (3)-(4), the data is pooled such that each observation is a participant and an outlet offered. The dependent variable is whether a participant subscribed to a specific outlet. The independent variables are based on the outlet's perceived ideology where ideology is measured on a 7-point scale from extremely liberal to extremely conservative with an additional option of 'do not know'. *Ideological Distance* is the standardized difference between the participant's self-reported ideology and the outlet's perceived ideology. Columns (2), (4) control for age, age squared, gender, and the set of potential outlets defined for a participant, and column (4) also controls for outlet fixed effects. Columns (1)-(2) use robust standard errors and columns (3)-(4) cluster standard errors at the individual level. *p<0.1 **p<0.05 ***p<0.01

Table 4: Segregation Measures

(a) Comscore

Category	Share	Isol.	Seg.	Slant, Abs.	Cong.	Extreme Pro	Mod. Pro	Mod.	Mod. Counter	Extreme Counter
1) All Browsing		0.012	0.190	0.264	0.054	0.090	0.299	0.327	0.221	0.062
2) Direct	49.9%	0.013	0.213	0.263	0.056	0.067	0.303	0.365	0.219	0.045
3) Social	5.1%	0.026	0.280	0.358	0.085	0.147	0.330	0.193	0.235	0.096
4) Search	37.3%	0.008	0.176	0.286	0.059	0.117	0.299	0.284	0.220	0.081
5) Other	7.6%	0.003	0.216	0.300	0.058	0.082	0.329	0.301	0.234	0.054
6) FB	4.2%	0.032	0.287	0.354	0.090	0.161	0.311	0.210	0.215	0.105
7) Non-FB	95.8%	0.011	0.188	0.263	0.053	0.088	0.299	0.332	0.221	0.061

(b) Extension Data

Category	Share	Isol.	Seg.	Slant, Abs.	Cong.	Extreme Pro	Mod. Pro	Mod.	Mod. Counter	Extreme Counter
1) Subscribed		0.513	0.361	0.554	0.519	0.502	0.307	0.091	0.068	0.035
2) FB Feed		0.229	0.218	0.373	0.322	0.279	0.405	0.158	0.120	0.037
3) Friends	48.2%	0.154	0.167	0.313	0.252	0.222	0.415	0.172	0.148	0.042
4) Pages	40.0%	0.379	0.286	0.449	0.401	0.352	0.384	0.137	0.097	0.030
5) Ads	11.9%	0.290	0.262	0.419	0.346	0.288	0.393	0.168	0.113	0.038
6) Browsing		0.172	0.196	0.325	0.264	0.203	0.424	0.215	0.132	0.025
7) Not FB	85.9%	0.152	0.194	0.320	0.254	0.190	0.429	0.223	0.131	0.025
8) FB	14.1%	0.261	0.227	0.361	0.307	0.264	0.399	0.183	0.127	0.030
9) Friends	59.5%	0.191	0.208	0.329	0.267	0.236	0.391	0.197	0.140	0.036
10) Pages	40.5%	0.443	0.290	0.429	0.390	0.345	0.376	0.153	0.110	0.020
11) Shared		0.292	0.250	0.412	0.358	0.304	0.416	0.124	0.129	0.026

These tables display segregation measures for online and social media news engagement. The first sub-table is based on the Comscore data and the second is based on control group participants in the extension subsample. The segregation measures are defined in Section 3. For more details on how Facebook data was processed and suspected ads were identified see Appendix A.5.

Appendix For Online Publication

A Data Collection and Processing

A.1 Leading News Outlets

Throughout the paper, I analyze participants' engagement with leading outlets. The list of outlets and their slant are based on a dataset constructed by Bakshy et al. (2015). The authors use Facebook's internal data and classify links to hard and soft news. Hard news articles are related to issues including national news, politics, or world affairs. Soft news includes issues such as sports and entertainment. The alignment of each website is determined according to the self-reported ideology of Facebook users who share hard news links from the website. While many of the sites in the list are traditional news outlets, such as washingtonpost.com, others are much more partisan organizations, such as occupydemocrats.com

I exclude from the dataset the following popular websites which are not directly related to news: Amazon, Barack Obama, The White House, Twitter, Vimeo, Wikipedia, and YouTube. I also exclude MSN and AOL since these sites are aggregators of a wide variety of content, they may serve as homepages, and they are often visited for reasons not related to news consumption (Peterson et al., 2019). After processing the data, the list of leading outlets contains 487 websites.[60]

A.2 Comscore Data

The Comscore Web Behavior Database Panel is a subset of Comscore's opt-in Media Matrix Panel, which is weighted to represent the US Internet population. Previous studies showed that the Web Behavior Database Panel is representative of online buyers in the United States (Hortacsu et al., 2012). For each calendar year, Comscore has a separate dataset with separate demographics. When combining data for multiple years, I assign each individual the zip code in the last year for which data exists.

When classifying the referral channel through which a news site was visited, the referring channel is defined as social if the referring domain is one of the following: "facebook.com", "live.com", "t.co", "reddit.com", "pinterest.com", "youtube.com", "linkedin.com", "twitter.com", "tumblr.com", "instagram.com". I classify any referral domain that includes the word google (e.g. "google.com" or "google.co.uk") as a search domain along with the following domains: "yahoo.com", "bing.com", "ask.com", "duckduckgo.com", "searchencrypt.com", "searchlock.com",

[60] I merge websites that appear twice in the dataset, with and without a web reference, into one entry. For example, washingtonexaminer.com and www.washingtonexaminer.com are merged, with the slant defined as the mean slant of the two entries.

"searchincognito.com", "search.com", "searchprivacy.co", "safesear.ch", "myprivatesearch.com", "netfind.com". I classify a site as visited directly is there is no referral domain or if the referral domain is the same domain as the domain visited.

A.3 Surveys

A.3.1 Recruitment Ads

The Facebook ads recruiting participants to the baseline survey mentioned that a research survey was conducted by Yale University and that participants could win Amazon gift cards (Appendix Figure A.13). One version of the ad suggested that the survey was about politics and the other suggested that it is about American society.[61]

Most participants were recruited through ads targeting all Facebook users living in the US who are over 18 years old. A subset of the ads targeted conservatives or moderate individuals who are often under-represented in Internet samples (Allcott and Gentzkow, 2017; Yeager et al., 2011). Since the majority of participants took the survey on a mobile phone, an additional subset of ads focused on desktop users, to ensure that a large enough sample of participants will be offered an option to install the Chrome extension. A very small minority of users seemed to have a technical issue when taking the users using the iOS operating system and therefore iOS users were excluded from the target audience once this was discovered (the sample still contains many iOS users). Using Facebook Pixel, the ads targeted Facebook users who were more likely to begin the survey. While the survey was open and participants could share the link or ad with anyone, the vast majority of participants entered the survey as a result of the ad.[62]

A.3.2 Baseline Survey

The baseline survey took place from early February to mid-March 2018. 40,504 responders took the survey and reached the screen where the intervention occurs. Of those, 37,494 are included in the final sample. Responders are excluded from the final sample for the following reasons: missing information on outlets the responder subscribes to either because the responder did not provide permissions to access that data or since the data was not collected properly in real-time (2.38%); the responder already subscribed to too many of the outlets such that it was not possible

[61] I do not find evidence for heterogeneous effects on political opinions or affective polarization by the type of ad used.

[62] To test whether participants entered the survey because someone shared it with them, I provided participants with a slightly modified link to the baseline survey after they completed the survey, and asked them to use this link if they wish to share the survey. Only 0.57% of participants entered the survey using this link. Any individual exposed to an ad could also share the ad or the link that appears in the ad with other individuals. Approximately 95% of exposures to the ads during the recruitment period were directly due to a sponsored ad appearing in one's Facebook feed and not due to someone sharing the ad. Therefore, it is likely that the vast majority of participants entered the survey since a sponsored ad appeared in their feed.

to define for the responder four potential liberal outlets and four potential conservative outlets (4.01%); technical issues with the Qualtrics survey which prevented some data from being collected (0.90%); taking the survey a second time (0.01%).; responding carelessly (0.12%). Careless responders are defined as responders who completed all survey sections until the intervention exceptionally quickly (in under three minutes where the median time was eleven minutes) and responders who did not answer at least half of the closed-ended, non-required questions, or who did not answer any question in the final page before the intervention. All the criteria determining whether to exclude a responder are based on survey data submitted before the intervention occurs. Finally, to slightly reduce the number of outlets, alternative outlets which are defined as potential outlets for fewer than 20 participants are excluded from the experiment, along with the participants for which these outlets were defined as potential outlets. This removes fewer than 0.1% of participants from the baseline sample.

A.3.3 Endline Survey

Participants were invited to the endline survey between mid-April and early June 2018. Participants were mostly recruited to the survey using emails and Facebook ads.[63] To match endline survey responses with baseline survey responses, participants were asked to log in to the endline survey through Facebook or supply an email address. I match endline responses based on the following criteria: email address the survey invitation was sent to, Facebook id, email address entered in the survey, combination of zip code, first and last name if the combination is unique, and combination of first and last name if the combination is unique. 98.73% of responses were matched with baseline responders.

17,635 participants are included in the endline survey subsample. If the same individual took the endline survey more than once, uncompleted surveys are excluded. If multiple observations still exist, only the first response is included for the individual. Overall, 0.41% of valid matched responses were excluded as duplicates. 0.02% of responses were also excluded for taking the survey carelessly if the survey was completed exceptionally quickly (spent less than 20 seconds per survey page, compared to a median time of 67 seconds).

A.4 Facebook Data on Subscriptions and Posts Shared

I collect data on outlets participants subscribed to (pages "liked") and posts they shared using a Facebook app, which provides an interface between a Facebook account and the survey.[64] The data allowed me to customize the survey by ensuring participants are not offered outlets they

[63] A small share of participants was recruited through an invitation in the browser extension or a Facebook notification.

[64] To minimize measurement error, data from the app was collected using several methods, including code running in the background of the baseline survey, a web service, and multiple scripts that ran for the duration of the experiment.

already subscribed to and including questions asking about the offered outlets. The app was approved through the standard Facebook review process.

I include in the analysis the following types of posts: link, note, status, and video. I focus on these posts since they are more likely to contain political content relevant to the experiment. In some cases, the outlets offered to participants published posts that contain only a photo and text (for example, Fox News published posts with quotes related to the news without an accompanying link or video). These posts are defined as photos and are excluded from the analysis. Therefore, the effect I find on the number of posts shared as a result of the experiment is probably slightly lower than the actual effects.

I match Facebook posts to leading outlets based on the Facebook page which shared a post. If a post is not matched with any Facebook page, I determine the slant of the post based on the domain of a link included in the post. For outlets offered in the experiment, I expand the list of domains in the Bakshy et al. (2015) dataset to decrease measurement error. For each outlet, I create a list of relevant domains by checking which domains were shared by the Facebook page associated with the outlet and including the most dominant domains and any other domain directly linked to the outlet. For example, in addition to associating "huffingtonpost.com" with the Huffington Post, I associate "huffpost.com" and other similar domains.

If a link refers to a short alias, created by URL-shortening services such as tinyurl.com, it cannot be directly matched to an outlet based on the domain. Therefore, each URL in a post shared is first converted to the final re-directed URL before being matched to the list of domains.

I also observe participants' gender and age on Facebook. I define participants' age as 2018 minus their birth year and replace any age above 90 with missing.

A.5 Extension Data

A.5.1 Browsing Behavior

I collect data on the Facebook feed and browsing behavior using the Chrome browser extension. I exclude URLs that were visited for less than one second before another URL in the same domain was visited, as it is likely that the user did not observe the content of the website. If a URL is visited more than once within a 20-minute window, only the first visit is included. News sites visited are matched to outlets based on their domain. A news site is determined to have been visited through Facebook if the website visited appeared in the participant's Facebook feed in the 20 minutes proceeding to the website being visited.[65] All URLs are first converted to the final re-directed URL before matching posts or news sites visited.

[65] The time window used is not particularly important. If a 5-minute window is used the number of sites determined to have been visited through Facebook in the two weeks following the intervention decreases by less than 3%, and if a 60-minute window is used, the number of sites increases by less than 3%.

A.5.2 Facebook Feed

I observe posts appearing in participants' Facebook feeds when participants have the extension installed and use their computer mouse to scroll down the Facebook feed. I do not observe posts unless they appear on the participants' screen. While the extension was designed to work with Google Chrome, it can also work with similar browsers and a very small number of users installed it on alternative browsers, such as Vivaldi.

I assign posts appearing in participants' Facebook feeds to outlets using the following hierarchy:

1. The post was created by a leading news outlet (e.g., a post by the New York Times)
2. The post shared a post created by a leading news outlet (e.g., a friend shared a post by the New York Times).
3. The post includes a link to a leading news outlet (e.g., a friend shared a New York Times link). If the post shares no link, but the text of the post contains a link, I use that link instead. I first convert all links to their final re-directed URL.

I exclude less than 1% of posts, where I cannot observe whether the post is shared by a page or a friend (these posts could be sharing content from other Facebook features such as a Facebook Game or Town Hall).

In my data, I cannot precisely identify whether a post is sponsored or organic. Instead, I use three techniques to identify ads. First, I assume that any post seen by at least two participants who did not subscribe to the post's page is sponsored. Second, I assume that any post seen at least five days after the first time it appears in my dataset is sponsored. The algorithm typically shows users only recent posts. However, a sponsored post will continuously appear in users' feeds as long as the advertiser continue to pay for more ads. Third, I assume that any post that appears more than twice in a participant's feed for at least two participants is an advertisement. Facebook's algorithm does not tend to show the same post many times to the same user, however, advertisers can choose to maximize impressions and thus may show the same post repetitively.

When determining whether a post is sponsored, I assume that two posts from the same page with the same text are the same post even if they have a different id since advertisers can use two posts to run two identical advertisements.[66]

While these criteria are far from perfect, they do seem to identify many ads. For example, based on my classification, the top ten words that are most likely to appear in posts identified as ads, compared to organic posts are: "get, now, free, new, today, just, time, one, us, help". In contrast, the top ten words most likely to appear in organic posts are: "trump, president, one, new, people, just, school, like, gun, now."[67]

[66] I make this assumption when the text is at least 20 characters long

[67] The terms exclude stop words along with the words http, can, said, see.

B Additional Details on Empirical Strategy

B.1 Segregation Measures

B.1.1 Isolation

Isolation is the difference between the mean share of conservatives that conservatives are exposed to in the outlets they visit and the mean share of conservatives that liberals are exposed to. Exposure in this context is defined as the share of conservative browsing the websitesamong all the site's visitors.

$$Isol_i = \sum_{i \in \{C_i\}} ShareCons_i * ConsExposure_i - \sum_{i \in \{L_i\}} ShareLib_i * ConsExposure_i$$

where $ShareCons_i$ is the share of outlets visited by individual i among all outlets visited by conservatives (i's weight among conservatives), $\{C_i\}$ is the set of conservative individuals, $\{L_i\}$ is the set of liberal individuals, and $ConsExposure_j$ is exposure to conservatives by individual i.

Exposure can be calculated as the average share of conservatives among all outlets visited by individual i. In order to prevent a small sample bias, the average share does not include the visits by individual i:

$$ConsExposure_i = \sum_j \frac{Visits_{ij}}{Visits_i} * \frac{Cons_j - Visits_{ij}}{Visits_j - Visits_{ij}}$$

where $Vis\dot{i}ts_{ij}$ is the number of visits of individual i to outlet j, $Visits_i$ is total visits by individual i, $Visits_j$ is total visits to site j, and $Cons_j$ is total conservative visits to site j.

B.1.2 Segregation

Segregation is defined as the scaled standard deviation of partisan news exposure. This can be interpreted as the square root of the expected square distance between the slant of news sites visited by two random participants in the sample (Flaxman et al., 2016):

$$Seg_i = \sqrt{2} * std.dev(Sl\bar{a}nt_i)$$

where $Sl\bar{a}nt_i$ is the mean slant of outlets visited by individual i. The slant of each outlet j is based on Bakshy et al. (2015) but first normalized to the unit interval (by adding one and dividing by two).

B.1.3 Absolute value of slant

To measure the extremity of news consumption, I calculate the absolute value of mean consumption slant as:

$$AbsSlant_i = \sum_i \frac{|\bar{Slant_i}|}{N}$$

where $\bar{Slant_i}$ is the mean slant of outlets visited by individual i. The slant of each outlet j is based on Bakshy et al. (2015) such that a middle-of-the-road outlet has a slant of zero, a completely conservative outlet has a slant of 1 and a completely liberal outlet has a slant of -1

B.1.4 Congruence

I define congruence as exposure to more extreme content matching the consumer's ideology:

$$Congruence_i = \sum_i \frac{(\bar{Slant_i} * IdeoLeaning_i)}{N}$$

where $\bar{Slant_i}$ has the same definition as in the previous measure and $IdeoLeaning$ is defined as 1 for a conservative participant and -1 for a liberal participant.

B.2 Pre-Analysis Plan

The main outcome and hypotheses tested in this study were pre-registered in the AEA RCT Registry.[68] The analysis deviates from the pre-analysis plan in two important ways. First, I use equal weights for the measures composing the indices, while the plan states that the weights for the index variables will be determined by the inverse of the covariance between the measures (Anderson, 2008). This method is not used since it generates negative weights. With negative weights, the interpretation of the index is less clear. For example, the question on President Trump's approval rating received a negative weight according to this index, which means that *ceteris paribus*, a participant who has a more favorable opinion on Trump would be considered more liberal.

Column (2) of Appendix Table A.17 shows that the effect on affective polarization is robust to using inverse-covariance weights. This method does not cleanly generate weights for individuals with missing outcomes. In column (3), weights are created using the inverse-covariance method based on participants with no missing outcomes and then renormalized to sum to one for each participant with missing outcomes. This creates an index value for all participants who have at least one non-missing outcome. The results remain very similar.

Appendix Table A.17a estimates the effect on the political opinions index using inverse-covariance weights. Since the inverse-covariance method generates negative weights, columns (4) and (5)

[68] AEA RCT Registry Trial 0002713.

repeat the analysis with negative weights replaced with zero and the weights renormalized accordingly. While there is some variation in the results, the most straight-forward comparison is between columns (1) and (5). These columns focus on the same participants, do not use different signs for the same weights, but assign different weights to the outcomes composing the index. In column (5), the effect of the conservative treatment is slightly larger but still economically small and not statistically significant.

The second important deviation from the pre-analysis plan is that the polarization index originally included five attitudinal measures and three behavioral measures, while only the attitudinal measures are analyzed in this paper. The behavioral measures were based on a question in the endline survey asking participants whether they would "like" or share a post stating that "In seeking truth, you have to get both sides of a story." The primary behavioral outcome is composed of an index of the following measures: did participants state they will share the post, did participants state they would "like" the post, did participants actually share the post. However, it was not possible to analyze the posts of a large share of participants by the time they took the endline survey, partly due to the unexpected Cambridge Analytica scandal, which led many individuals to revoke access to the posts they share. Furthermore, the behavioral measure turned out not to measure polarization well. While a measure of polarization should typically be correlated with partisanship, there was almost no correlation between being partisan and the behavioral outcomes.[69]

Column (1) of Appendix Table A.18 shows that the primary estimate is still significant when using all eight variables in the polarization index.[70] Column (3) measures the effect only on the behavioral outcomes (for most participants data not exist on whether posts were shared so this index is mostly based on the self-reported survey answers). The effect of the treatments is small and not statistically significant. While this result does not change the conclusions regarding affective polarization, it is interesting to note that exposure to counter-attitudinal outlets does not affect participants' self-reported willingness to share or like a post regarding the importance of seeking both sides of a story.

There are various other minor changes compared to the pre-analysis plan, include the following. In the plan, I stated that I will estimate the results excluding the first two days after the intervention. Instead, I estimate the results for each week (or month separately). In the plan, I stated I would control for the randomization block and for whether the participant used the iOS operating system. I exclude the iOS variable for simplicity (this does not affect the primary endline survey results). I do not control for the randomization blocks (strata) since due to attrition, some strata have only one or two respondents instead of the original three respondents defined for each block.

[69] The correlation between the behavioral polarization measures and the absolute value of a baseline scale of partisan affiliation (where 0 is no party identification, 1 is leaning toward a party, 2 is identifying with a party and, 3 is strongly identifying with a party) is only 0.04-0.06. The correlation between the affective polarization measures and partisan affiliation is 0.22-0.46.

[70] The effect when all eight variables are used to construct a polarization index is smaller in index points than the effect when the five attitudinal measures are used. When standardizing the indices with respect to the control group, the effects are similar since the index created when using all eight variables has less variation in the control group.

When controlling for the block, I am only able to analyze a subset of participants. The results for that subset are essentially the same with and without controlling for strata. I do not report raw or adjusted p-values for each index component of the political opinions and affective polarization measures, as I do not focus on the individual components. Instead, I present each component visually in appendix figures.

In the pre-analysis plan ideological leaning is defined first by self-reported ideology and then by party affiliation. I prefer using party affiliation as the main variable defining ideological leaning to make the study comparable to other papers, which tend to focus on party affiliation (Druckman and Levendusky, 2019). The results are robust to the original definition. I control for ideological leaning in the primary endline survey regressions. In contrast to the plan, I do not present several demographic variables in the balance table since they suffer from post-treatment bias and do not impute them since I already have rich survey and social media data.

B.3 Controls

To increase power, when estimating the effect on political opinion and affective polarization, I control for a set of pre-registered covariates. I control for self-reported ideology, party affiliation, approval of President Trump, ideological leaning, age, age squared, gender. Age and gender are included in the Facebook data provided when participants log in to the survey and the remaining covariates are based on the baseline survey. Self-reported ideology is a nominal variable with seven ideological options from very liberal to very conservative and an option for participants who have not thought much about this. Party affiliation is a nominal variable with seven affiliation options ranging from strong Democrat to strong Republican along with an option of "other party". Approval of Trump is a nominal variable with four options ranging from strongly disapprove to strongly approve.

When estimating the effect on political opinions, I also control for the following baseline survey questions: feeling toward President Trump (0-100 integer); worry about illegal immigration (nominal variable with the options not at all, only a little, fair amount, great deal); does the participant believes Mueller is conducting a fair investigation (nominal variable with the options yes, no, do not know), and whether the participant thinks Trump has attempted to obstruct the investigation into Russian interference in the election (nominal variable with the options yes, no, do not know).

When estimating the effect on affective polarization, I also control for the baseline values of the *feeling thermometer* and *difficult perspective* measures (defined in Section 2.4.2).

In all regressions, if a covariate includes missing values, the missing values are coded to a constant and an additional dummy control is added to the regression indicating whether a value is missing. Regressions testing for heterogeneous effects also control for each participant's potential outlets since individuals who were assigned the alternative outlet may have different characteristics than individuals who were assigned the primary outlets.

C Additional Analysis

C.1 Survey Purpose

At the end of the baseline survey, participants were presented with the following question: "If you had to guess, what would you say is the primary purpose of this study?" Appendix Table A.19 shows the most common three-word expressions participants mentioned according to their treatment assignment. Unsurprisingly, participants understood that the study is on media and politics, as most questions focused on these topics and the consent form stated that this is the topic of the study. Among the most common expressions, there are not many substantial differences between the treatments.

Appendix Table A.20 presents the expressions with the largest differential usage between the treatment arms and the control group. While participants in both the pro- and counter-attitudinal treatments mentioned terms such as "echo chamber"and "social media" more than the control group (probably due to the text of the intervention encouraging participants to "Like" Facebook pages), the differences between the two treatment arms in the usage of these terms is small. When comparing the pro- and counter-attitudinal treatments to each other, almost no substantial differences stand out. One exception is that some participants in the counter-attitudinal treatment thought the purpose of the survey was to get them to like liberal Facebook pages. These participants probably were not pleased with the experimenter trying to "push liberal" content (that was not the actual purpose of the experiment, of course) and therefore it is unlikely that they expressed opinions aligned with these outlets in order to make an impression on the experimenter. In any case, while these expressions represent a relatively large difference between the treatments, they are not mentioned often.

Overall, this section suggests that participants in the counter-attitudinal treatment did not perceive the experimenter's expectations substantially differently than participants in the pro-attitudinal treatment. This conclusion does not rule out that experimenter effects are driving some of the results. It is possible, for example, that participants in the pro- and counter-attitudinal treatments understood that the study attempts to analyze the effect of news outlets on political opinions, they remembered which outlets they were offered, and tried in the endline survey to convey attitudes more similar to the outlets offered (e.g., a more positive opinion toward the Republican party if they were offered conservative outlets). However, at least it is unlikely that differential expectations of the experimenter's objective are driving the main results.

C.2 Analysis of the Content Participants Engaged With

This section shows that participants in the counter-attitudinal treatment were slightly less likely to engage with political content, but that in both treatments, the most common content participants engaged with seems to be political in nature. I analyze the content of posts participants were exposed to in their feed, links they visited, and posts they shared as a result of the intervention using three methods. First, I show the most common phrases mentioned in the posts. Second, I define certain terms as political and analyze the share of political words by the outlet's slant and whether it matched the participant's ideology. Third, I analyze the section and outlet where each article appeared among the URLs appearing in the posts.

An important challenge in this analysis is that the posts affected by the treatment cannot be cleanly identified. For example, participants in the control group visited the news sites of their potential counter-attitudinal outlets approximately 1.7 times in the two weeks following the intervention, while participants in the counter-attitudinal treatment visited these websites approximately 2.9 times (as shown in Figure 4). While clearly the participants were affected by the treatment, I cannot identify which of their visits to counter-attitudinal news sites would have occurred in a counterfactual with no intervention. I focus on posts affected directly by the intervention by analyzing only posts shared by pages participants subscribed to in the experiment (excluding suspected ads). While this decreases the likelihood of including posts that participants would have engaged with without the intervention, it also excludes some posts that were affected by the intervention. For example, participants often visited the websites of the offered outlets indirectly, even when they did not observe the specific link to an article in their feed (as shown in Figure A.4).

Throughout this section, I focus on the eight weeks following the intervention to increase the number of data points. Still, the analysis of posts with links participants clicked on is based on a relatively small sample of 2,243 pro-attitudinal and 1,262 counter-attitudinal posts, and therefore should be interpreted cautiously (posts participants were exposed to and posts participants shared have larger sample sizes). To reduce variability in the text analyzed, I include in the analysis only posts from the four primary outlets and first alternative outlet that were offered to participants. This excludes less than 3% of posts participants were exposed to.

Before discussing the results, an important caveat is in order. This section is descriptive and its purpose is to show what content participants engaged with according to whether the outlets they were offered were pro- or counter-attitudinal. When comparing the content shared by liberals who subscribed to liberal outlets (pro-attitudinal) with content shared by conservatives who subscribed to liberal outlets (counter-attitudinal), I am *not* estimating the causal effect of the treatments, as the compositions of the two groups compared are different by definition.

C.2.1 Most Common Phrases

Appendix Table A.21 shows the most common phrases in posts participants were exposed to in their feed, in posts with links participants visited, and in posts shared by participants. I first remove punctuation, terms that appear in only one outlet, media-related terms or terms that were likely to be covered mostly by specific outlets (e.g., "write" or "New York"), and then stem the words appearing in the posts.[71]

The most common phrases participants were exposed to are political and are usually related either to President Trump, the aftermath of the Parkland school shooting, or the Mueller investigation. The phrases appearing in posts participants clicked are similar to the phrases in posts participants were exposed to.

The posts shared should not be directly compared to the posts participants were exposed to or clicked since the data is based on two different subsamples. Regardless, it is clear that posts shared are often political even when participants shared posts in the counter-attitudinal treatment. However, the response to scandals may be heterogeneous. For example, liberals are more likely to share articles mentioning Robert Mueller in both the pro- and counter-attitudinal treatments. Similarly, liberals in the liberal treatment are more likely to share articles mentioning Stormy Daniels and conservatives in the conservative treatment are more likely to share articles mentioning Hillary Clinton.

C.2.2 Share of Posts Mentioning Political Words

Focusing on the most common words allows us to understand which topics were most prominent but does not provide a complete analysis of the posts, especially if there is a lot of variability in the posts' content. In this subsection, I use a simple measure to determine a lower bound for the share of political posts. I define a post as political if it contains terms related to political figures ("biden, bolton, carson, clinton, devos, kushner, manafort, mccabe, mcconnell, michael cohen, obama, pelosi, pence, pruitt, tillerson, trump"), political parties ("conservative, democrat, dnc, gop, liberal, republican, the left, the right"), political institutions ("congress, elect, politic, senate, vote, white house") or political issues ("ar 15, daca, gun control, gun law, gun right, immigration, mass shooting, nra, parkland, sanctuary city, sanctuary state, school shooting, tax cut, walkout"). I search for the terms in the post's text, its URL, and any commentary on the post if it is shared.

Remarkably, more than half of the posts observed, clicked, and shared, are political. This is probably a lower bound for the actual number of political terms since posts including the terms I mentioned are almost always political but there are other political posts not captured by these terms (e.g., posts about race relations, gender issues, climate change and additional posts about gun legislation that do not include a unique term that can be clearly identified as political).

Appendix Figure A.14 shows that participants in the pro-attitudinal treatment were generally more likely to engage with political posts. However, the difference between the pro- and counter-attitudinal treatments is surprisingly small with one notable exception. Among liberals who shared posts from liberal outlets they were offered, 68% of posts were political, compared to 41% among conservatives who shared posts from the offered liberal outlets.

[71] In addition to stop words, I remove the following terms: bit, breaking news, can, comment, fox friend, fox news, http, https, journal, last week, new york, new york time, news, nyt, opinion, said, say, times, wall street journal, washington post, write, write the editori board, wsj, year old.

Still, it may be surprising that a large portion of the counter-attitudinal posts shared by participants was political. Why do participants share these posts? Anecdotally, there seem to be various reasons. Some posts are written by moderate columnists in a counter-attitudinal outlet (e.g., William A. Galston at the Wall Street Journal), others focus on rare bipartisan topics (e.g., a bill against sex trafficking), or report topical news without expressing strong opinions. In other cases, the posts may tackle issues where the outlet does not completely share the party's line, or where the participants may not agree with the party (e.g., conservatives who oppose the NRA's positions). There were also cases where participants share the posts with a negative comment, even though these are less common than might be expected. Finally, in a few cases, participants admitted they are sharing posts from outlets they usually would not share. This suggests that typically participants did not start sharing partisan news completely supporting the other side, but they may have shared articles with more nuanced positions in counter-attitudinal outlets.

C.2.3 Outlets and Sections

Instead of determining the posts' topics based on words in the post, I can analyze the content participants engaged with using the outlets' own classification of their articles. I use the link appearing in posts to determine their section. Most outlets classify articles into sections, such as News, Business, and Arts, and mention the section on their website or the website's HTML. For others, such as the Washington Post, the section is not clearly defined but can be determined simply by the article's URL.[72] This method is not perfect. MSNBC usually does not classify articles and videos into sections and Slate often creates short links for its URLs which were no longer available when I determined the link's section. Still, the advantage of this method is that it completely relies on internal decisions by the outlets.

Appendix Figure A.15 shows the most common outlets and sections participants were exposed to. The figure mostly reflects the different preferences of participants when subscribing to outlets. Liberals mostly avoided "liking" Fox News when it was offered and preferred the Wall Street Journal. They were more likely to already subscribe to one of the primary liberal outlets in baseline, and therefore, more likely to be offered to subscribe to Washington Post, the first alternative liberal outlet.

Appendix Figure A.16 suggests that participants clicked a larger share of posts about culture or arts compared to the share observed in the feed. For example, entertainment articles from Huffington Post and cultural articles from the Washington Times are more prominent in this figure. Interestingly, this holds both for participants in the pro- and counter-attitudinal treatments. However, posts with links to politics and national news are still most likely to be clicked in both treatments.

The differences between the posts shared by participants are more stark. For example, Appendix Figure A.17 shows that conservatives shared Huffington Post articles in the parenting, women,

[72] For example:, https://www.washingtonpost.com/**politics**/the-unending-campaign-of-donald-trump

or queer voices sections, while among posts shared by liberals, these sections form a very small minority.[73] Still, within each outlet, the dominant sections among posts shared are typically the political or national news sections, even in the counter-attitudinal treatment.

C.3 Heterogeneous Effects

In the pre-analysis plan, I stated that I will test for heterogeneous effects based on whether participants are ideological, whether they are in an echo chamber, the openness of participants, and whether they are sophisticated.

I define participants as *Ideological* if the absolute value of their self-reported ideology on the 7 point scale (from -3 for very liberal to +3 for very conservative) is above or equals the median.

I use two measures of being in an echo chamber. The variable *Echo Chamber* is whether the answer to "Thinking about the opinions you see people post about government and politics on Facebook, how often are they in line with your own views" is above or equals the median. *Seen Counter Att.* is whether the share of potential counter-attitudinal outlets, among all potential outlets, participants reported seeing in their feed in baseline is above or equals the median.

I measure whether a participant has an *Open Personality* according to whether her average agreement with the following statements is above or below the median: "I see myself as open to new experiences, complex" and the reverse values of "I see myself as conventional, uncreative." The questions based on Gosling et al. (2003). I define participants as *Certain* in their opinions if their answer to "Generally speaking, how certain are you of your political opinions?" is above or equals the median.

I define participants as *Sophisticated* if they answered one of the following questions correctly: "Suppose 110 members of a local government voted on an infrastructure bill. The bill passed by a margin of 100 votes. How many members voted against the bill", "Suppose the number of US citizens on the internet doubles every month. If it took 48 months for the entire US population to have internet access, how many months did it take for half the population to have internet access". These questions are based on the Cognitive Reflection Test (Shane, 2005).

In addition to the pre-registered tests, I explore the effect of several additional moderators. *Most News Social Media* is whether participants reporting getting most of their news about government and politics through social networking sites. Participants have *High News Subscriptions* if their baseline subscription to news outlets on Facebook is above or equals the median. Participants are considered *Exposed to Outlets* if their self-reported exposure to posts from the eight potential outlets in baseline is above or equals the median. Participants are considered to *Know Outlets Slant* if the distance between their perceived slant of the potential outlets and the average perceived slant by participants with the same self-reported ideology is below the median. Participants are

[73] Interestingly, almost no articles shared were in the sports section (less than 1% of articles for which a section could be identified).

considered to *Follow the News* if their answer to "how often do you pay attention to what's going on in government and politics?" is above the median. Participants are considered to have a *High Feeling Thermometer Difference* if the difference between their feeling toward their own party and the opposing party is above or equals the median. Finally, participants are considered *Conservative* if their ideological leaning is conservative, *Older* if their age is above or equal to the median age, and *Female* if they identify in Facebook as female.

When analyzing heterogeneity in the effects of the pro- and counter-attitudinal treatments, I do not distinguish between heterogeneity due to differences in the participants' ideology and heterogeneity due to differences in the outlets offered. For example, if conservatives are affected more by the pro-attitudinal treatment, that could be due to conservatives being more persuadable or because Fox News is more persuasive than New York Times.

Appendix Figures A.18 and A.19 estimate heterogeneous effects on subscribing to outlets, exposure to posts from outlets, and visiting the outlets' websites. Each row represents a separate regression estimating the effect of interacting the pro- or counter-attitudinal treatment with the specified variable, where the reference group is the control group.[74] A higher value means individuals were more likely to engage with the pro- or counter-attitudinal potential outlets as a result of the pro- or counter-attitudinal treatment, respectively.

Ideological individuals are more likely to subscribe to pro-attitudinal outlets and less likely to subscribe to counter-attitudinal outlets. Participants who are more certain in their opinions, and who follow the news are also less likely to subscribe to counter-attitudinal outlets. Similarly, ideological participants, along with participants following the news and participants who are more polarized in baseline, are less likely to visit these outlets. Finally, participants who subscribe to many outlets in baseline are more likely to subscribe to counter-attitudinal outlets. Interestingly, even though they subscribe at higher rates, they are *less* likely to be exposed to these outlets in their feed as a result of the intervention, probably since there is more competition for space in their feed.

The left panel of Appendix Figure A.20 shows that the effect on political opinions is mostly homogeneous (i.e., most participants were not persuaded by the treatments). The right panel of Appendix Figure A.20 does not show strong heterogeneous effects on affective polarization according to most covariates tested. The strongest heterogeneous effect found is based on the baseline feeling thermometer measure for affective polarization. The effect on affective polarization is weaker among participants who were more polarized in baseline. However, this result is significant at the 10% level and the results are not adjusted for multiple hypothesis testing, and therefore more research is needed to explore heterogeneity in affective polarization.

[74] The results of most heterogeneous effects are similar when estimating all the heterogeneous effects on either political opinions or affective polarization simultaneously in one regression.

C.4 Reweighting for National Representativeness

C.4.1 Data sources

To reweight the sample to match the US population, I use the following data sources. The medium where Americans get most of their news is based on the Pew American Trends Panel Wave 23 (November-December 2016). All other US data is based on the 2016 American National Election Survey (ANES). The estimates are based on pre-election ANES questions, besides vote or support for a presidential candidate, which is based on the post-election survey.

In Table 2, I also present demographics for Facebook users. Data on whether the opinions Facebook users see about government and politics on Facebook are in line with their views is based on a question in the Pew American Trends Panel Wave 1 (March-April 2014) asked among respondents who pay attention to posts about government and politics on Facebook. All other data on Facebook users is based on the 2018 Pew Core Trends Survey.

C.4.2 Analysis

In this section, I reweight the sample to match the national population using the entropy weighting procedure (Hainmueller, 2012). I match the following subset of control covariates: self-reported ideology (mean value on a scale of 1-7), the share of participants identifying as Democrats, Republicans, and Independents, the difference between the participants feeling toward their party and the opposing party, age, and the share of females. For the feeling thermometer, self-reported ideology, age, and gender covariates, missing variables are first replaced with the mean value (less than 5% of observations are missing for each of these variables). When analyzing the effects of the pro- and counter-attitudinal treatments, I compare the sample to the US population for which an ideological leaning can be defined and use those means to reweight the sample.[75]

Appendix Tables A.22 and A.23 show that reweighting the sample does not change the main conclusions of the study. The effect on the slant of posts participants were exposed to increases slightly. The effect on sites visited, posts shared, political opinion, and affective polarization remain essentially the same, although the confidence intervals are wider. These tables should be interpreted with caution. It is likely that even after reweighting, the sample is still different than the national population on unobservables or covariates not used when reweighting the sample. Still, the tables show that it is unlikely that an effect on affective polarization is only found because the survey sample is more liberal or more polarized than the rest of the population.

[75] I include respondents who identify or lean toward one of the parties, who define themselves as liberal or conservative, or who voted, intended to vote or preferred Donald Trump or Hillary Clinton, according to the ANES pre-election survey. Overall, 94% of respondents in the ANES survey are included.

C.5 Predicted Treatment Effect for the Full Baseline Sample

In the previous section, I reweighted participants to match the US population. In this section, I predict the main treatment effect for the entire baseline sample. While the baseline sample is not nationally representative, such an estimation provides several advantages. First, it estimates the same results among a larger group of participants that are more representative than the extension and endline survey subsamples, using a large set of Facebook and survey covariates. Second, it alleviates concerns that differential attrition by some observable characteristics is driving the results.

I first estimate heterogeneous effects on the slant of posts observed, the slant of news sites visited, the political opinions index, and the affective polarization index. The effects on media engagement are estimated in the extension subsample and the effects on self-reported opinions and attitudes are measured in the endline survey subsample.[76] I exclude the control group in these estimates so the interpretation is the effect of the conservative treatment on conservative media consumption and conservative opinions, compared to the liberal treatment, or the effect of the pro-attitudinal treatment on polarization, compared to the counter-attitudinal treatment. I estimate heterogeneous effects using causal forests (Wager and Athey, 2018). The intuition behind causal forests is that one part of the sample is used to determine how to split each tree and another part is used to estimate heterogeneity. If the same sample was used for both processes, heterogeneity would be overestimated due to overfitting, as the sample would be split according to the covariates that happen to predict heterogeneous effects in this particular sample.

I use a large set of covariates including almost all close-ended baseline survey questions and data from Facebook on the age, age squared, and gender of the participant, the number of pages liked by the participant in baseline, and the number of pages the participant liked in 2017. In addition, I include covariates for whether each of the outlets in the experiment could have been potentially offered to the participants and whether the participant liked a set of popular pages on Facebook (for example, one variable is whether the participant liked The Beatles on Facebook). I include all pages liked by at least 10% of participants in baseline. In total, 255 covariates are used. I then use these covariates to predict the ITT effect among all participants in the baseline sample.

Appendix Table A.24 shows that the results predicted among the entire baseline sample are very similar to the results found among the subsamples of participants who completed the endline survey or installed the Chrome extension for at least two weeks. Based on the analysis of heterogeneity throughout this paper, the fact that the effects on opinions and attitudes are stable is not surprising, as the effects on the primary outcomes are generally homogeneous and the differences between participants in the baseline and endline surveys are not dramatic.

While these results are reassuring, two caveats should be noted. First, I control for many observable variables, but there could be unobservables differentiating the subsamples. Second, when

[76] I do not analyze the effect on posts shared because the access posts subsample already includes a large share of the baseline sample.

estimating heterogeneous effect in the extension subsample, I cannot control for one important difference between the groups - the device with which the survey was taken - since participants could only install the extension when taking the survey on a computer using Google Chrome.

C.6 Effects on Knowledge

While this paper focuses on persuasion and polarization, the survey included several questions related to political knowledge. The two primary measures of political knowledge are self-reported familiarity, measured according to whether participants reported hearing of news events and political figures, and accurate political knowledge, measured according to participants' answers to several true/false questions on recent events. For some questions, participants were expected to gain knowledge when assigned to the liberal outlet (heard of Michael Cohen, heard about the Stephon Clark shooting, believed the Russian government tried to influence the 2016 elections, believed a wall is not being built at the US-Mexico border) and for other measures, the conservative treatment was expected to have an effect (heard of Louis Farrakhan, heard about a controversial speech by Hillary Clinton in India, believed Trump is not a criminal target of the Mueller investigation, believed Trump's tax cuts would increase most people's income).

Appendix Table A.25 presents the effect of the treatments on knowledge for the four primary self-reported familiarity outcomes and the four primary accurate knowledge outcomes. The coefficients of interest are the effects of the liberal treatment on liberal outcomes and conservative treatment on conservative outcomes. The treatments seem to have little to no effect on the knowledge outcomes.

Appendix Table A.26 uses the browser extension data to show that the intervention affected news exposure. The regressions measure the effect of the treatments on the number of posts that appeared in the participants' social media feeds and referred to relevant topics.[77] For all four topics, the treatments had a significant effect in the expected direction when the relevant treatment is compared to the control group, and for three of the four topics, the effect is also significant when the treatments are compared to each other.[78]

The results presented in this section suggest that while the slant of one's social media feed can determine the news events an individual is exposed to on social media, that exposure does not necessarily affect their political awareness of topics. One possible explanation is that individuals consume news also outside their social media feed. In any case, this result should not be interpreted as definitive evidence of a null effect. Participants were asked questions about very specific issues, the range of possible answers was limited, and answers to true/false questions could be

[77] Posts are defined as referring to Michael Cohen, Louis Farrakhan, or the shooting of Stephon Clark if they include the expressions "michael cohen", "louis farrakhan" and "stephon clark," respectively. Posts refer to Hillary Clinton's speech in India suggesting that many white women voted for Trump since they took their voting cues from their husbands if they include the words "clinton," "vote," and either "india" or "husband."

[78] For both tables mentioned in this section, the results are similar when running the regressions only among participants who installed the extension for at least two weeks and completed the endline survey.

driven by motivated reasoning and not by participants' true beliefs. Furthermore, previous studies have shown that the effect of media on political knowledge is complex, and depends on the context and the issue covered (Schroeder and Stone, 2015).

C.7 Exposure to Posts From the Offered Pro- and Counter-Attitudinal Outlets

In this section, I provide more details on the decomposition exercise for the primary specification, analyze several alternative decompositions, and test whether there is a gap in exposure to pro and counter-attitudinal articles within outlets.

C.7.1 Decomposition Calculations

I include in this analysis participants in the pro- and counter-attitudinal treatments for which I can observe posts in the Facebook feed in the two weeks after the intervention and for whom at least one post is observed. Overall, the sample includes 521 participants in the pro-attitudinal treatment and 538 participants in the counter-attitudinal treatment.

I define the number of posts observed in the counter-attitudinal treatment as:

$$S_C * A_C * U_C$$

where S_C is the mean number of new subscriptions to the offered counter-attitudinal outlets. A_C is the share of posts from the subscribed counter-attitudinal outlets among all posts observed in the feed, and U_C is the total number of posts observed in the feed in the counter-attitudinal treatment. I define the number of posts observed in the pro-attitudinal treatment as:

$$S_P * A_P * U_P = (S_C + S_\Delta) * (A_C + A_\Delta) * (U_C + U_\Delta)$$

I then decompose the difference in exposure to four separate expressions as described in Equation 3. To calculate S_Δ and A_Δ, I use the following regressions:

$$TotalSub_i = S_\Delta ProTreat_i + \varepsilon_i$$

$$TotalPosts_i = T_\Delta ProTreat_i + X_i + \xi_i$$

where $TotalSub_i$ and $TotalPosts_i$ are the number of offered outlets the participant subscribed to and the total number of posts observed, respectively. These regressions are presented in Appendix Table A.27, columns (1) and (2). X_i controls for Facebook usage before the intervention to increase precision.

To calculate the effect of subscribing to a post on exposure, I pool the two groups of potential outlets such that for each participant there are two observations: one observation with the four potential pro-attitudinal outlets and one observation with the four potential counter-attitudinal outlets. I calculate the share of posts the participants observed from each group of outlets among the total number of posts from all sources the participant observed in the two weeks following the intervention. I only include posts shared directly by the outlet to isolate any effect of friends sharing specific posts. I use the share of posts as the outcome variable instead of the total number of posts since users may observe more posts from pro-attitudinal outlets due to increased Facebook usage, and I account for this effect separately. A_C and A_Δ are estimated using the following regression:

$$SharePosts_{ij} = A_C * Sub_{ij} + A_\Delta * Sub_{ij} \times Pro_{ij} + \delta * Pro_{ij} + \nu_{ij} \tag{4}$$

where $SharePosts_{ij}$ is the share of posts participant i observed from group j, Sub_{ij} is the number of outlets participant i subscribed to from group j. Pro_{ij} is whether the outlets in the group matched the consumer's ideology. I instrument for Sub_{ij} and $Sub_{ij} \times Pro_{ij}$ with $Offer_{ij}$ and $Offer_{ij} \times Pro_{ij}$, where $Offer_{ij}$ is whether participant i is was offered outlets from group j in the intervention. This regression is presented in column (3) of Appendix Table A.27. Conceptually, it can be easier to think of this regression as two separate regressions. One regression includes only the potential counter-attitudinal outlets, and measure the effect of subscribing to an outlet on exposure to the outlet (A_C). I exploit the fact that for some participants the counter-attitudinal outlets were offered and for others they were not offered. In a second regression, I repeat this exercise for the pro-attitudinal outlets. A_Δ is the difference between the coefficients.

C.7.2 Alternative Decompositions

Appendix Figure A.21 presents the decomposition exercise using several alternative estimations. The x-axis is the gap in daily exposure to posts from the pro- and counter-attitudinal outlets, in the two weeks following the intervention. Most of these specifications lead to similar results, although I am often underpowered to detect precise effects. The first row of the figure is the primary specification shown in Figure 8. The second row adds fixed effects for the potential outlets defined for each participant. This assures that the estimates are derived from comparing participants who could have been offered the same set of outlets. The rest of the decompositions are described below.

Exclude Unsubscriptions Participants in the counter-attitudinal treatment are more likely to unsubscribe from outlets. Therefore, they may observe fewer posts due to their direct decision, but since they initially subscribed to the outlet, this could be accounted for as an algorithmic effect. In the third row of Appendix Figure A.21, only outlet subscriptions lasting two weeks are defined as subscriptions (this estimation only includes participants for which I observe two weeks of subscriptions data). The results do not change substantially.

Exclude Suspected Ads In the primary decomposition, I assume that participants observe posts they subscribe to as determined by Facebook's algorithm. This typically holds for organic posts. However, participants also observe sponsored posts (ads) which are different in several important aspects. First, they can appear in a user's feed even if she did not subscribe to the outlet. Second, the placement of sponsored posts can be determined by the advertiser. For example, an outlet can decide to show posts to a subset of users who subscribed to its Facebook page. This means that part of the effect attributed to the algorithm may result from the behavior of advertisers.[79] When excluding suspected ads, the gap between exposure to pro- and counter-attitudinal outlets slightly decreases. This suggests that ads target users whose ideology matches the outlet they subscribe to. Still, even when ads are excluded, the gap between the two groups of outlets remains large and the decomposition does not change substantially.

Reweight Based on Compliance P is estimated using two IV estimators, and thus its causal interpretation relies on the assumption that there is no essential heterogeneity (Heckman et al., 2006). Otherwise, the difference between exposure in the pro- and counter-attitudinal treatments might be due to the combination of heterogeneity in the treatment effect and selection into compliance, and not due to different treatment effects. In the fifth row panel of Appendix Figure A.21, I re-weight the IV estimators, such that participants predicted to comply receive a lower weight. I first calculate the probabilities of compliance with the pro-attitudinal treatment and counter-attitudinal treatment, by regression compliance on the following covariates using a logit regression: age, female, self-reported ideology, party (dummy variables for Democrat, Republican and Independent), and the difference between the participant's feeling toward her party and the opposing party. I then predict the probability of compliance for each participant and define the participant's weight as the inverse of the predicted probability.

The panel shows that reweighting the compliers does not change the result substantially. The reweighted estimates measure the treatment effect under the conditional effect ignorability assumption (Angrist and Fernandez-Val, 2013; Aronow and Carnegie, 2013). This assumes that conditional on the covariate (the compliance score), subscribing to outlets has the same ATE for compliers on non-compliers. There could still be essential heterogeneity based on other variables differentiating the compliers, but at least this suggests that the result does not stem from differences in compliers and heterogeneous effects by ideology or baseline affective polarization, for example. The results are stable not because the effect is homogeneous, but rather because the compliers are not dramatically different from non-compliers in both treatments.

[79] Even with sponsored posts, the algorithm may still play an important role. For example, advertisers can target a broad array of users and pay for each click on a post. This creates an incentive for Facebook to place the posts among users who are likely to click them, and thus the incentives in determining where to place sponsored posts becomes similar to the incentives when placing organic posts.

Reweight to Match Population Demographics In the sixth row of the figure, I reweight the participants to match population means on the same set of variables mentioned in the previous section and using the entropy weighting procedure. Reweighting decreases the gap between the number of posts observed, largely due to a smaller effect of platforms algorithms. One possible explanation for the result is that when analyzing the results separately for conservatives and liberals, I find that the algorithm's tendency to increase exposure to matching news outlets is driven by the liberals in my sample, though I am underpowered to estimate these results precisely. This difference could be due to the ideology of participants or differences in the outlets offered.

Excluding Facebook Usage The effect on Facebook usage is only marginally significant. In the sixth row of Appendix Figure A.21, I assume that the exposure gap only stems from subscriptions and the platform algorithm, and exclude the usage dimension. For this decomposition, I change the calculation of A in equation 4, and instead of estimating the effect on the share of posts in the feed, I estimate the effect on the number of posts observed by participant i from outlets in group j.

Decomposition Over Time In the final two rows of Appendix Figure A.21, I decompose the gap in exposure for the first and second week after the intervention. I use the same estimate for subscriptions in both weeks, but calculate exposure to posts and Facebook usage according to each week's specific activity. The overall gap in the number of posts is greater in the first week, but this reflects the fact that participants were generally exposed to more posts from the offered outlets in the first week. The relative difference between pro- and counter-attitudinal posts is greater in the second week (approximately 140% more pro-attitudinal posts) compared to the first week (106%). The effect associated with subscriptions becomes smaller over time and the effect associated with the algorithm slightly increases. This suggests that Facebook's algorithm learns from participants' behavior that they prefer pro-attitudinal content. However, the effect of the algorithm is still strong in the first week suggesting that either the algorithm learns very quickly (e.g., based on engagement with the first posts from an offered outlet shown to a participant) or that the algorithm uses other baseline information (such as subscriptions to other outlets) to determine that participants are more interested in pro-attitudinal content.

C.7.3 Differential Exposure to Articles Within an Outlet

To estimate whether participants were exposed to news more likely to match their opinions within an outlet, I focus on the subset of articles that were shared on Facebook or Twitter by at least one member of Congress in January-November 2018. I define the slant of an article according to the mean DW-Nominate score of congressmembers who shared the article.[80] Using this measure, I

[80] The list of the Facebook pages of congressmembers is based on the Congress Members Project (https://github.com/unitedstates/congress-legislators). Based on this list, I collected all posts shared by congressmembers in 2018. The list of tweets shared by congressmembers is from the Tweets of Congress Project (https://github.com/alexlitel/congresstweets). The datasets were downloaded on December 2018.

find that in general conservative participants are exposed to more conservative articles on Facebook, even when controlling for the outlet. This is not surprising as a conservative is likely to have more conservative friends, who are likely to share more conservative articles within an outlet. However, when I focus only on posts shared by the eight potential outlets defined for each participant, I do not find any correlation between the slant of the articles and consumers' ideologies. This suggests that Facebook's algorithm does not lead to conservatives being supplied with more conservative articles, *within* the set of posts shared by an outlet. It also suggests that conservatives and liberals were exposed to similar content from the outlets they subscribed to in the intervention, conditional on posts from the outlet appearing in their feed.

D Interpretation

How should we interpret the fact that the intervention affected attitudes toward parties, while political opinions remained stable? In this section, I compare two frameworks explaining affective polarization and examine which is most consistent with the data.

Consider the following model: consumer i's prior on state k of the world is $\theta_{ik} \sim (\theta_{ik}^0, \frac{1}{h_{ik}})$, where θ_{ik}^0 is the consumer's initial belief and h_{ik} is the precision of the belief (the consumer's certainty). I extend classic media persuasion models by introducing the concept of affective polarization and assuming that a consumer's political opinion, γ_i, is a weighted average of K beliefs:

$$\gamma_i = \sum_{k \in \{1..K\}} w_{ik} \theta_{ik} \tag{5}$$

where $w_{ik} \in \{0, 1\}$ is the weight consumer i places on belief k when determining her political opinion. A weight can be thought of as the priority the consumer places on a specific belief. For example, a consumer's support for a climate bill can depend on two beliefs: the consumer's belief on whether the bill will decrease or increase emissions and the belief on whether the bill will increase or decrease electricity prices. A conservative may place a positive weight only on the effect on prices and a liberal may place a positive weight on the effect on emissions.[81] A political party uses the same framework and its opinion is a weighted average of various beliefs.

Outlet j receives signal s_{jk} on the state of the world: $s_{jk} \sim N(\theta_k^*, \frac{1}{h_{jk}})$, where θ_k^* is the true state of the world and h_{jk} is the precision of the signal received. Media outlets act as delegates for their con-

[81] In a 2019 Pew survey, 74% of Democrats stated that the environment should be a top priority for President Trump and Congress in 2019, compared to only 31% of Republicans. On the other hand, 79% of Republicans said the economy should be a top priority, compared to 64% of Democrats (the sample includes respondents leaning toward the Democratic and Republican parties). Pew Research Center January 2019 Political Survey.

As a clarifying example for the framework, I intentionally focus on a general topic–support for climate change policy. Some of the questions forming the political opinions index are on more specific topics, but the same logic holds. For example, the favorability of the March for Our Lives Movement could depend on participants' belief on whether banning certain weapons will decrease gun violence and their belief on whether the movement will prevent most gun owners from purchasing their preferred guns.

sumers by covering issues according to the weights their consumers place on them.[82] Therefore, pro-attitudinal outlets cover issues more when $w_{own} > w_{opposing}$ and counter-attitudinal outlets cover issues more when $w_{opposing} > w_{own}$, where w_{own} are the weights used by the individual's party and $w_{opposing}$ are the weights used by the opposing party. Indeed, Figure 2 suggests that there is substantial differentiation in the topics news outlets cover. Returning to the climate change example, data from the outlets offered in the experiment also demonstrates this differential coverage: for every post from a conservative outlet mentioning the word "environment," 2.74 posts mentioned the word "economy," while for liberal outlets, the ratio was 0.82.[83]

I assume that consumers exposed to a new outlet update their beliefs in the direction of the outlet. This type of movement is expected if media outlets are biased in their reporting and consumers are naive and do not completely take the bias into account (DellaVigna and Kaplan, 2007).[84]

A straightforward way to model affective polarization is to define attitudes as a linear function of the distance between the political opinion of party p and a benchmark for the "correct" opinion according to individual i:

$$A_{ip} = g(\gamma_p - \hat{\gamma}_{ip}) \tag{6}$$

where A_{ip} is the attitude of individual i toward party p, γ_p is the political opinion of party p and $\hat{\gamma}_{ip} = \phi(\theta_{i1}, ..., \theta_{ik}, w_{i1}, ..., w_{ik}, \theta_{p1}, ..., \theta_{pk}, w_{p1}, ..., w_{pk})$, is the benchmark opinion that individual i thinks party p should hold. I consider two benchmark opinions: either individuals use their own opinion as the benchmark or they determine the benchmark opinion based on their beliefs weighted by the weights party p places on the beliefs.

Affective polarization due to political distance: $A_{ip} = g(\gamma_p - \sum_k w_{ik}\theta_{ik})$

Consumers may determine their attitudes toward a party based solely on the distance between their opinion and the party's opinion. Without loss of generality, I will focus on the position of a liberal consumer toward the Republican party ($\gamma_i < \gamma_p$). When the individual's political opinion changes from γ_i^0 to γ_i^1, the following change is expected in her attitude toward party p:

$$\Delta A_{ip} = g(\gamma_p - \gamma_i^1) - g(\gamma_p - \gamma_i^0) = g(\sum_k w_{ik}(\theta_{ik}^0 - \theta_{ik}^1)) \tag{7}$$

[82] Delegation has long been suggested as an explanation for why consumers prefer like-minded news (Chan and Suen, 2008; Suen, 2004). Yuksel (2018) shows that specialization can also increase polarization by allowing voters to learn about issues they care about and thus respond less to party platforms.

[83] This calculation is based on the ratio between the number of times the words "economy" and "environment" appeared in the description of all posts shared by each outlet in February-November 2018. Duplicate posts with the same description were excluded.

[84] An alternative explanation for why consumers' posteriors move toward the opposing party when exposed to counter-attitudinal news is that individuals' priors tend to support their political opinion. In other words, liberals tend to have more liberal priors than the true state of the world and conservatives tend to have more conservative priors. When exposed to counter-attitudinal outlets, liberals and conservatives receive more signals on issues for which they have weak prior and their beliefs move toward the true state of the world.

According to this theory, increased affective polarization can be explained by ideological divergence (Rogowski and Sutherland, 2016), and an update in the consumer's beliefs should only affect attitudes toward a party through its effect on the consumer's political opinions. Returning to the climate bill example, a consumer would determine her attitude toward a political party based on the distance between her support for the climate bill and the party's support for the bill. This theory is not consistent with the experiment since attitudes changed without a corresponding change in political opinions.

Affective polarization due to unreasonable opinions: $A_{ip} = g(\gamma_p - \sum_k w_{pk}\theta_{ik})$

Alternatively, the attitude of a consumer toward a party may depend on whether the political opinion of a party is reasonable according to the party's weights. Hence, the benchmark opinion is the opinion the party would hold according to the consumer's beliefs regarding the state of the world, weighted by the weights party p places on those beliefs. In other words, affective polarization increases when consumers cannot rationalize the parties' political opinions and perceive that the party is not adhering to its own values.[85] The change in affective polarization following an update to the consumer's beliefs is:

$$\Delta A_i = g(\gamma_p - \sum_k w_{pk}\theta^1_{ik}) - g(\gamma_p - \sum_k w_{pk}\theta^0_{ik}) = g(\sum_k w_{pk}(\theta^0_{ik} - \theta^1_{ik})) \tag{8}$$

If the consumer and the party place the same weight on beliefs ($w_{pk} = w_{ik}$), there is no difference between the two theories. However, with heterogeneous weights, political opinions and affective polarization may be differentially affected. In the climate bill example, a liberal who believes the climate bill will mitigate emissions and *decrease* consumer prices will support the bill. The consumer will have a negative attitude toward a party opposing the bill since even if the party places a zero weight on decreasing emissions, it should still support the bill. If the liberal is exposed to conservative outlets and learns that the bill is more likely to increase prices, she may still support the bill since she places a positive weight only on mitigating emissions but will develop a less negative attitude toward a party that places a positive weight on consumer prices and thus opposes the bill.[86]

This theory is consistent with the results of the experiment if the consumers updated beliefs on

[85] Another way to interpret affective polarization according to this framework is that the consumer attributes malicious motives to the party. Since the consumer infers that the party should have a different political opinion according to its weights and the correct beliefs, she concludes that there is an additional unethical consideration determining the party's stance. For example, the consumer might assume that the party supports a policy because it is corrupt or because the policy will have negative implications for the party's opponents.

[86] Stone (2020) shows that affective polarization could increase due to limited strategic thinking or a false consensus bias. In the context of this experiment and theoretical framework, a false-consensus bias is similar to consumers having the wrong priors regarding the weights the opposing party places on beliefs. Exposure to counter-attitudinal news allows consumers to learn those weights and thus rationalize the opinions of the opposing party. I focus on beliefs regarding issues and not beliefs regarding the opposing party's weights because I suspect that weights are more likely to be common knowledge. However, both theories are consistent with the results of my experiment.

which they place zero weights, but at least one of the parties places positive weights.[87] This would result in consumers' political opinions remaining constant, but attitudes toward parties changing.[88]

To further test these theories, I analyze the effect of the experiment on participants' attitudes toward the opposing party. If affective polarization is simply a function of political distance, attitudes toward parties will be affected when consumer i updates beliefs on which she places positive weights (Equation 7). Therefore, attitudes toward both parties are more likely to be affected by pro-attitudinal outlets that cover these beliefs. On the contrary, if affective polarization is a function of unreasonable opinions, attitudes toward party p will be affected more by beliefs on which p places positive weights (Equation 8). As a result, pro-attitudinal outlets are more likely to affect attitudes toward one's own party, while counter-attitudinal outlets are more likely to affect attitudes toward the opposing party. Table A.16 shows that attitudes toward the opposing party are indeed more likely to be affected by exposure to counter-attitudinal outlets, consistent with the theory that affective polarization is due to perceived unreasonable opinions.

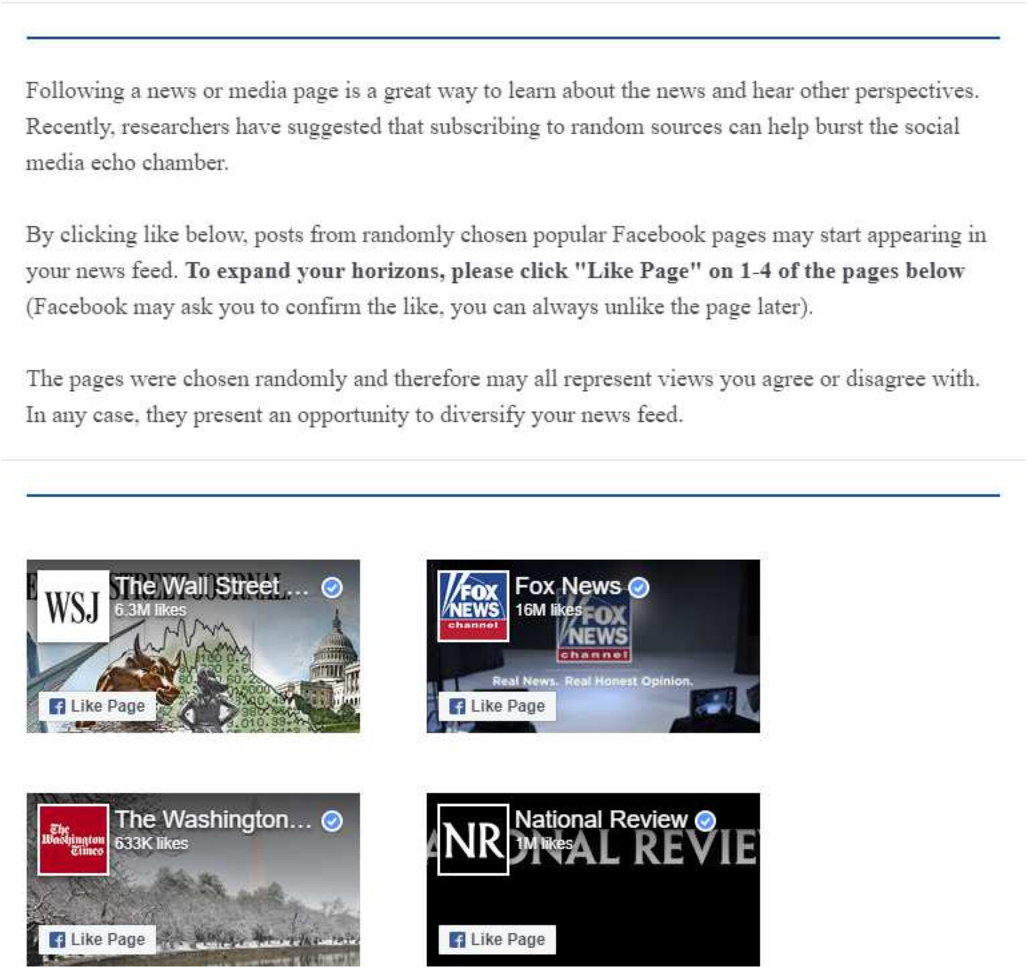
To conclude, there is still limited evidence on whether exposure to pro- and counter-attitudinal news has an effect on affective polarization, let alone an understanding of the channels explaining this effect. I present a parsimonious theory that is consistent with the results: consumers determine their attitudes toward a party based on the distance between the party's opinions and the opinion the party should hold according to the consumers' beliefs and the party's weights. While I provide evidence supporting the theory, there could be other explanations for the change in affective polarization, and more research is needed to pinpoint the precise mechanisms explaining how affective polarization evolves.

E Additional Figures and Tables

[87] It is plausible that as a result of the experiment consumers updated beliefs on which they place zeros weights since they are less likely to have been exposed to counter-attitudinal outlets covering these beliefs. Thus, they are expected to have weaker priors regarding those beliefs. Indeed, Appendix Figure A.12 shows that participants assigned to the counter-attitudinal treatment were more likely to say that they modified their views in the past two months because of something they saw on social media, compared to participants assigned to the pro-attitudinal treatment.

[88] The stability of political opinions relies on a strong assumption that consumers place zero weights on some beliefs or that they determine their political opinions based on lexicographic orderings of beliefs. This assumption is plausible in certain cases. For example, individuals who do not believe climate change is happening may place a zero weight on whether a climate bill decreases greenhouse gas emissions. More importantly, the logic behind the theory still holds if consumers place a positive but small weight on beliefs. In that case, we would expect political opinions to be slightly affected when those beliefs change, but the effect could still be much smaller than any change in affective polarization (indeed, the point estimate of the effect of the treatments on political opinions is positive, but economically very small).

Figure A.1: Example for the Conservative Treatment Intervention



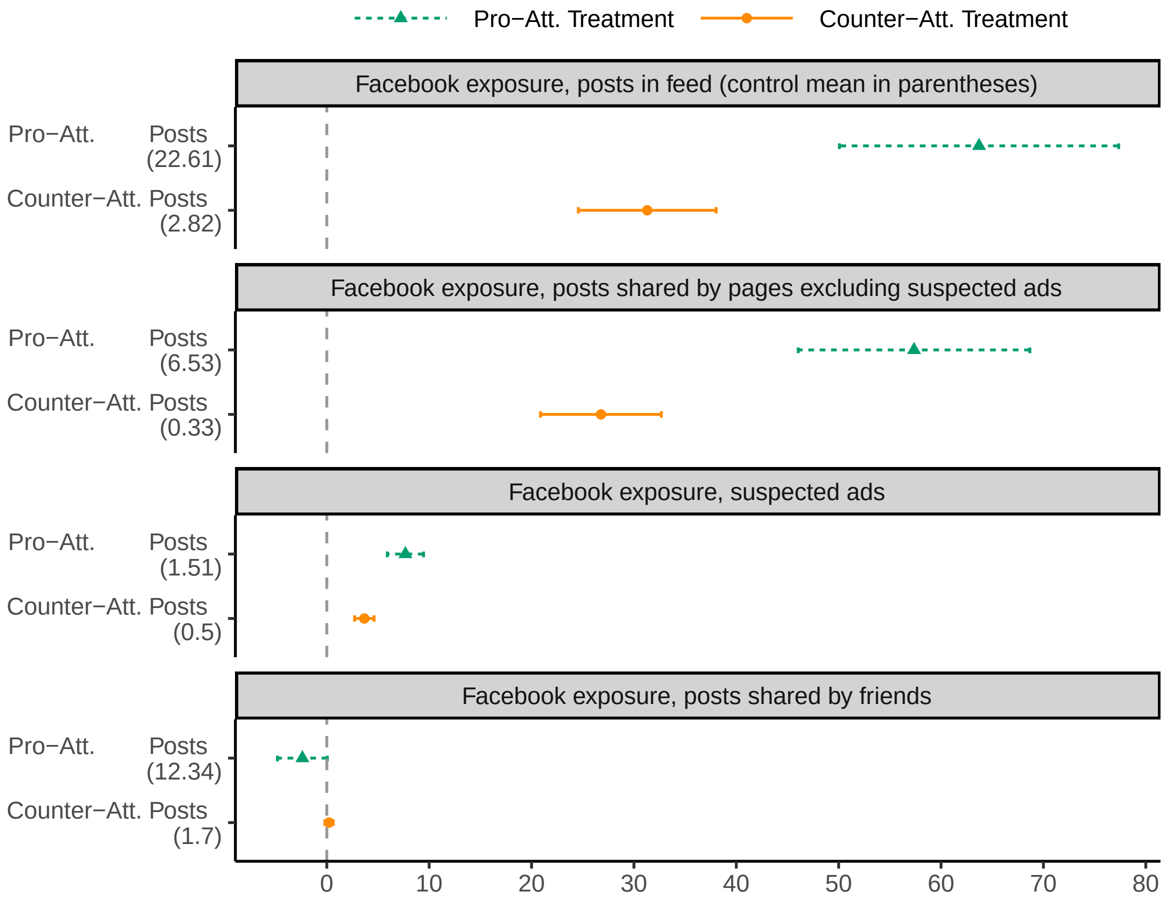
Following a news or media page is a great way to learn about the news and hear other perspectives. Recently, researchers have suggested that subscribing to random sources can help burst the social media echo chamber.

By clicking like below, posts from randomly chosen popular Facebook pages may start appearing in your news feed. **To expand your horizons, please click "Like Page" on 1-4 of the pages below** (Facebook may ask you to confirm the like, you can always unlike the page later).

The pages were chosen randomly and therefore may all represent views you agree or disagree with. In any case, they present an opportunity to diversify your news feed.

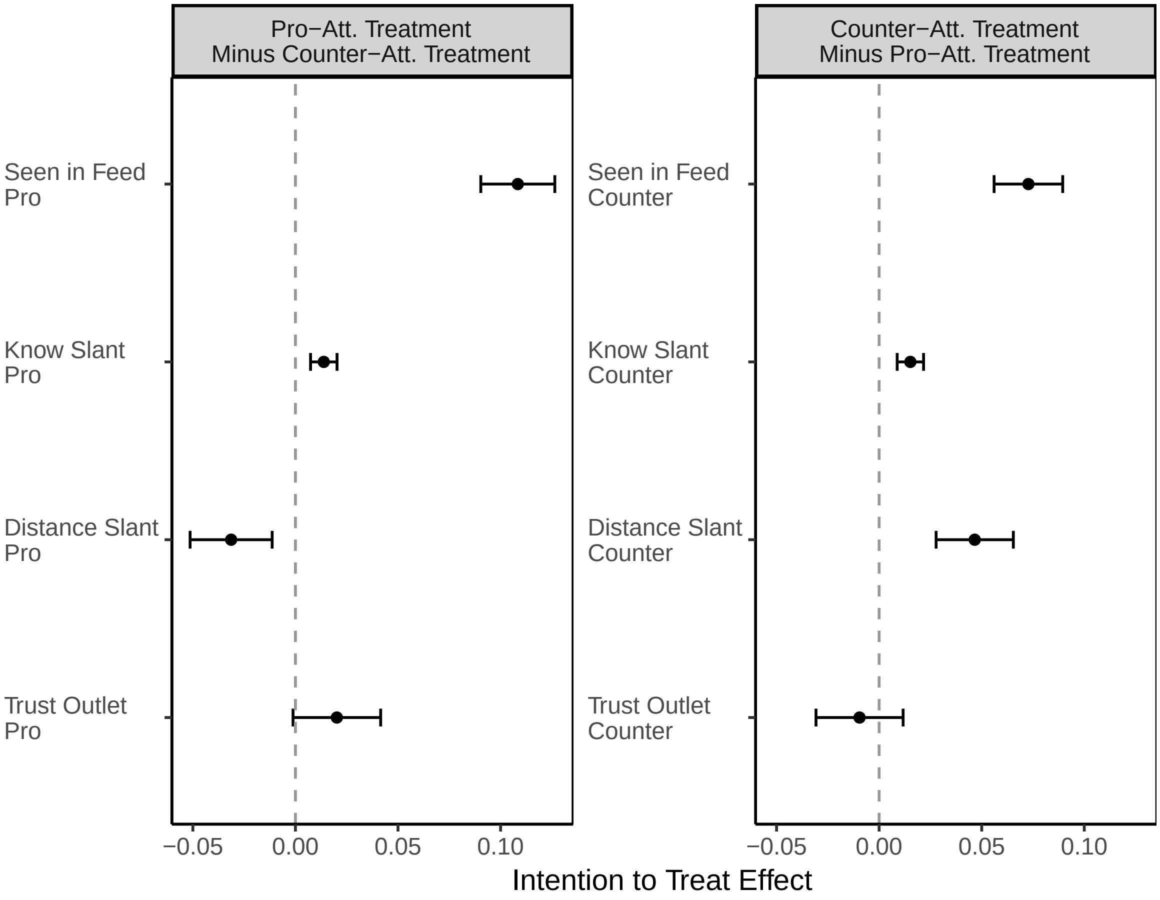
This figure shows the survey page asking participants to subscribe to four conservative outlets. Participants randomly assigned to the conservative treatment, who have not already subscribed to the four primary outlets, were shown a page similar to this figure. The image in the background of each outlet is dynamically updated according to the outlet's Facebook page, and the order of the outlets was determined randomly.

Figure A.2: Effect of the Pro- and Counter-Attitudinal Treatments on Exposure to the Potential News Sites, by Type of Post



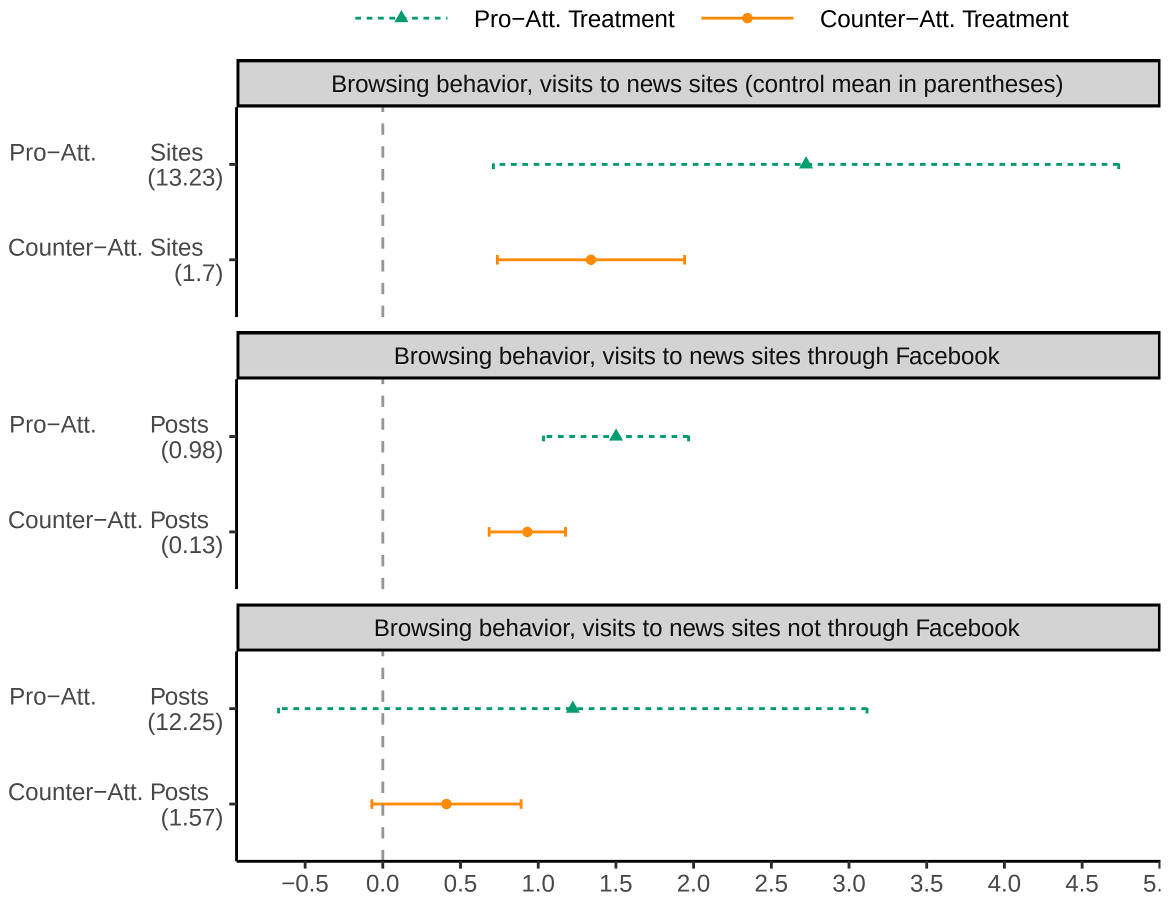
This figure shows the effect of the pro-attitudinal and counter-attitudinal treatments on total exposure to the potential outlets in the two weeks following the intervention. The first panel shows total exposure and is identical to the second panel in Figure 4. The second panel shows the effect on posts shared by Facebook pages organically. This includes all posts shared by the potential outlets, or other Facebook pages referring to the potential outlets, besides posts which are likely to be sponsored (ads). The third panel shows the effect on exposure to suspected ads related to the outlets. The fourth panel shows the effect on posts shared by Facebook friends. Appendix Section A.5 explains how ads were identified. Error bars reflect 90 percent confidence intervals.

Figure A.3: Effects on Survey Responses Related to the Potential Outlets



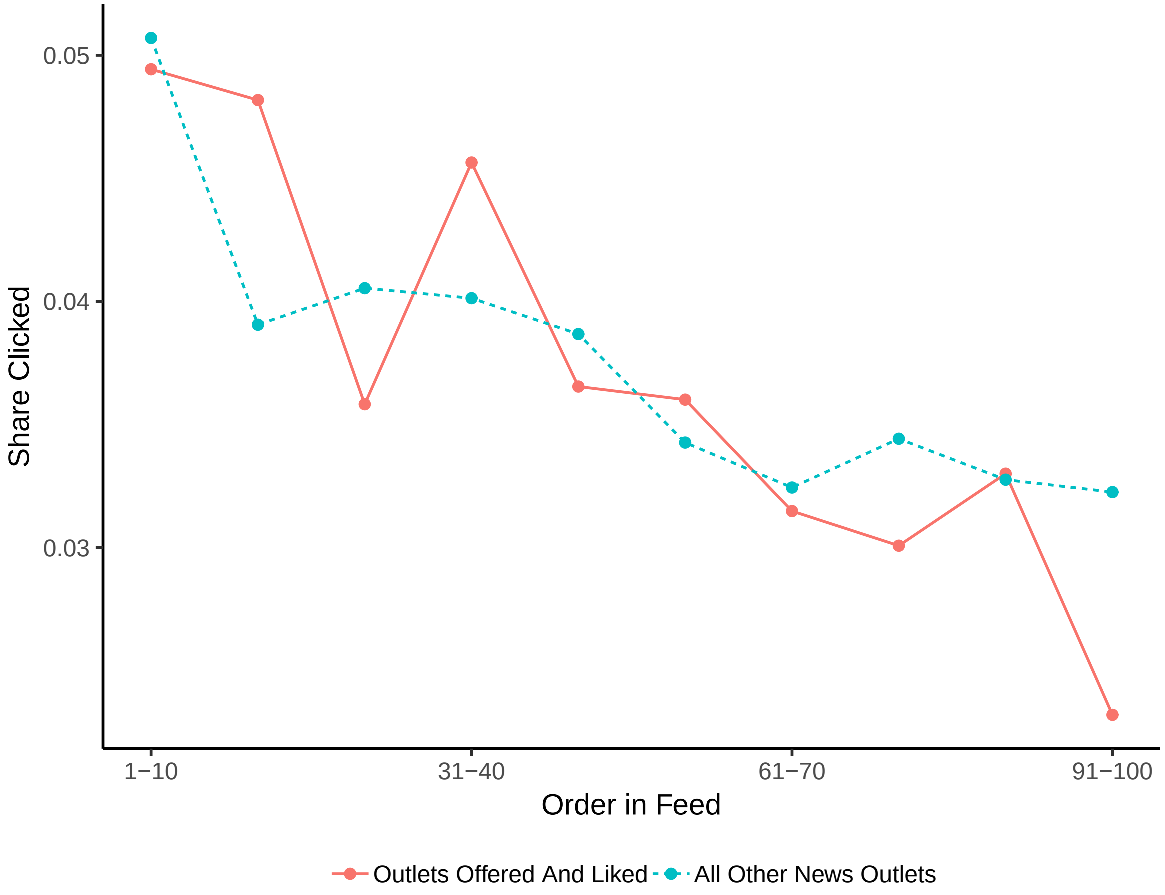
This figure shows the effect of the experiment on attitudes toward the potential outlets. Each row represents a regression pooling the opinions of participants in the endline survey on the eight potential outlets defined for each participant. The regressions control for the standard covariates: ideology, party, 2016 candidate supported, ideological leaning, age, age squared, gender, and the outcome in baseline when it exists. In addition, the regressions control for outlet fixed effects and for the set of potential outlets defined for each participant. *Seen in Feed* is whether the participant reported seeing news from the outlets in their Facebook feed over the past week more than five times (3), 3-5 times (2), 1-2 times (1), or reported seeing no posts (0). *Know Slant* is whether the participants did not mark "do not know" when asked what is the outlet's slant. *Distance Slant* is the difference between the participant's baseline ideology and the perceived ideology of the outlet. *Trust Outlet* is whether the participant perceived the outlet as very trustworthy (2), trustworthy (1), not trustworthy nor untrustworthy (0), untrustworthy (-1), or very untrustworthy (-2). Non-binary outcomes are standardized by subtracting the control group mean and dividing by the control group standard deviation. The left panel shows the effects of the pro-attitudinal treatment on the pro-attitudinal outlets (the counter-attitudinal treatment is the reference group). The right panel shows the effects on the counter-attitudinal treatment on counter-attitudinal outlets. Standard errors clustered at the individual level. Error bars reflect 90 percent confidence intervals.

Figure A.4: Effects of the Pro- and Counter-Attitudinal Treatments on News Sites Visited, by Source



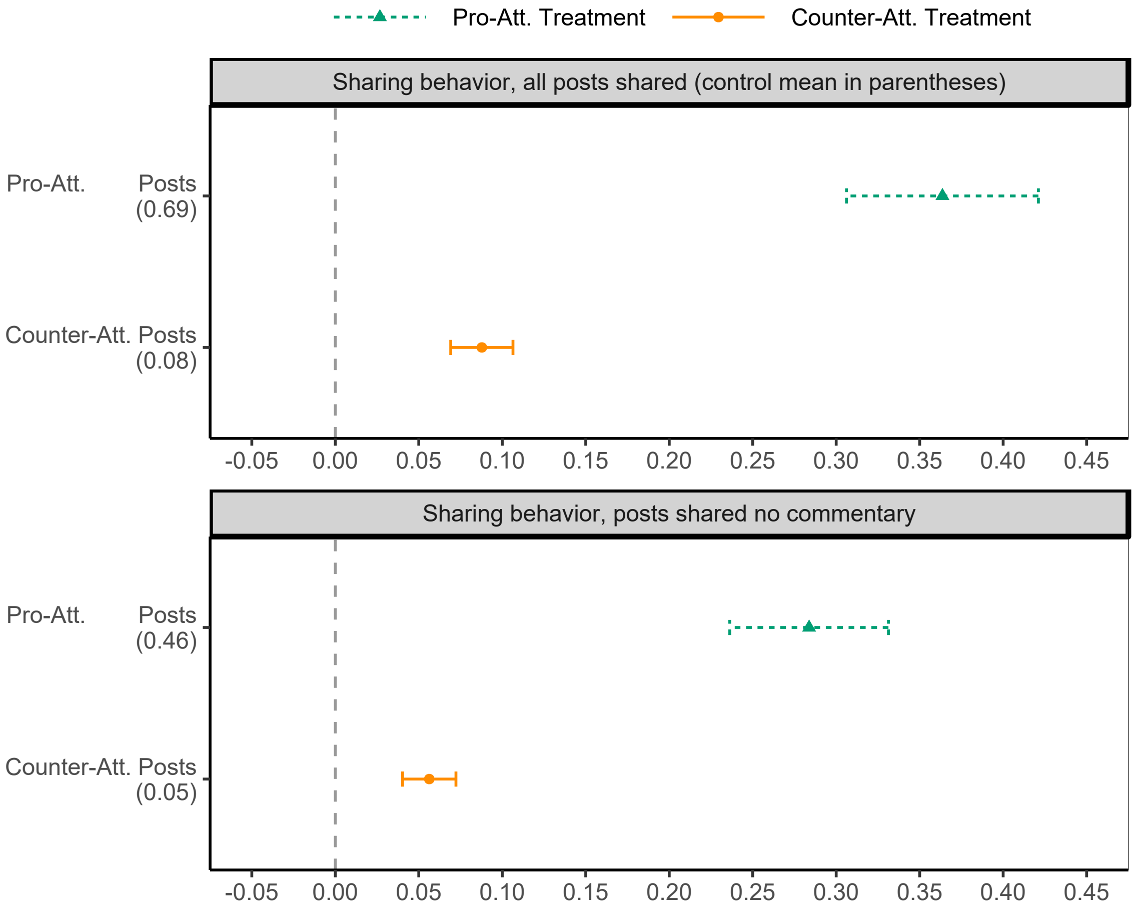
This figure shows the effect of the pro- and counter-attitudinal treatments on total visits to the potential outlets' websites in the two weeks following the intervention. The first panel shows total visits and is identical to the third panel in Figure 4. The second panel shows the effect on visits to websites that could be matched with a URL appearing in a Facebook post. The third panel shows the effect on all other visits. Appendix Section A.4 explains how posts were matched with visits to news sites. Error bars reflect 90 percent confidence intervals.

Figure A.5: Share of Links Visited by Order in Feed



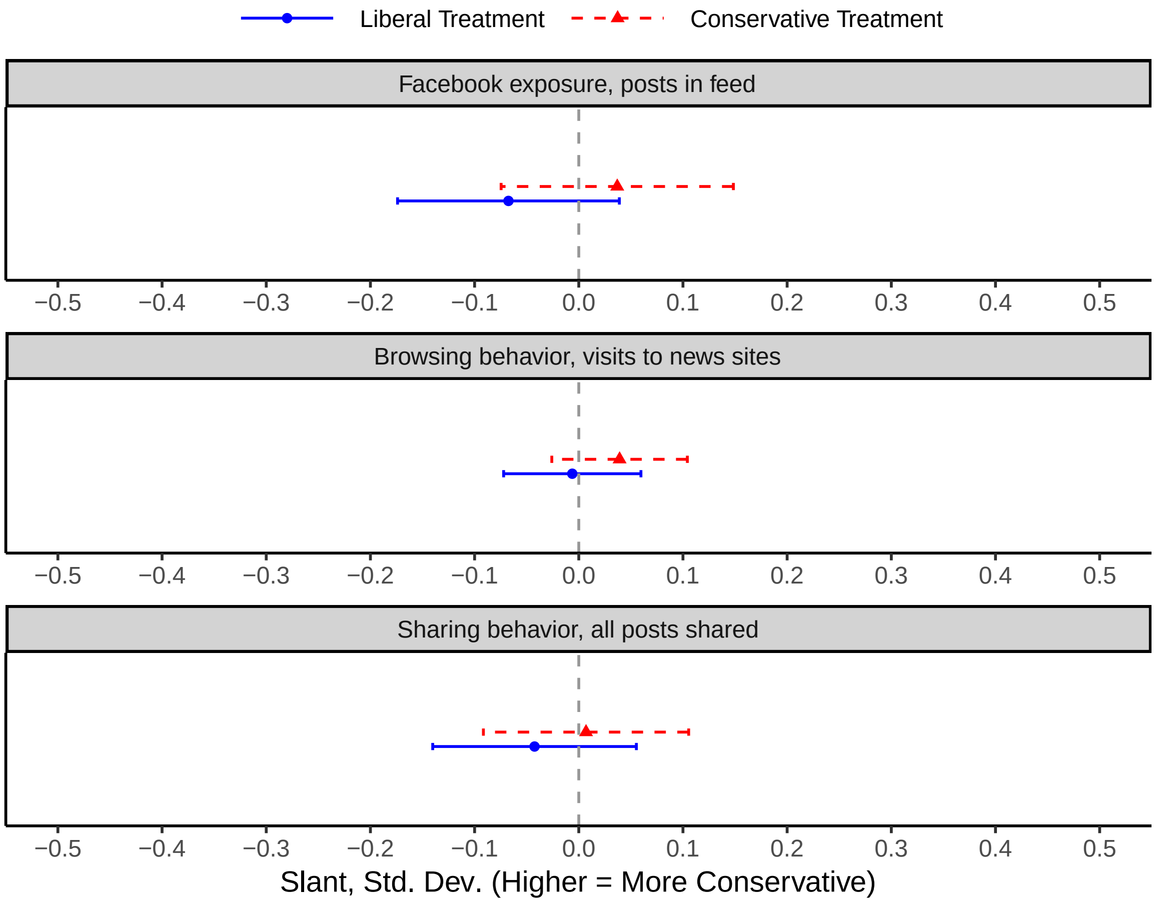
This figure shows the share of links clicked in posts from the pages of leading news outlets, excluding suspected ads. To determine the order of posts, a Facebook feed session is defined to begin when a participant views a post on Facebook at least 30 minutes after viewing a previous post. To smooth the results, posts are grouped into groups of ten based on their order. Appendix A.4 explains how posts were matched with visits to news sites and Appendix A.5 explains how suspected ads were identified.

Figure A.6: Effects of the Pro- and Counter-Attitudinal Treatments on Number of Posts Shared, Access Posts Subsample



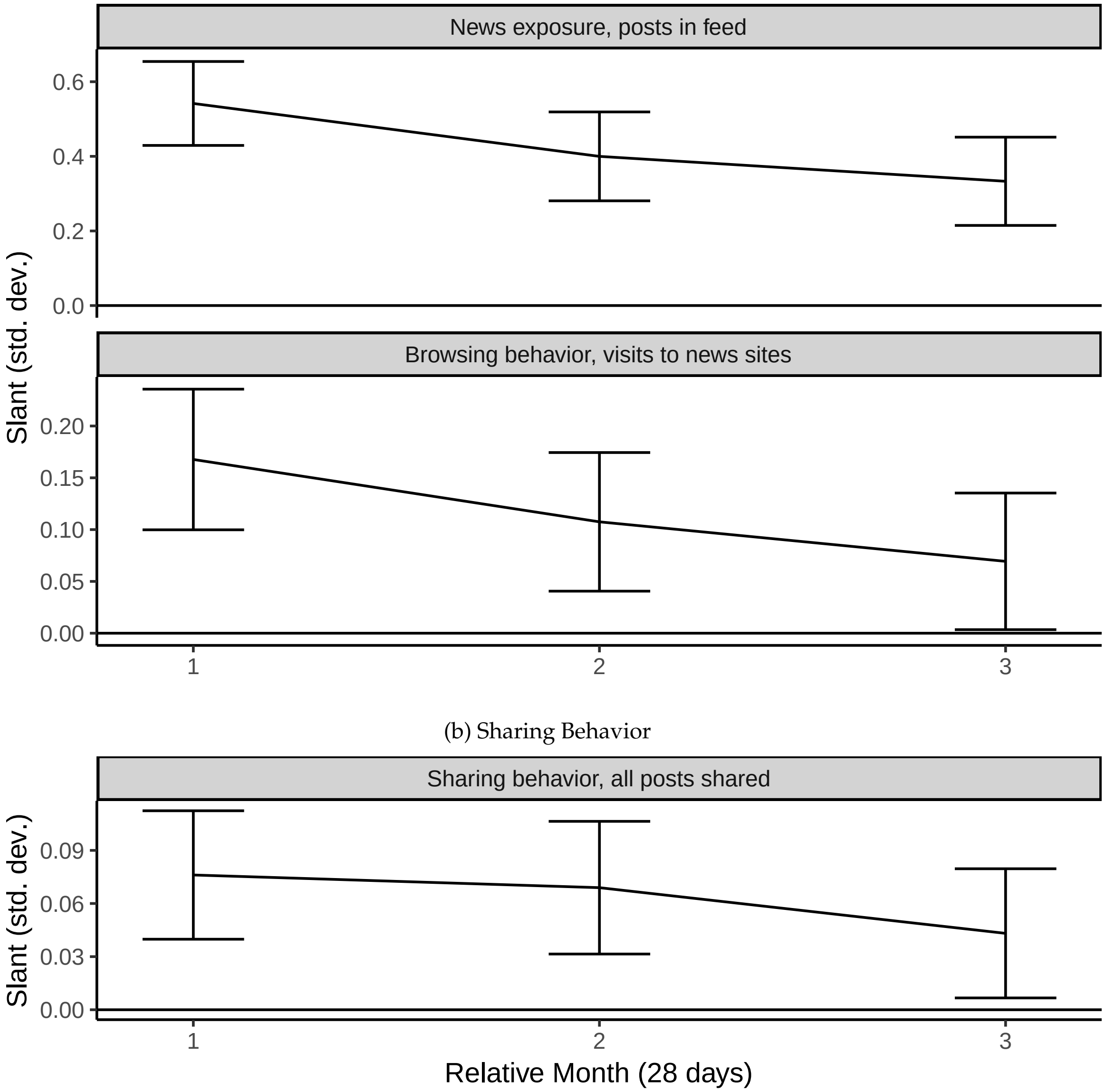
This figure shows the effect of the pro- and counter-attitudinal treatments on the number of posts participants shared from the four potential pro-attitudinal outlets and four potential counter-attitudinal outlets in the two weeks following the intervention. The first panel includes all posts and the second panel includes only posts that were shared without any commentary by the participant. The regressions control for the outcome measure in baseline. The data is from the access posts subsample: 33,532 participants with a liberal or conservative ideological leaning who provided access to their posts for at least two weeks following the intervention. Error bars reflect 90 percent confidence intervals.

Figure A.7: Effect of the Liberal and Conservative Treatments on Slant, Excluding each Participant's Eight Potential Experimental Outlets



This figure shows the effect of the liberal and conservative treatments on the mean slant, in standard deviations, of all news participants engaged with, excluding the four potential liberal outlets and the four potential conservative outlets defined for each participant. Each row in the figure is estimated by regressing engagement with the four potential conservative outlets or four potential liberal outlets on treatment assignment. The regressions control for the outcome in baseline if it exists. The sample includes 1,699 participants who installed the extension and provided permissions to access their posts for at least two weeks following the intervention. Error bars reflect 90 percent confidence intervals.

Figure A.8: Effects of the Conservative Treatment on Mean Slant by Month, Compared to Liberal Treatment

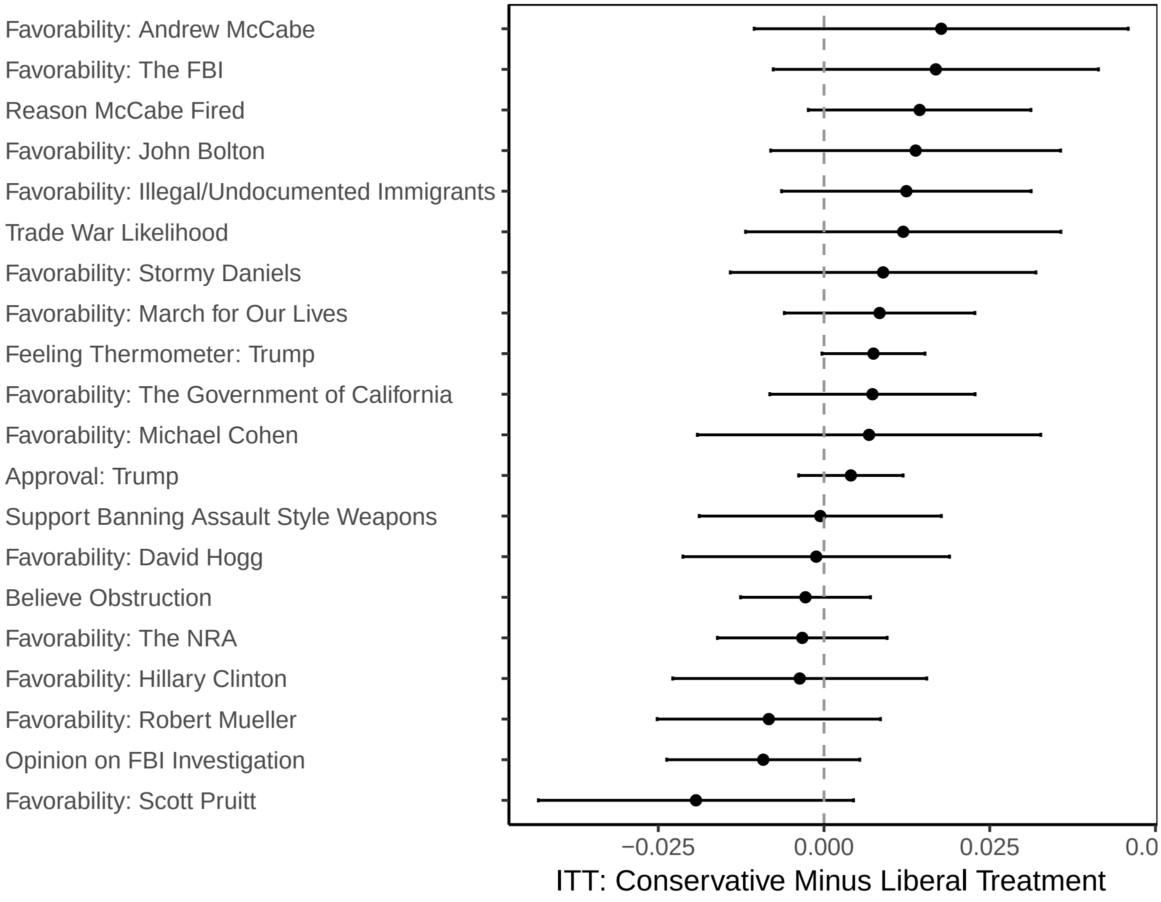


(a) News Exposure and Browsing Behavior

(b) Sharing Behavior

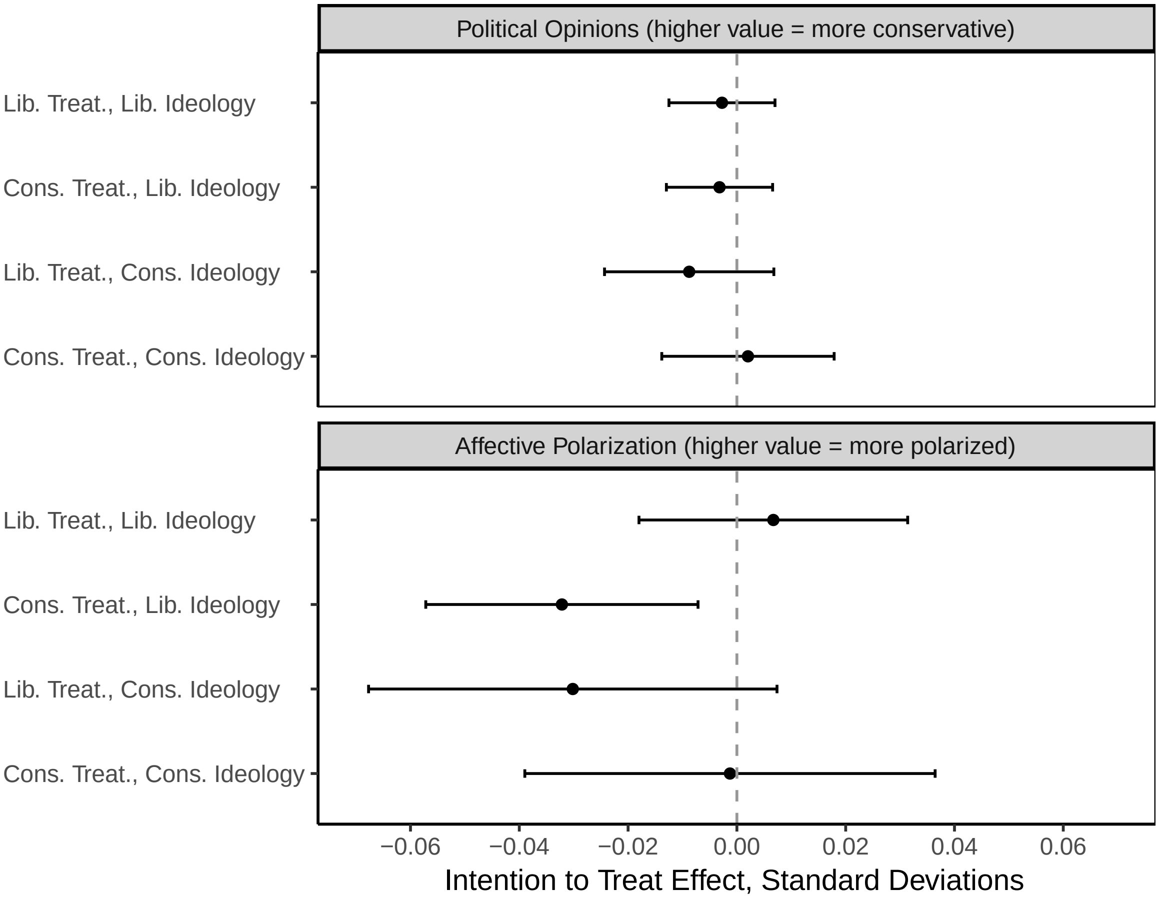
These figures show the difference between the effect of the liberal and conservative treatments on the mean slant over time. Each panel presents a series of regressions, where the dependent variable is the slant of outlets in a specific month. In the x-axis, relative month 1 is defined as 28 days immediately following the intervention. In sub-figure (a), the data is based on 1,351 participants who kept the extension installed for at least 84 days following the intervention. In sub-figure (b), the data is based on 9,932 participants who provided access to posts they shared for at least 84 days following the intervention. The regressions control for the outcome in baseline, if it exists. Error bars reflect 90 percent confidence intervals.

Figure A.9: Effect on Components of the Political Opinion Index



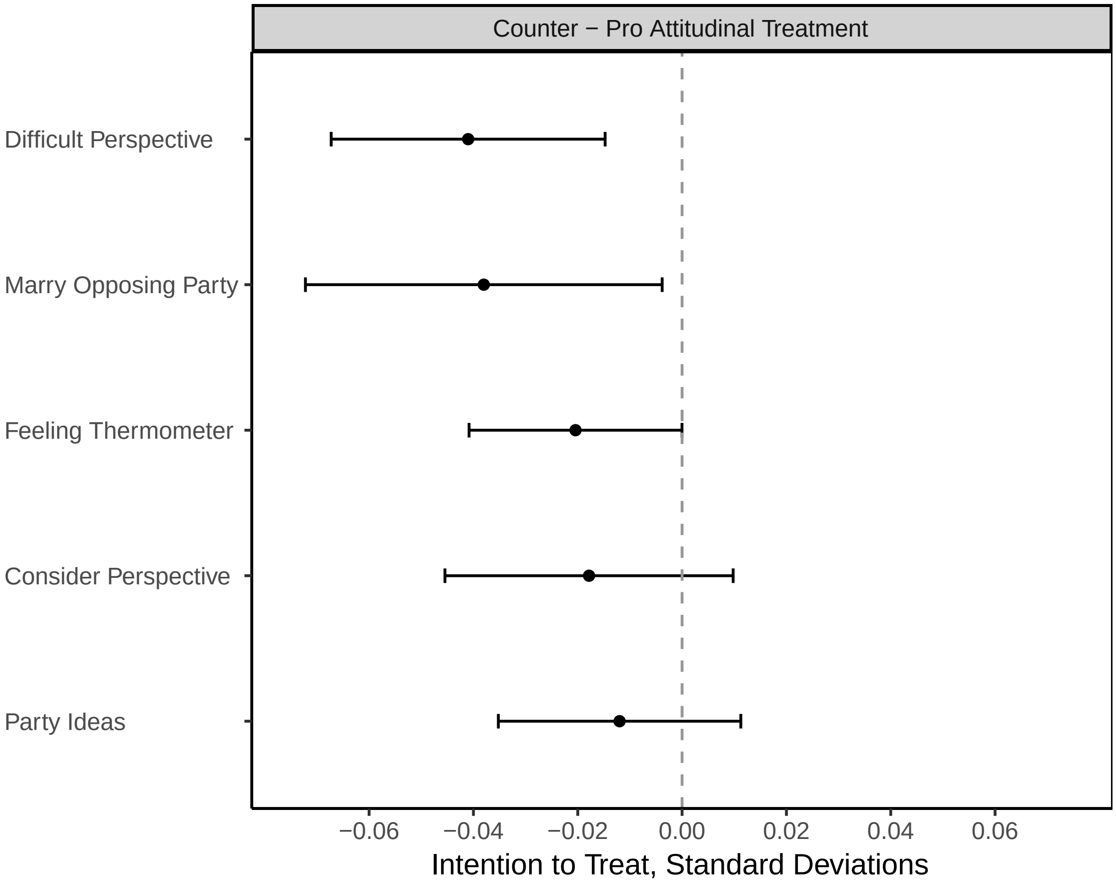
This figure shows the effect of the conservative treatment, compared to the liberal treatment on outcomes composing the political opinions index. Each row represents a separate regression as specified in Section 2.5. Outcomes are defined such that a higher value is associated with a more conservative opinion and then standardized with respect to the control group. *Favorability* outcomes are based on questions asking participants whether they have a very favorable, favorable, unfavorable, or very unfavorable opinion on specific individuals or organizations. *Approval: Trump* is whether participants strongly approve, somewhat approve, somewhat disapprove, or strongly disapprove of the job Donald Trump is doing as President. *Feeling Thermometer: Trump* is feeling toward Trump on a 0-100 degrees scale. *Believe Obstruction* is whether participants believed that President Trump has attempted to derail or obstruct the investigation into the Russian interference in the 2016 election. *Opinion on FBI Investigation* is whether participants think the FBI investigation into Trump campaign officials' contacts with Russian government officials is a serious attempt to find out what really happened, a politically-motivated attempt to embarrass Donald Trump or equally-motivated by both of these. *Reason McCabe Fired* is whether participants believe McCabe was fired because of improper actions while serving as Deputy Director of the FBI, as a way to damage McCabe's credibility in any evidence he might give to the Robert Mueller investigation, or as an act of revenge (multiple choice question). *Trade War Likelihood* is whether participants believe it is very likely, somewhat likely, somewhat unlikely, or very unlikely that a trade war will develop between the United States and foreign countries in the next year. *Support Banning Assault Style Weapons* is whether participants strongly support, support, oppose, or strongly oppose banning assault-style weapons. Error bars reflect 90 percent confidence intervals.

Figure A.10: Effect of the Treatments on Political Opinions, by Ideological Leaning



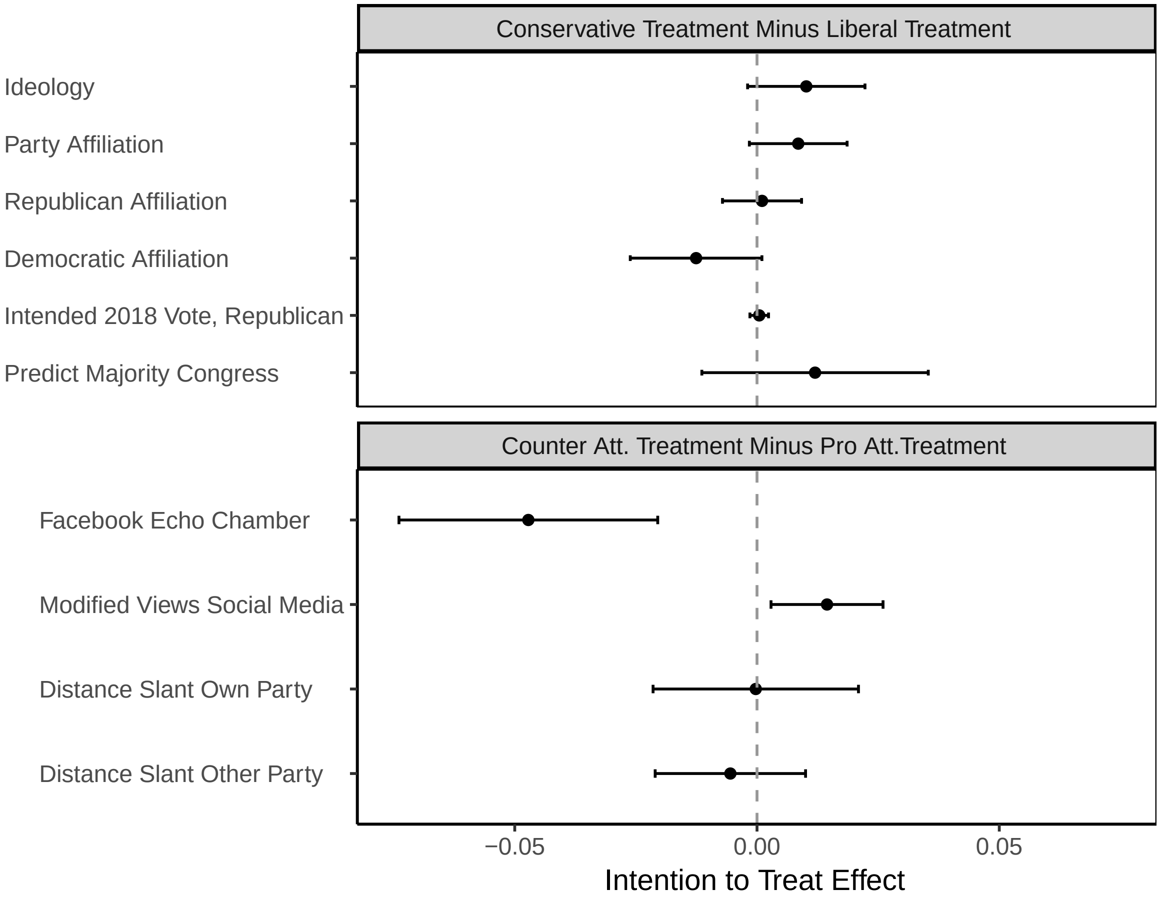
This figure shows the effect of the treatments among ideological subgroups based on the following regression: $Y_i = \beta_1 T_i^L I_i^L + \beta_2 T_i^L I_i^C + \beta_3 T_i^C I_i^L + \beta_4 T_i^C I_i^C + I_\mathrm{i} + \alpha X_i + \varepsilon_i$
where: T_i^C, T_i^L are binary indicators for the conservative and liberal treatments and I_i^C,I_i^L are binary indicators for whether the participant's ideological leaning is conservative or liberal. The reference group is the control group. The controls and the definition of ideological leaning are specified in Section 2.5. In the first panel, the x-axis is the ITT effect on the political opinions index, where a higher value is a more conservative outcome. In the second panel, the x-axis is the ITT effect on the affective polarization index, where a higher value is a more polarized outcome. Error bars reflect 90 percent confidence intervals.

Figure A.11: Effect of the Treatments on Components of the Affective Polarization Index



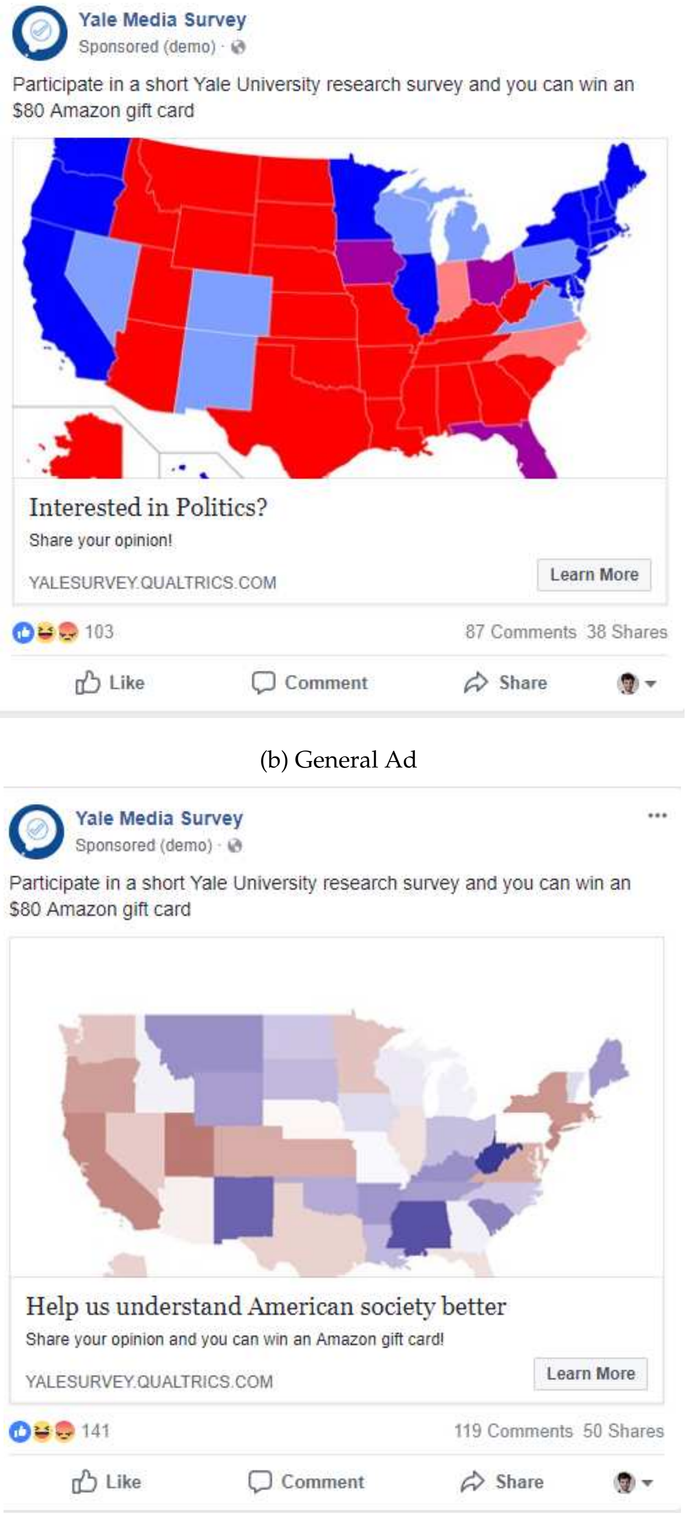
This figure shows the effect of the counter-attitudinal treatment on the measures composing the affective polarization index, compared to the pro-attitudinal treatment. Each row presents the result of a regression estimating the effect of the treatment on one dependent variable where a higher value is associated with a more polarized outcome. *Difficult Perspective* and *Consider Perspective* measure political empathy. The former is the difference in how difficult it is to see things from each party's point of view, and the latter is the difference in how important it is to consider the perspective of each party. *Marry Opposing Party* is how participants would feel if their son/daughter married someone from the opposing party. *Feeling Thermometer* is the difference in how warm participants feel toward each party. *Party Ideas* is the difference in how many good ideas each party has. The outcomes are described in more detail in Section 2.4.2 and the regressions are specified in Section 2.5. Error bars reflect 90 percent confidence intervals.

Figure A.12: Effect of the Treatments on Additional Survey Outcomes



This figure shows the effect of the experiment on additional endline survey outcomes. *Ideology* is self-reported on a 7-point scale. *Party Affiliation* is the party the participant identifies with on a 7-point scale from strong conservative to strong liberal. *Republican/Democrat Affiliation* is whether the participant is a strong Republican/Democrat (3), is a Republican/Democrat (2), leans toward the Republican/Democratic party (1), or does not identify with both parties (0). *Intended 2018 Vote* is whether the participant intends to vote for the Republican Party candidate (1) or the Democratic Party candidate (0) in her district if the election was held the day the survey was taken. *Predict Majority Congress* is the party the participant's predicts will hold the majority of seats in Congress after the 2018 vote: Republican Party (1) not sure (0), or the Democratic Party (-1). *Facebook Echo Chamber* is whether opinions seen about government and politics on Facebook are in line with participants' views always or nearly all the time (3), most of the time (2), some of the time (1), not too often (0). *Modified Views Social Media* is whether the participant modified her views in the past two months about a political or social issue because of something she saw on social media. *Distance Slant* is the difference between the participant's baseline ideology and the perceived ideology of a party. Non-binary outcomes are standardized by subtracting the control group mean and dividing by its standard deviation. In addition to the standard controls (Section 2.5), the regressions control for baseline outcomes when they exist. Error bars reflect 90 percent confidence intervals.

Figure A.13: Recruitment Ads



(a) Political Ad

Yale Media Survey

Sponsored (demo) ·

Participate in a short Yale University research survey and you can win an $80 Amazon gift card

Interested in Politics?

Share your opinion!

YALESURVEY.QUALTRICS.COM

Learn More

103

87 Comments 38 Shares

Like Comment Share

(b) General Ad

Yale Media Survey

Sponsored (demo) ·

Participate in a short Yale University research survey and you can win an $80 Amazon gift card

Help us understand American society better

Share your opinion and you can win an Amazon gift card!

YALESURVEY.QUALTRICS.COM

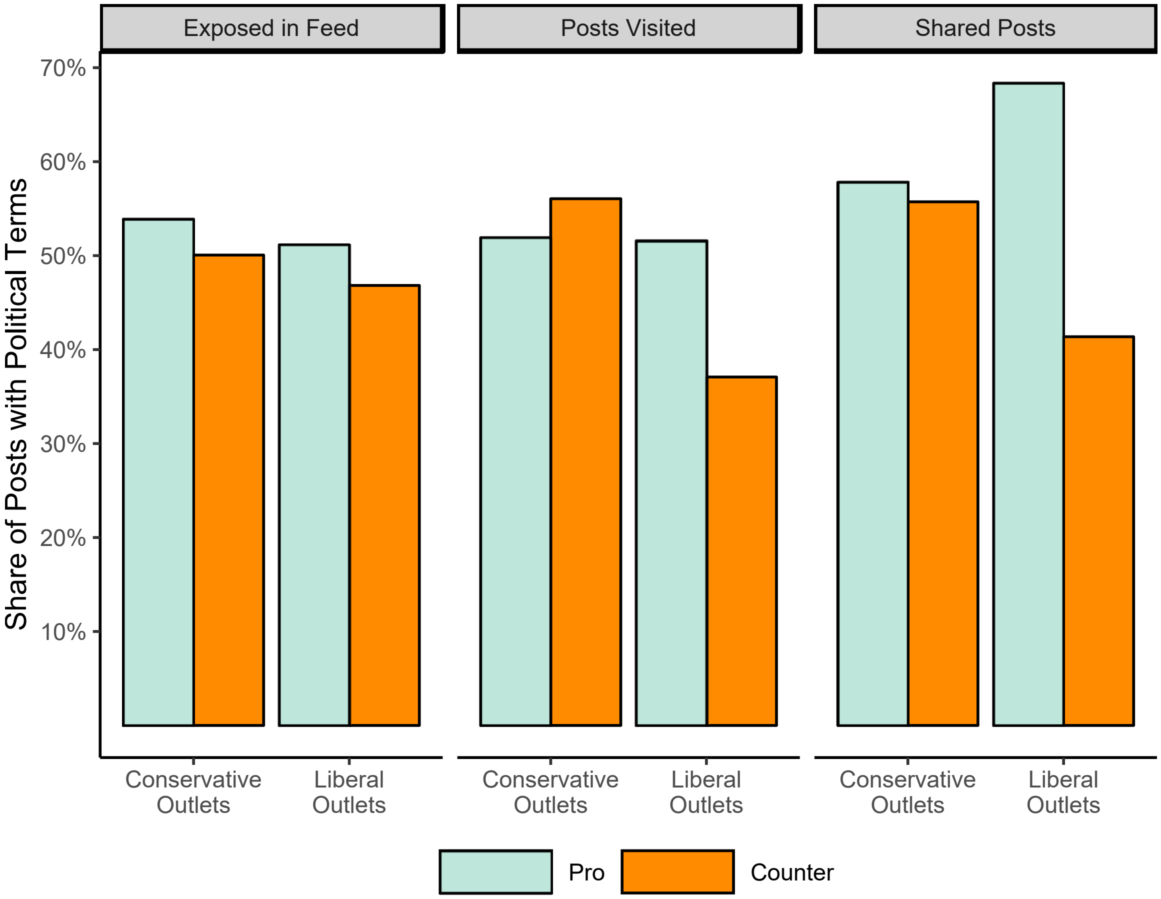
Learn More

141

119 Comments 50 Shares

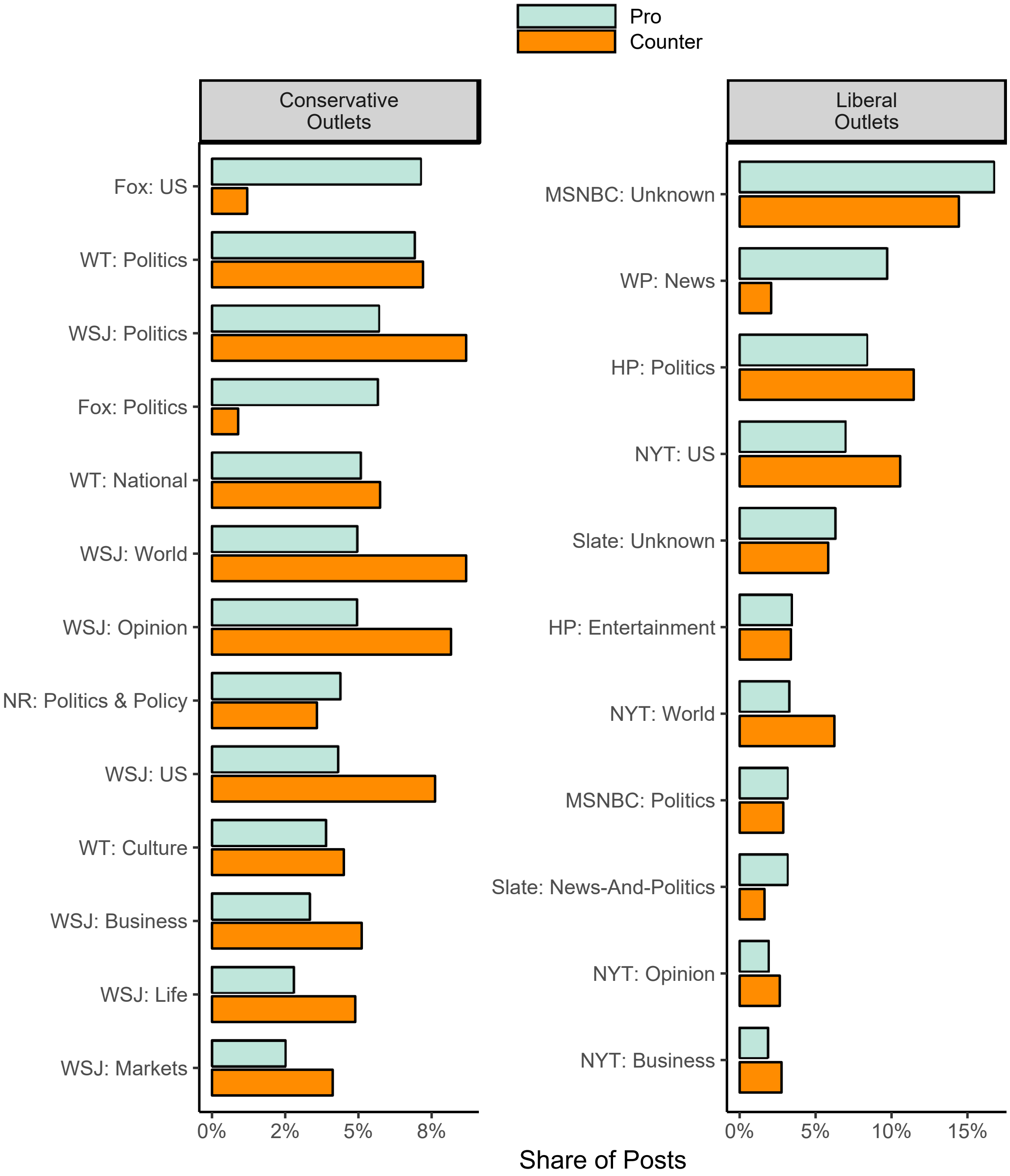
Like Comment Share

Figure A.14: Share of Posts Mentioning Political Terms



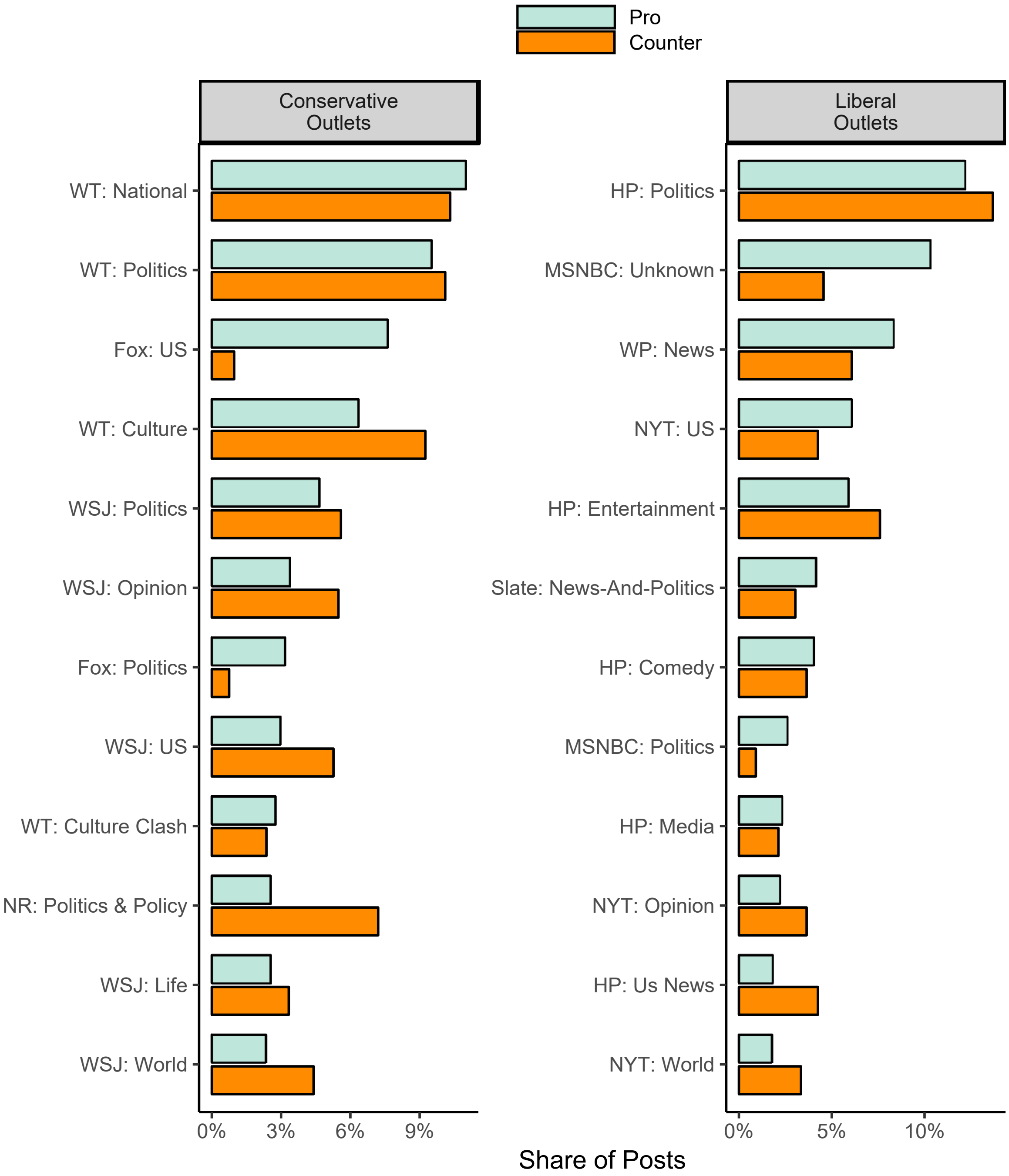
This figure shows the share of posts mentioning political terms mentioned in posts from outlets participants subscribed to. Posts are defined as political if they contain the following terms: ar 15, biden, bolton, carson, clinton, congress, conservative, daca, democrat, devos, dnc, elect, gop, gun control, gun law, gun right, immigration, kushner, liberal, manafort, mass shooting, mccabe, mcconnell, michael cohen, nra, obama, parkland, pelosi, pence, politic, pruitt, republican, sanctuary city, sanctuary state, school shooting, senate, tax cut, the left, the right, tillerson, trump, vote, walkout, white house. Posts from the pages of the four primary and first alternative outlets (excluding suspected ads) in the first eight weeks following the intervention are included. Political terms are searched for in the post's description, the URL, and the commentary left by the participants for shared posts.

Figure A.15: Links in Posts Observed in the Feed, by Outlet and Section



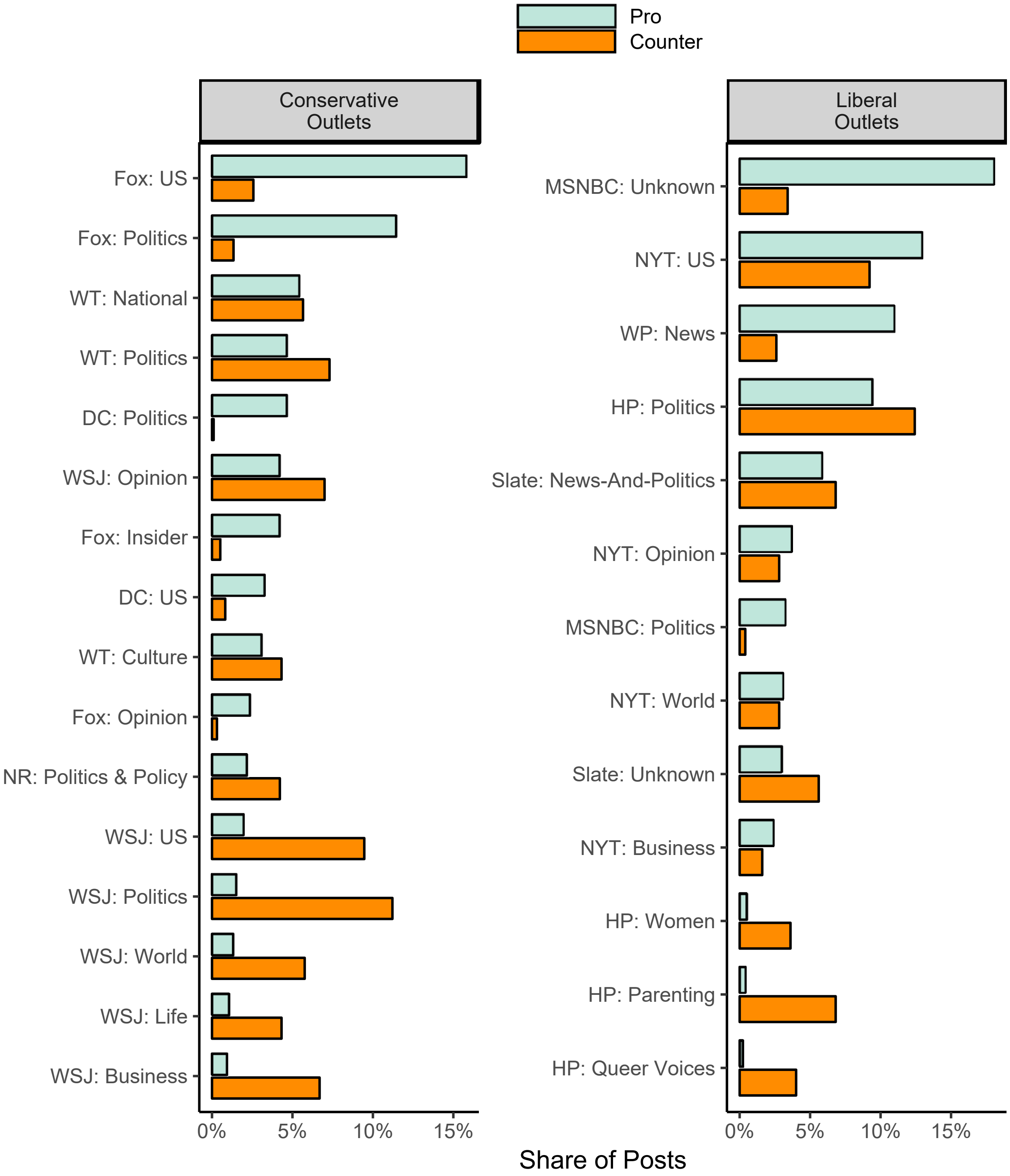
This figure shows the most common outlets and sections of links participants were exposed to in their feed. Data is from the eight weeks following the intervention. Posts from the pages of the four primary and first alternative outlets (excluding suspected ads) are included: Daily Caller (DC), Fox News (Fox), Huffington Post (HP), MSNBC, Slate, National Review (NR), New York Times (NYT), Wall Street Journal (WSJ), Washington Post (WP), and Washington Times (WT).

Figure A.16: Links Visited by Participants, by Outlet and Section



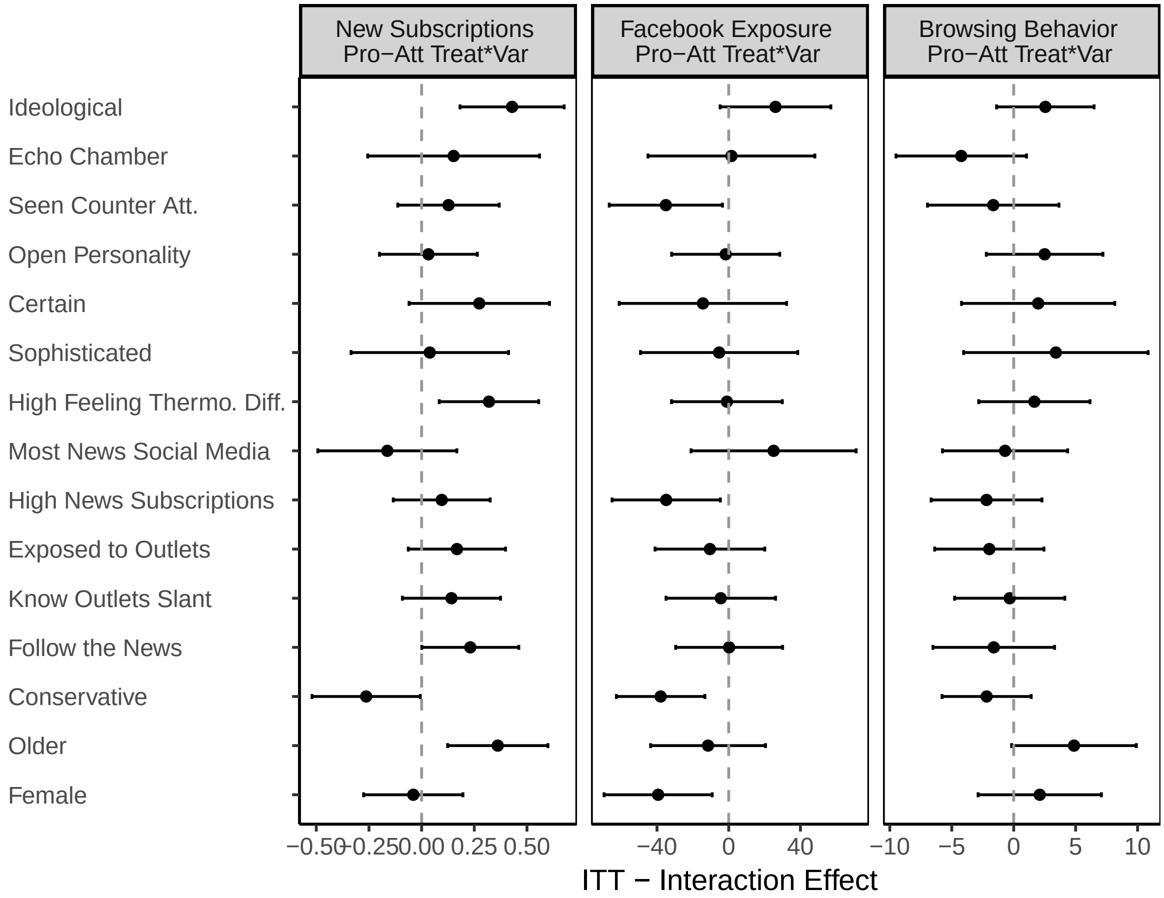
This figure shows the most common outlets and sections participants visited through links shared by the outlets they subscribed to. For more details see Figure A.15.

Figure A.17: Links in Posts Shared by Participants, by Outlet and Section



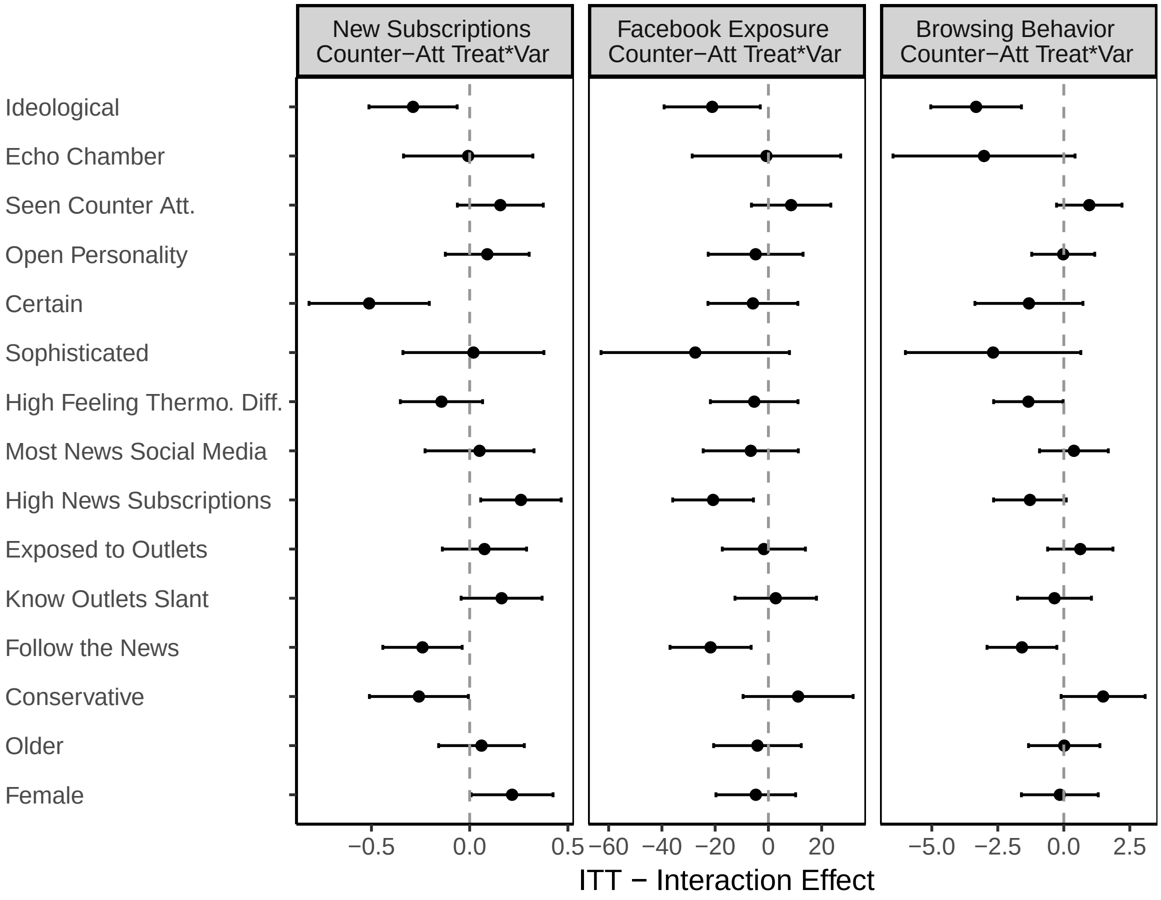
This figure shows the most common outlets and sections of the links participants shared when sharing posts from the outlets they subscribed to. For more details see Figure A.15.

Figure A.18: Heterogeneous Effects on Engagement with Pro-Attitudinal Outlets



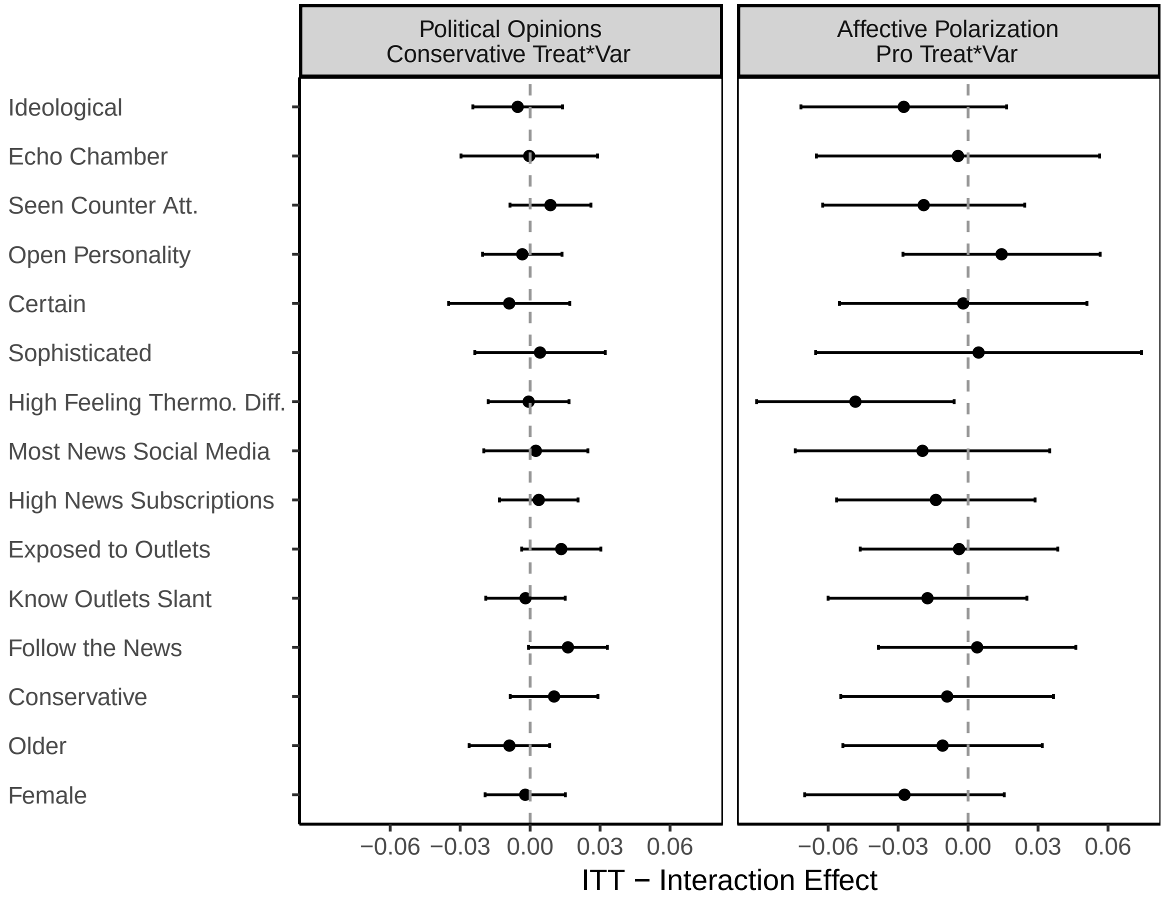
This figure shows heterogeneous effects of the pro-attitudinal treatment on engagement with the pro-attitudinal outlets. Each row presents the β coefficient in the following regression:
$Y_i = \alpha T_i^P + \beta T_i^P \times Var + \gamma Var + \delta X_i + \varepsilon_i$,
where the dependent variables are the number of potential pro-attitudinal outlets participants subscribed to (left panel), the number of posts from these outlets appearing in their feed (center panel), and the number of websites associated with these outlets that they visited (right panel). The regressions control for the set of potential outlets defined for each participant and baseline outcomes if they exist. A higher value means individuals were more likely to engage with pro-attitudinal outlets as a result of the pro-attitudinal treatment, compared to the control group. The definitions of the variables analyzed are described in Section C.3. Error bars reflect 90 percent confidence intervals.

Figure A.19: Heterogeneous Effects on Engagement with Counter-Attitudinal Outlets



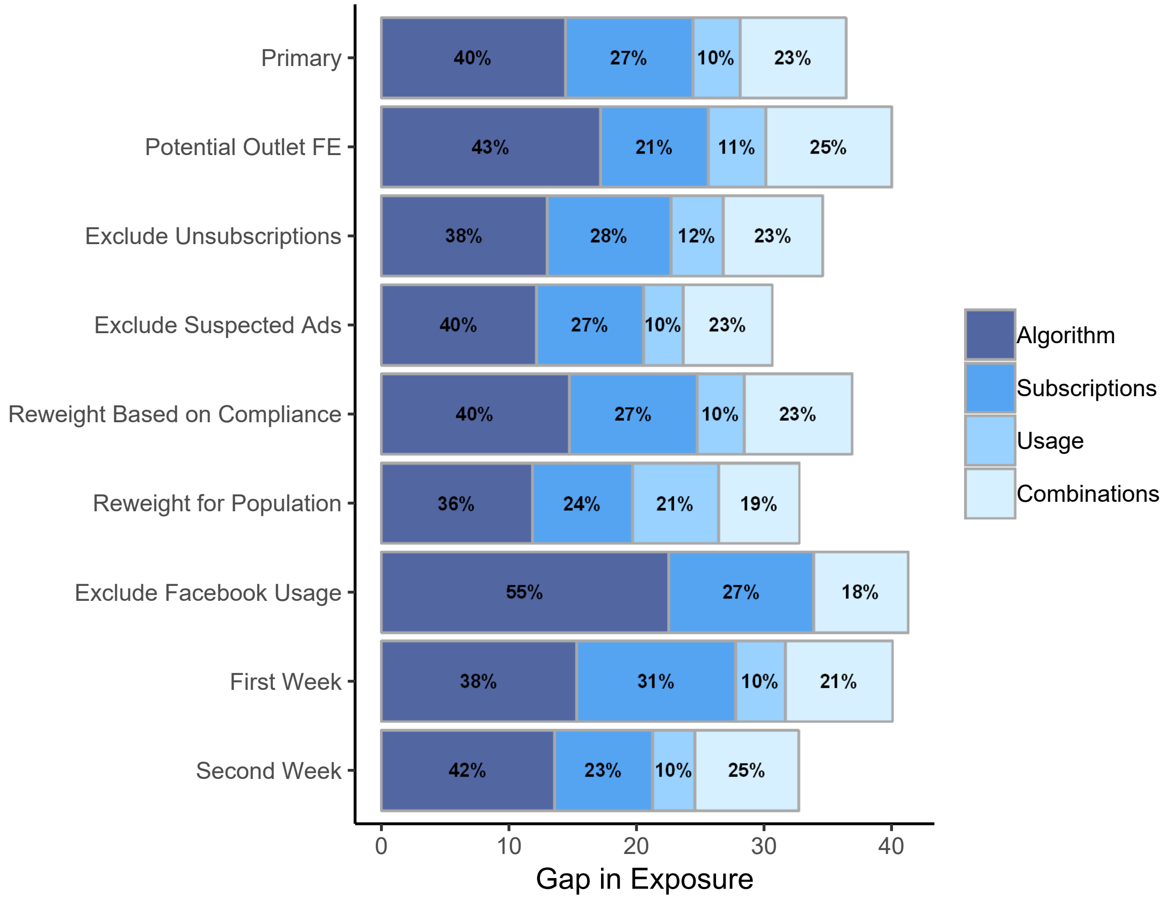
This figure shows heterogeneous effects of the counter-attitudinal treatment on engagement with the counter-attitudinal outlets. Each row presents the β coefficient in the following regression: $Y_i = \alpha T_i^A + \beta T_i^A \times Var + \gamma Var + \delta X_i + \varepsilon_i$, where the dependent variables are the number of potential counter-attitudinal outlets participants subscribed to (left panel), the number of posts from these outlets appearing in their feed (center panel), and the number of websites associated with these outlets that they visited (right panel). The regressions control for the set of potential outlets defined for each participant and baseline outcomes if they exist. A higher value means individuals were more likely to engage with counter-attitudinal outlets as a result of the counter-attitudinal treatment, compared to the control group. The definitions of the variables analyzed are described in Section C.3. Error bars reflect 90 percent confidence intervals.

Figure A.20: Heterogeneous Effects on Political Opinions and Affective Polarization



This figure shows heterogeneous effects on political opinions and affective polarization. In the left panel, each row represents the β coefficient in the following separate regression:
$Y_i = \alpha T_i^C + \beta T_i^C \times Var + \gamma Var + \delta X_i + \varepsilon_i$,
where the dependent variable is the political opinion index, and the independent variable is the full interaction of the conservative treatment and the variable analyzed in the row. A higher value means individuals were more likely to become more conservative by the conservative treatment, compared to the liberal treatment.
In the right panel, each row presents the β coefficient in the following regression:
$Y_i = \alpha T_i^P + \beta T_i^P \times Var + \gamma Var + \delta X_i + \varepsilon_i$,
where the dependent variable is the affective polarization index, and the independent variable is the full interaction of the pro-attitudinal treatment and the variable analyzed in the row. A higher value means individuals were more likely to become polarized as a result of pro-attitudinal treatment, compared to the counter-attitudinal treatment. The regressions control for the covariates specified in Section 2.5 along with the potential outlets defined for each participant. The definitions of the variables analyzed are described in Section C.3. Error bars reflect 90 percent confidence intervals.

Figure A.21: Decomposing the Gap Between Exposure to Posts from the Offered Pro-attitudinal and Counter-attitudinal Outlets, Additional Estimations



This figure decomposes the gap between the number of posts participants were exposed to from the offered pro-attitudinal and counter-attitudinal outlets. The first row repeats the main specification described in Figure 8. The second row controls for the potential outlets defined for each participant. The third row defined subscriptions as subscribing to the outlet for at least two weeks. The fourth row excludes posts that are likely to be sponsored (ads). The fifth row reweights the participants in each treatment such that the compliers resemble the entire sample. The sixth row reweights the participants such that the entire sample resembles the US population. The seventh row excludes differences in usage between the groups. The final two rows decompose the results separately for the first and second week after the intervention. Each row is described in more detail in Section C.7.2.

Table A.1: List of Outlets Offered and Subscriptions

Outlet	Group	Slant	Potential	Offered	Sub.	Share
The Washington Times	Conservative	0.70	37,120	12,366	3,278	0.27
The National Review	Conservative	0.90	36,168	12,057	2,953	0.24
The Wall Street Journal	Conservative	0.28	35,406	11,805	4,059	0.34
Fox News	Conservative	0.78	32,566	10,842	1,425	0.13
The Daily Caller	Conservative	0.87	4,522	1,471	323	0.22
Washington Examiner	Conservative	0.81	1,719	607	133	0.22
The Western Journal	Conservative	0.90	1,531	509	153	0.30
Townhall	Conservative	0.93	397	135	37	0.27
The Blaze	Conservative	0.89	221	80	25	0.31
The Conservative Tribune	Conservative	0.89	204	72	34	0.47
Newsmax	Conservative	0.77	114	32	14	0.44
Slate	Liberal	-0.68	35,206	11,738	3,008	0.26
MSNBC	Liberal	-0.81	35,091	11,688	2,786	0.24
HuffPost	Liberal	-0.62	31,927	10,643	2,359	0.22
The New York Times	Liberal	-0.55	30,337	10,145	3,376	0.33
Washington Post	Liberal	-0.26	8,234	2,824	1,341	0.47
Salon	Liberal	-0.88	5,119	1,668	595	0.36
Daily Kos	Liberal	-0.90	2,015	661	232	0.35
The Atlantic	Liberal	-0.54	636	203	116	0.57
Mother Jones	Liberal	-0.87	515	150	59	0.39
NPR	Liberal	-0.61	431	119	70	0.59
The New Yorker	Liberal	-0.76	317	105	65	0.62
PBS	Liberal	-0.54	134	40	23	0.57

This table shows the list of outlets included in the experiment. *Slant* is the slant from -1 to 1 of the domain associated with each outlet according to Bakshy et al. (2015). *Potential* is the number of participants for whom the outlet was defined as a potential outlet. These participants were offered the outlet if they were assigned to the treatment associated with the outlet's ideological group. *Offered* is the number of participants who were offered to subscribe to the outlet. *Sub.* is the number of participants who subscribed to each outlet in the intervention. *Share* is subscribed divided by offered. The first four liberal outlets and the first four conservative outlets are the primary outlets offered in the experiment and the rest of the outlets are the alternative outlets offered if a participant already subscribed to a primary outlet. Data is from the baseline sample.

Table A.2: Descriptive Statistics by Sample

		Baseline Sample	Access Posts Subsample	Endline Survey Subsample	Extension Subsample
1)	Ideology (-3, 3)	-0.61	-0.61	-0.71	-0.95
2)	Ideology, Abs. Value (0, 3)	1.75	1.75	1.80	1.81
3)	Feeling Therm., Rep.	29.07	29.22	27.54	22.86
4)	Feeling Therm., Dem.	46.99	47.02	47.79	51.21
5)	Feeling Therm., Difference	50.22	50.27	50.32	51.08
6)	Difficult Pers., Difference	1.92	1.92	1.96	1.92
7)	Most News Social Media	0.18	0.18	0.17	0.16
8)	Took Survey Mobile	0.67	0.67	0.63	0.00
9)	Female	0.52	0.52	0.52	0.49
10)	Age	47.69	47.65	48.78	52.47
11)	Total Subscriptions	474	474	472	481
12)	News Outlets Subscriptions	8.11	8.11	8.28	8.61
13)	Compliance	0.53	0.53	0.58	0.76
14)	N	37,494	34,592	17,635	1,835

This table presents descriptive statistics by subsample. *Baseline Sample* includes all participants. *Access-Posts Subsample* includes participants who provided access to posts they share for at least two weeks. *Endline Survey Subsample* includes participants who completed the baseline survey. *Extension Subsample* includes participants who installed the browser extension for at least two weeks. *Ideology, Abs. Value* is the absolute value of self-reported ideology. *Feeling Therm., Difference* is the difference between feelings toward the participant's party and the opposing party according to the feeling thermometer questions. *Difficult Pers., Difference* is the difference in whether participants find it difficult to see things from the opposing party and their own party's point of view. For all other variables, see Table 2.

Table A.3: Balance Table, Pro- and Counter-Attitudinal Treatments

	Mean		Difference		
Variable	Sample N=36,330	US	Control - Pro.	Control - Counter.	Pro. - Counter.
Baseline Survey					
Ideology, Abs. Value (0, 3)	1.80	1.31	0.00	-0.00	-0.00
Democrat	0.39	0.37	0.01	0.00	-0.01
Republican	0.17	0.30	0.00	-0.01	-0.01
Independent	0.36	0.29	-0.01*	0.00	0.01**
Vote Support Clinton	0.54		-0.00	-0.00	0.00
Vote Support Trump	0.27		0.00	0.00	0.00
Feeling Therm., Difference	50.22	38.44	0.36	0.41	0.05
Difficult Pers., Difference	1.92		0.03	0.02	-0.02
Facebook Echo Chamber	1.20		0.00	-0.01	-0.01
Follows News	3.36	2.48	0.01	0.01	0.01
Most News Social Media	0.17	0.12	0.00	-0.00	-0.01
Device					
Took Survey Mobile	0.67		-0.01*	-0.00	0.01*
Facebook					
Female	0.52	0.52	-0.01	-0.00	0.00
Age	47.91	47.70	0.02	0.08	0.06
Total Subscriptions	473		6.91	3.16	-3.75
News Outlets Slant, Abs. Value	0.54		-0.00	-0.00	0.00
Access Posts, Pre-Treat.	0.98		0.00	0.00	-0.00
Attrition					
Took Followup Survey	0.47		0.03***	0.03***	0.00
Access Posts, 2 Weeks	0.92		0.01	0.00	-0.00
Extension Install, 2 Weeks	0.05		0.00	-0.00	-0.00
F-Test			1.23	0.80	0.99
P-value			[0.20]	[0.74]	[0.48]

This table presents descriptive statistics by whether participants were assigned to the pro-attitudinal treatment, counter-attitudinal treatment, or control group. The second column shows summary statistics for American adults for whom an ideological leaning can be defined. *Ideology, Abs. Value* is the absolute value of self-reported ideology. *Feeling Therm., Difference* is the difference between the feeling toward the participant's party and the opposing party. *Difficult Pers., Difference* is the difference in whether participants find it difficult to see things from the opposing party and their own party point of view. *News Outlets Slant, Abs. Value* is the absolute value of the mean slant of all outlets participants subscribed to on Facebook in baseline. Slant ranges from -1 to 1 and is based on Bakshy et al. (2015). For all other variables see Table 2. Data sources for the US are specified in Appendix Section C.4.1.*p<0.1 **p<0.05 ***p<0.01

Table A.4: Balance Table, Liberal and Conservative Treatments, Among Participants Who Completed the Follow-up Survey

	Mean			Difference		
Variable	Sample N=17,635	US	FB Users	Control - Lib.	Control - Cons.	Cons. - Lib.
Baseline Survey						
Ideology (-3, 3)	-0.71	0.17		-0.01	-0.02	0.01
Democrat	0.40	0.35	0.30	0.01	0.01	0.01
Republican	0.16	0.28	0.21	0.00	0.00	0.00
Independent	0.36	0.32	0.35	-0.02*	-0.01	-0.01
Vote Support Clinton	0.55			-0.00	-0.00	-0.00
Vote Support Trump	0.25			0.01	-0.00	0.01
Feeling Therm., Rep.	27.54	43.06		0.20	-0.04	0.24
Feeling Therm., Dem.	47.79	48.70		0.43	0.68	-0.25
Difficult Pers., Rep. (1, 5)	3.18			0.04	0.01	0.04
Difficult Pers., Dem. (1, 5)	2.35			-0.01	-0.03	0.03
Facebook Echo Chamber	1.20		1.12	0.01	-0.01	0.01
Follows News	3.38	2.42		0.02	0.02	-0.00
Most News Social Media	0.17	0.13		-0.01**	-0.00	-0.01*
Device						
Took Survey Mobile	0.63			-0.01	0.01	-0.01
Facebook						
Female	0.52	0.52	0.55	-0.01	-0.00	-0.00
Age	48.78	47.30	42.86	0.55*	-0.31	0.86**
Total Subscriptions	472			2.37	15.27	-12.90
News Outlets Slant (-1, 1)	-0.20			0.00	-0.01	0.01
Access Posts, Pre-Treat.	0.98			0.00	0.00*	-0.00
F-Test				1.15	0.97	1.31
P-Value				[0.29]	[0.49]	[0.16]

This table presents descriptive statistics by whether participants were assigned to the liberal treatment, conservative treatment, or control group among participants who completed the endline survey. The variables are explained in the notes for Table 2. *p<0.1 **p<0.05 ***p<0.01

Table A.5: Balance Table, Pro- and Counter-Attitudinal Treatment, Among Participants Who Completed the Follow-up Survey

	Mean		Difference		
Variable	Sample N=17,130	US	Control - Pro.	Control - Counter.	Pro. - Counter.
Baseline Survey					
Ideology, Abs. Value (0, 3)	1.84	1.31	-0.00	0.00	0.00
Democrat	0.41	0.37	0.02*	0.01	-0.01
Republican	0.16	0.30	0.00	0.00	-0.00
Independent	0.35	0.29	-0.02**	-0.00	0.01
Vote Support Clinton	0.57		-0.00	0.00	0.00
Vote Support Trump	0.25		0.00	0.01	0.01
Feeling Therm., Difference	50.32	38.44	0.96*	1.10**	0.14
Difficult Pers., Difference	1.96		0.05*	0.04	-0.01
Facebook Echo Chamber	1.22		0.00	0.00	-0.00
Follows News	3.39	2.48	0.02	0.03*	0.00
Most News Social Media	0.17	0.12	-0.00	-0.01	-0.00
Device					
Took Survey Mobile	0.63		-0.01	0.01	0.01
Facebook					
Female	0.52	0.52	-0.01	-0.01	0.00
Age	48.96	47.70	0.12	0.20	0.08
Total Subscriptions	471		4.99	3.30	-1.69
News Outlets Slant, Abs. Value	0.55		-0.00	0.00	0.00
Access Posts, Pre-Treat.	0.98		-0.00	0.00	0.00
F-Test			0.63	0.75	0.57
P-value			[0.89]	[0.78]	[0.94]

This table presents descriptive statistics by whether participants were assigned to the pro-attitudinal treatment, counter-attitudinal treatment, or control group among participants who completed the endline survey. The variables are explained in the notes for Tables 2 and A.3. *p<0.1 **p<0.05 ***p<0.01

Table A.6: Descriptive Statistics by Compliance

		Control	All Comply:		Pro-Att. Comply:		Counter-Att. Comply:		Liberal Comply:		Conservative Comply:	
			Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
1)	Ideology (-3, 3)	-0.62	-0.92	-0.27	-0.86	-0.31	-1.05	-0.25	-1.13	-0.04	-0.71	-0.51
2)	Ideology, Abs. Value (0, 3)	1.80	1.77	1.73	1.83	1.75	1.78	1.82	1.78	1.72	1.75	1.75
3)	Democrat	0.40	0.43	0.32	0.44	0.32	0.46	0.34	0.47	0.27	0.40	0.37
4)	Republican	0.17	0.13	0.21	0.15	0.21	0.12	0.23	0.11	0.25	0.16	0.18
5)	Independent	0.35	0.36	0.37	0.35	0.38	0.36	0.35	0.35	0.38	0.37	0.36
6)	Vote Support Clinton	0.54	0.60	0.44	0.60	0.46	0.64	0.46	0.65	0.39	0.55	0.50
7)	Vote Support Trump	0.27	0.20	0.34	0.23	0.34	0.17	0.36	0.15	0.38	0.25	0.29
8)	Feeling Therm., Difference	50.47	50.24	49.92	51.23	48.52	49.03	51.02	50.70	49.33	49.79	50.51
9)	Difficult Pers., Difference	1.93	1.93	1.88	1.97	1.81	1.89	1.95	1.94	1.89	1.92	1.88
10)	Facebook Echo Chamber	1.20	1.21	1.15	1.23	1.14	1.22	1.19	1.23	1.13	1.19	1.17
11)	Most News Social Media	0.17	0.18	0.17	0.17	0.17	0.19	0.17	0.18	0.17	0.17	0.17
12)	Took Survey Mobile	0.67	0.67	0.67	0.67	0.68	0.68	0.66	0.69	0.67	0.66	0.67
13)	Female	0.52	0.57	0.46	0.56	0.47	0.60	0.45	0.59	0.45	0.56	0.47
14)	Age	47.94	48.32	46.95	49.03	46.32	47.86	47.86	48.18	46.74	48.46	47.16
15)	Total Subscriptions	476	509	430	496	431	521	429	515	428	504	431
16)	News Outlets Subscriptions	8.16	8.77	7.41	8.87	7.26	8.79	7.73	8.78	7.40	8.75	7.42
17)	Certain (0, 4)	3.16	3.12	3.18	3.14	3.17	3.11	3.20	3.11	3.17	3.13	3.19
18)	Open Personality (1, 7)	5.62	5.70	5.54	5.67	5.55	5.72	5.52	5.71	5.53	5.68	5.55
19)	Seen Counter-Att. Share	0.42	0.42	0.41	0.41	0.42	0.43	0.40	0.41	0.41	0.43	0.41
20)	N	12,104	13,258	11,734	7,115	4,985	5,791	6,335	6,604	5,893	6,654	5,841

This table presents descriptive statistics on compliance by treatment arm for the entire baseline sample. *Certain* is whether participants are extremely certain (4), very certain (3), somewhat certain (2), slightly certain (1), or not at all certain (0) of their political opinions. *Open Personality* is agreement with "I see myself as open to new experiences, complex" and the reverse values of "I see myself as conventional, uncreative." (both on a 7-point scale) (Gosling et al., 2003). *Seen Counter-Att. Share* is the share of potential counter-attitudinal outlets the participants reported seeing in their feed among all potential outlets. The rest of the variables are explained in Table 2 and Appendix Table A.3.

Table A.7: Additional Segregation Measures

(a) Segregation Measures Among Comscore Users Visiting News Sites Through Facebook

Category	Share	Isol.	Seg.	Slant, Abs.	Cong.	Extreme Pro	Mod. Pro	Mod.	Mod. Counter	Extreme Counter
1) All Browsing		0.026	0.194	0.244	0.073	0.127	0.308	0.264	0.219	0.081
2) Direct	45.3%	0.023	0.217	0.252	0.071	0.098	0.316	0.306	0.219	0.061
3) Social	27.6%	0.034	0.260	0.321	0.091	0.178	0.303	0.198	0.210	0.111
4) Search	21.7%	0.008	0.147	0.252	0.058	0.120	0.297	0.279	0.217	0.087
5) Other	5.4%	0.002	0.224	0.290	0.062	0.094	0.335	0.272	0.241	0.060
6) FB	26.3%	0.035	0.264	0.325	0.092	0.180	0.301	0.200	0.206	0.113
7) Non-FB	73.7%	0.020	0.186	0.236	0.066	0.109	0.309	0.291	0.221	0.071

(b) Segregation Measures Over Time, Comscore Data

Category	Share	Isol.	Seg.	Slant, Abs.	Cong.	Extreme Pro	Mod. Pro	Mod.	Mod. Counter	Extreme Counter
1) All: 2007-2008		0.021	0.174	0.256	0.032	0.063	0.379	0.187	0.317	0.054
2) All: 2017-2018		0.012	0.190	0.264	0.054	0.090	0.299	0.327	0.221	0.062

(c) Segregation Measures, Extension Data, Ideology Proxied using Zip Code

Category	Share	Isol.	Seg.	Slant, Abs.	Cong.	Extreme Pro	Mod. Pro	Mod.	Mod. Counter	Extreme Counter
1) Subscribed		0.024	0.361	0.554	0.161	0.356	0.228	0.091	0.148	0.177
2) FB Feed		0.011	0.218	0.373	0.131	0.206	0.332	0.158	0.195	0.108
3) Friends	48.2%	0.007	0.167	0.313	0.113	0.168	0.352	0.172	0.212	0.094
4) Pages	40.0%	0.010	0.286	0.449	0.152	0.253	0.305	0.137	0.177	0.126
5) Ads	11.9%	0.011	0.262	0.419	0.130	0.212	0.315	0.168	0.192	0.113
6) Browsing		0.008	0.196	0.325	0.117	0.144	0.358	0.215	0.200	0.082
7) Not FB	85.9%	0.004	0.194	0.320	0.114	0.133	0.362	0.223	0.201	0.079
8) FB	14.1%	0.006	0.227	0.361	0.138	0.194	0.323	0.183	0.202	0.097
9) Friends	59.5%	-0.007	0.208	0.329	0.112	0.179	0.330	0.197	0.204	0.092
10) Pages	40.5%	0.007	0.290	0.429	0.164	0.242	0.319	0.153	0.163	0.121
11) Shared		-0.024	0.250	0.412	0.144	0.195	0.351	0.124	0.197	0.132

These tables display additional measures of segregation. The first sub-table includes only individuals in the Comscore panel who visited multiple news sites through Facebook and through other means. The second sub-table includes the 2007-2008 and 2017-2018 Comscore panels. Ideology for 2007-2008 is based on the 2006 and 2008 election cycles FEC donation data and ideology for 2017-2018 is based on the 2016 and 2018 data. In the third sub-table, ideology for control group participants is proxied based on zip codes instead of survey answers. The segregation measures are defined in Section 3.

Table A.8: Segregation Measures, Visit-Level

(a) Comscore

Category	Share	Isol.	Seg.	Slant, Abs.	Cong.	Extreme Pro	Mod. Pro	Mod.	Mod. Counter	Extreme Counter
1) All Browsing		0.023	0.348	0.412	0.065	0.119	0.340	0.223	0.240	0.078
2) Direct	65.5%	0.030	0.359	0.424	0.064	0.109	0.353	0.217	0.250	0.071
3) Social	7.3%	0.040	0.412	0.500	0.095	0.227	0.266	0.172	0.185	0.150
4) Search	20.0%	0.009	0.264	0.352	0.074	0.124	0.320	0.266	0.217	0.074
5) Other	7.3%	0.006	0.318	0.380	0.025	0.088	0.344	0.216	0.276	0.076
6) FB	6.0%	0.045	0.422	0.513	0.096	0.240	0.255	0.172	0.173	0.160
7) Non-FB	94.0%	0.021	0.342	0.406	0.063	0.111	0.345	0.226	0.245	0.072

(b) Extension Data

Category	Share	Isol.	Seg.	Slant, Abs.	Cong.	Extreme Pro	Mod. Pro	Mod.	Mod. Counter	Extreme Counter
1) Subscribed		0.573	0.454	0.624	0.520	0.517	0.299	0.090	0.062	0.032
2) FB Feed		0.297	0.311	0.475	0.394	0.340	0.407	0.155	0.085	0.014
3) Friends	34.3%	0.206	0.291	0.429	0.322	0.257	0.452	0.162	0.108	0.021
4) Pages	57.2%	0.477	0.322	0.500	0.436	0.392	0.376	0.151	0.072	0.009
5) Ads	8.4%	0.353	0.306	0.496	0.406	0.325	0.432	0.146	0.079	0.018
6) Browsing		0.217	0.303	0.438	0.328	0.241	0.472	0.157	0.112	0.017
7) Not FB	90.8%	0.198	0.302	0.434	0.320	0.229	0.480	0.158	0.115	0.018
8) FB	9.2%	0.368	0.311	0.478	0.401	0.365	0.389	0.149	0.083	0.014
9) Friends	46.4%	0.230	0.290	0.434	0.329	0.278	0.435	0.165	0.101	0.021
10) Pages	53.6%	0.560	0.333	0.523	0.472	0.453	0.344	0.130	0.065	0.008
11) Shared		0.409	0.322	0.456	0.360	0.316	0.403	0.133	0.136	0.013

These tables display segregation measures based on visit-level data instead of aggregating data first at the user-level. In these tables users who visit more websites implicitly receive more weight. The first sub-table is based on Comscore data and the second is based on control group participants in the extension subsample. The segregation measures are defined in Section 3.

Table A.9: Effects of the Treatments on News Exposure, News Sites Visited and Sharing Behavior, Two Weeks Following the Intervention, Poisson Regression

	Pro-Att. Outlets Facebook Exposure	Pro-Att. Outlets Browsing Behavior	Pro-Att. Outlets Sharing Behavior	Counter-Att. Outlets Facebook Exposure	Counter-Att. Outlets Browsing Behavior	Counter-Att. Outlets Sharing Behavior
	(1)	(2)	(3)	(4)	(5)	(6)
Pro-Att. Treat.	1.34^{***}	0.29^{**}	0.57^{***}	0.33^{**}	0.19	0.17
	(0.13)	(0.14)	(0.21)	(0.16)	(0.25)	(0.31)
Counter-Att. Treat.	-0.06	-0.03	0.26	2.49^{***}	0.54^{***}	1.27^{***}
	(0.13)	(0.14)	(0.21)	(0.16)	(0.19)	(0.31)
Pro-Att. exponentiated	3.82	1.33	1.77	1.39	1.21	1.18
Counter-Att. exponentiated	0.94	0.97	1.3	12.11	1.71	3.56
Observations	1,648	1,648	1,648	1,648	1,648	1,648

This table presents the effects of the pro- and counter-attitudinal treatments on engagement with the potential pro- and counter-attitudinal outlets in the two weeks following the intervention, estimated using Poisson regressions. The sample includes participants with a liberal or conservative ideological leaning who installed the extension and provided permission to access their posts for at least two weeks following the intervention. The regressions control for the outcome measure in baseline if it exists. Robust standard error.
$^{*}p<0.1$ $^{**}p<0.05$ $^{***}p<0.01$

Table A.10: Effect of the Treatments on News Slant by Subsample

	News Exposure			Browsing Behavior			Shared Posts		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Liberal Treatment	-0.237^{***}	-0.234^{***}	-0.191^{***}	-0.092^{**}	-0.080^{**}	-0.100^{**}	-0.021^{*}	-0.106^{*}	-0.045
	(0.060)	(0.063)	(0.073)	(0.037)	(0.039)	(0.046)	(0.012)	(0.056)	(0.065)
Conservative Treatment	0.355^{***}	0.365^{***}	0.462^{***}	0.102^{**}	0.105^{**}	0.107^{**}	0.046^{***}	0.054	0.131^{*}
	(0.067)	(0.070)	(0.082)	(0.040)	(0.041)	(0.050)	(0.013)	(0.060)	(0.073)
Cons. Treat. - Lib. Treat.	0.59^{***}	0.60^{***}	0.65^{***}	0.19^{***}	0.19^{***}	0.21^{***}	0.07^{***}	0.16^{***}	0.18^{**}
	(0.06)	(0.07)	(0.08)	(0.04)	(0.04)	(0.05)	(0.01)	(0.06)	(0.07)
Ext. Subsample	X			X					
Posts Subsample							X		
Ext. + Posts Subsample		X			X			X	
Ext. + Posts + Endline Subsample			X			X			X
Observations	1,556	1,433	1,010	1,785	1,652	1,166	18,328	979	685

This table presents the effect of the treatments on the slant of outlets participants engaged with across various subsamples. The dependent variables are the mean slant in standard deviations of news participants were exposed to in their feed (column 1-3), of news sites they visited (columns 4-6), and of news they shared (columns 7-9). *Ext. Subsample* refers to the extension subsample, i.e., participants who installed the extension for at least two weeks. *Posts Subsample* refers to the access posts subsample, i.e., participants who provide permissions to access their posts for at least two weeks. *Ext + Posts Subsample* refers to participants in both these subsamples. *Ext + Posts + Endline Subsample* refers to participants who installed the extension, provided access to posts, and completed the endline survey. The regressions control for outcome variables in baseline when they exist. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.11: Effect of the Treatments on Primary Outcomes, Controlling for Covariates

(a) Effect of the Treatments on the Political Opinions Index

	(1)	(2)	(3)	(4)
Conservative Treatment	0.010	−0.002	−0.001	−0.001
	(0.018)	(0.006)	(0.005)	(0.005)
Liberal Treatment	−0.006	−0.009	−0.006	−0.006
	(0.018)	(0.006)	(0.005)	(0.005)
Conservative - Lib. Treatment	0.017	0.007	0.005	0.005
	(0.019)	(0.006)	(0.005)	(0.005)
Common Controls		X	X	X
Baseline Political Opinions Controls			X	X
Ex. Last Control Group Responders				X
Observations	17,635	17,635	17,635	17,237

(b) Effect of the Treatments on the Affective Polarization Index

	(1)	(2)	(3)	(4)
Pro-Att. Treatment	−0.022	−0.003	0.005	0.005
	(0.019)	(0.015)	(0.012)	(0.012)
Counter-Att. Treatment	−0.055***	−0.039**	−0.028**	−0.028**
	(0.019)	(0.015)	(0.012)	(0.012)
Pro-Att. Lower Lee Bound	-0.132	-0.072	-0.03	-0.012
Pro-Att. Upper Lee Bound	0.086	0.076	0.065	0.018
Counter-Att. Lower Lee Bound	-0.172	-0.115	-0.064	-0.041
Counter-Att. Upper Lee Bound	0.06	0.045	0.037	-0.016
Pro-Att. - Counter-Att. Treat	0.033*	0.035**	0.033***	0.033***
	(0.019)	(0.015)	(0.012)	(0.012)
Common Controls		X	X	X
Baseline Polarization Controls			X	X
Ex. Last Control Group Responders				X
Observations	16,896	16,896	16,896	16,514

These tables present the effects on the political opinions and affective polarization indices. Column (1) does not control for any covariates. Column (2) controls for self-reported ideology, party affiliation, 2016 candidate supported, ideological leaning, age, age squared, and gender. Column (3) also controls for baseline questions similar to endline questions composing each index. Column (4) excludes control group participants recruited to the follow-up survey with the last email sent or ad published. Without these participants, attrition is similar across treatments. In the specifications with control variables, I first trim the excess observation and then run the regressions with the controls. The specification and controls are described in more detail in Section 2.5. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.12: Effect of the Treatments on the Affective Polarization Index, Excluding Each Index Component

	(1)	(2)	(3)	(4)	(5)	(6)
Pro-Att. Treatment	0.005	0.001	0.008	0.005	0.002	0.010
	(0.012)	(0.013)	(0.013)	(0.012)	(0.013)	(0.012)
Counter-Att. Treatment	-0.028^{**}	-0.033^{**}	-0.018	-0.029^{**}	-0.035^{***}	-0.020^{*}
	(0.012)	(0.013)	(0.013)	(0.012)	(0.013)	(0.012)
Pro - Counter	0.033^{***}	0.034^{**}	0.025^{**}	0.034^{***}	0.038^{***}	0.030^{**}
	(0.012)	(0.014)	(0.013)	(0.012)	(0.013)	(0.012)
Excluded Measure		Feeling Thermometer	Difficult Perspective	Consider Perspective	Party Ideas	Marry Opposing Party
Observations	16,896	16,896	16,896	16,896	16,895	16,896

This table presents the effect of the treatments on the affective polarization index. Column (1) is the primary specification. In columns (2)-(6), the index is created with four of the five affective polarization index components. The specification and controls are described in more detail in Section 2.5. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.13: Effect of the Treatments on Primary Outcomes, According to Outlets Offered

(a) Effect of the Treatments on the Political Opinions Index

	(1)	(2)	(3)
Liberal Treatment	-0.006	-0.010	-0.007
	(0.005)	(0.007)	(0.005)
Conservative Treatment	-0.001	-0.007	-0.002
	(0.005)	(0.007)	(0.005)
Cons. Treat - Lib. Treat	0.005	0.003	0.005
	(0.005)	(0.007)	(0.005)
Controls	X	X	X
Only Primary Outlet		X	
Potential Outlets FE			X
Observations	17,635	9,630	17,635

(b) Effect of the Treatments on the Affective Polarization Index

	(1)	(2)	(3)
Pro-Att. Treatment	0.005	-0.001	0.004
	(0.012)	(0.016)	(0.013)
Counter-Att. Treatment	-0.028^{**}	-0.031^{*}	-0.032^{**}
	(0.012)	(0.016)	(0.013)
Pro-Att. Treat. - Counter-Att. Treat	0.033^{***}	0.029^{*}	0.036^{***}
	(0.012)	(0.017)	(0.013)
Controls	X	X	X
Only Primary Outlet		X	
Potential Outlets FE			X
Observations	16,896	9,125	16,896

These tables present the effects of the treatments on the political opinions index and the affective polarization index. Column (1) is the primary specification and includes all participants. Column (2) includes only participants who did not subscribe in baseline to any of the four primary liberal outlets or the four primary conservative outlets. Thus, in this column, all participants in the liberal treatment were offered the same four primary liberal outlets and all participants in the conservative treatment were offered the same conservative outlets. Column (3) controls for the set of eight potential liberal and conservative outlets defined for each participant. The specification and controls are described in more detail in Section 2.5. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.14: Effect of the Treatments on Primary Outcomes, by Subsample

(a) Effect of the Treatments on the Political Opinions Index

	(1)	(2)	(3)	(4)
Liberal Treatment	−0.006	−0.007	−0.011	−0.020
	(0.005)	(0.005)	(0.018)	(0.019)
Conservative Treatment	−0.001	−0.003	0.002	−0.001
	(0.005)	(0.005)	(0.018)	(0.018)
Conservative Treat - Lib. Treat	0.005	0.004	0.013	0.018
	(0.005)	(0.005)	(0.018)	(0.018)
Controls	X	X	X	X
Sample	Endline	Endline+ Posts	Endline+ Ext	Endline+ Posts+Ext
Observations	17,635	16,339	1,286	1,196

(b) Effect of the Treatments on the Affective Polarization Index

	(1)	(2)	(3)	(4)
Pro-Att. Treatment	0.005	0.008	0.015	0.027
	(0.012)	(0.013)	(0.044)	(0.046)
Counter-Att. Treatment	-0.028^{**}	-0.027^{**}	-0.072^{*}	−0.056
	(0.012)	(0.013)	(0.043)	(0.045)
Pro-Att. Treat. - Counter-Att. Treat	0.033^{***}	0.035^{***}	0.087^{**}	0.083^{*}
	(0.012)	(0.013)	(0.043)	(0.045)
Controls	X	X	X	X
Sample	Endline	Endline+ Posts	Endline+ Ext	Endline+ Posts+Ext
Observations	16,896	15,647	1,241	1,151

These tables present the effects of the treatments on the political opinions index and the affective polarization index. Column (1) is the primary specification and includes all participants who completed the endline survey (the endline survey subsample). Column (2) includes participants who also provided permissions to access their posts for at least two weeks. Column (3) includes only participants who installed the extension for at least two weeks. Column (4) includes participants who both provided access to their posts and installed the extension. The specification and controls are described in more detail in Section 2.5. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.15: Effect of News Exposure on Affective Polarization

(a) Causal Effect Based on Experimental Variation

	IV Affective Polarization	
	(1)	(2)
FB Counter-Att. Share, Std. Dev.	-0.130^{*} (0.067)	
FB Congruence Scale, Std. Dev.		0.105^{*} (0.057)
Controls	X	X
First Stage F	65.1	65.22
Observations	1,072	1,072

(b) Cross-Sectional Correlation in Control Group

	OLS	OLS
	Affective Polarization	
	(1)	(2)
FB Counter-Att. Share, Std. Dev.	-0.385^{***} (0.052)	
FB Congruence Scale, Std. Dev.		0.407^{***} (0.054)
Data	Control Group	Control Group
Observations	352	352

These tables measure the association between exposure to pro- and counter-attitudinal news and affective polarization. *FB Counter-Att. Share* is the share of news form counter-attitudinal outlets participants were exposed to on Facebook between the baseline and endline surveys, among all news from pro- and counter-attitudinal outlets. *FB Congruence Scale* is the mean slant of all news exposed to on Facebook, multiplied by (-1) for liberal participants. Sub-table (a) shows the results of IV regressions, where the independent variables are instrumented with the treatment. Sub-table (b) presents the results of regressions run only among control group participants, where the dependent variable is the affective polarization index and the independent variables are the two summary statistics (with no controls). The regressions control for the covariates specified in Section 2.5 and include all participants who are both in the endline and extension subsamples. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.16: Effect of the Treatments on Attitudes Toward Each Party

	Attitude Own Party	Attitude Opposing Party
	(1)	(2)
Pro-Att. Treatment	0.008	−0.003
	(0.013)	(0.014)
Counter-Att. Treatment	0.001	0.031**
	(0.014)	(0.014)
Pro - Counter	0.007	-0.035**
	(0.014)	(0.014)
Observations	16,896	16,896

This table presents the effect of the pro and counter-attitudinal treatments on attitudes toward the party the participant is associated with and the opposing party. Participants whose ideological leaning is defined as liberal are assumed to be associated with the Democratic Party and participants whose ideological leaning is defined as conservative are assumed to be associated with the Republican Party. The outcome for each party is an index composed of the following four questions: the feeling thermometer, how difficult it is to see things from each party's point of view, how important it is to consider the perspective of the party, and whether the party has good ideas. The controls and the definition of ideological leaning are specified in Section 2.5. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.17: Primary Outcomes Using Different Index Methods

(a) Political Opinions

	(1)	(2)	(3)	(4)	(5)
Liberal Treatment	−0.006	−0.008	−0.054	−0.006	−0.004
	(0.005)	(0.017)	(0.090)	(0.007)	(0.005)
Conservative Treatment	−0.001	0.025	−0.072	0.007	0.003
	(0.005)	(0.017)	(0.091)	(0.007)	(0.005)
Cons. - Lib. Treatment	0.005	0.033*	-0.018	0.013*	0.008
	(0.005)	(0.017)	(0.050)	(0.007)	(0.005)
Controls	X	X	X	X	X
Index Method	Standard	Inv-Cov	Inv-Cov	Inv-Cov	Inv-Cov
Include Missing Outcomes	-	No	Yes	No	Yes
Replace Negative Weights With 0	-	Yes	Yes	No	No
Observations	17,635	9,434	17,635	9,434	17,635

(b) Affective Polarization

	(1)	(2)	(3)
Pro-Att. Treatment	0.005	0.004	0.001
	(0.012)	(0.017)	(0.010)
Counter-Att. Treatment	−0.028**	−0.031*	−0.026***
	(0.012)	(0.017)	(0.010)
Pro-Att. Treat. - Counter-Att. Treatment	0.033***	0.035**	0.027***
	(0.012)	(0.017)	(0.010)
Controls	X	X	X
Index Method	Standard	Inv-Cov	Inv-Cov
Include Missing Outcomes	-	No	Yes
Observations	16,896	10,059	16,896

These tables estimate the effects of the treatments on the primary outcomes using different summary indexes. Column (1) uses equal weights for all outcomes in the index. Column (2) uses inverse covariate weights and excludes participants with missing values for any of the index components. In Column (3), participants with missing outcomes are included with weights renormarlized to sum to one, such that an outcome measure is created for all participants who have at least one non-missing outcome. Columns (4) and (5) repeat columns (2) and (3) with non-negative weights replaced with zeros and all weights renormalized to sum to one. The specification and controls are described in Section 2.5. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.18: Effect of the Treatments on Behavioral and Attitudinal Polarization Measures

	All	Affective	Behavior
Pro-Att. Treatment	0.006	0.005	−0.001
	(0.014)	(0.012)	(0.018)
Counter-Att. Treatment	−0.028**	−0.028**	−0.010
	(0.014)	(0.012)	(0.018)
Counter-Att. Treatment - Pro-Att. Treat.	0.035**	0.033***	0.009
	(0.014)	(0.012)	(0.019)
Controls	X	X	X
Observations	17,159	16,896	16,637

This table estimates the effects of the treatments on polarization indices. Column (1) includes the five affective components and the three behavioral components. Column (2) is the primary outcome analyzed in the paper and includes the five affective components. Column (3) includes the three behavioral components. The specification and controls are described in Section 2.5. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.19: Common Phrases Mentioned When Describing the Baseline Survey's Objective

(a) Common Three-Word Expressions by Treatment Assignment

Rank	Control	Counter	Pro
1	social media polit (0.91%)	social media polit (1.20%)	social media polit (1.36%)
2	media influenc polit (0.75%)	media influenc polit (0.94%)	media influenc polit (0.90%)
3	peopl get news (0.70%)	effect social media (0.85%)	peopl get news (0.78%)
4	peopl polit view (0.53%)	peopl get news (0.83%)	effect social media (0.66%)
5	social media influenc (0.49%)	social media influenc (0.73%)	peopl polit view (0.61%)
6	effect social media (0.46%)	social media news (0.57%)	media polit view (0.57%)
7	influenc social media (0.46%)	peopl polit view (0.56%)	social media news (0.56%)
8	media affect polit (0.44%)	media echo chamber (0.53%)	social media influenc (0.53%)
9	current polit climat (0.40%)	media polit view (0.52%)	influenc social media (0.46%)
10	social media news (0.38%)	influenc social media (0.46%)	media echo chamber (0.46%)
11	media polit view (0.38%)	media affect polit (0.41%)	polit view media (0.41%)
12	correl polit view (0.37%)	social media affect (0.40%)	social media affect (0.41%)
13	see social media (0.34%)	social media echo (0.40%)	social media effect (0.39%)
14	polit view media (0.33%)	impact social media (0.39%)	current polit climat (0.37%)
15	affect polit view (0.32%)	influenc polit view (0.38%)	influenc polit view (0.37%)

(b) Common Two-Word Expressions by Treatment Assignment

Rank	Control	Counter	Pro
1	polit view (8.31%)	social media (9.67%)	social media (9.77%)
2	social media (7.47%)	polit view (8.41%)	polit view (8.40%)
3	polit opinion (4.20%)	polit opinion (4.13%)	polit opinion (4.13%)
4	polit lean (3.39%)	news sourc (3.92%)	news sourc (3.58%)
5	news sourc (2.63%)	polit lean (3.10%)	polit lean (3.57%)
6	media polit (2.31%)	media polit (2.43%)	media polit (2.83%)
7	polit climat (1.91%)	echo chamber (2.34%)	echo chamber (1.97%)
8	polit parti (1.90%)	media influenc (1.95%)	see peopl (1.96%)
9	get news (1.69%)	see peopl (1.80%)	media influenc (1.84%)
10	media influenc (1.67%)	get news (1.74%)	media bias (1.69%)
11	media bias (1.64%)	peopl polit (1.61%)	polit parti (1.69%)
12	see peopl (1.54%)	polit parti (1.58%)	get news (1.61%)
13	liber conserv (1.47%)	polit affili (1.54%)	polit affili (1.55%)
14	peopl polit (1.45%)	polit belief (1.54%)	polit belief (1.55%)
15	polit affili (1.43%)	media bias (1.49%)	polit climat (1.55%)

These tables show words participants mentioned often when asked "If you had to guess, what would you say is the primary purpose of this study?" at the end of the baseline survey. I first process the text by removing non-ascii characters, converting all characters to lowercase, remove common stop words and stemming words to their roots. The share of responses that include the phrase appears in parenthesis.

Table A.20: Expressions with Highest Usage Differential When Describing the Survey's Purpose

(a) Control Group and the Pro-Attitudinal Treatment

Expression	Share Among Phrases with the Same Length		
	Control	Pro	Counter
chamber	0.2%	0.4%	0.5%
divers	0.0%	0.1%	0.1%
echo	0.2%	0.4%	0.5%
echo chamber	0.20%	0.51%	0.58%
media echo	0.02%	0.12%	0.13%
media echo chamber	0.022%	0.153%	0.167%
open	0.0%	0.2%	0.2%
page	0.0%	0.1%	0.2%
social	1.7%	2.2%	2.1%
social media	1.91%	2.56%	2.40%

(b) Control Group and the Counter-Attitudinal Treatment

chamber	0.2%	0.4%	0.5%
divers	0.0%	0.1%	0.1%
echo	0.2%	0.4%	0.5%
echo chamber	0.20%	0.51%	0.58%
like	0.2%	0.3%	0.5%
open	0.0%	0.2%	0.2%
page	0.0%	0.1%	0.2%
percept	0.9%	0.6%	0.5%
promot	0.0%	0.1%	0.1%
willing	0.0%	0.0%	0.1%

(c) Pro-Attitudinal Treatment and Counter-Attitudinal Treatment

connect polit	0.04%	0.07%	0.02%
like	0.2%	0.3%	0.5%
peopl identifi	0.02%	0.04%	0.01%
percept media polit	0.035%	0.042%	0
polit	10.6%	10.4%	9.7%
push	0.0%	0.1%	0.1%
push liber	0.02%	0.03%	0.09%
rang	0.0%	0.0%	0.0%
seem like	0.01%	0	0.03%
social media bias	0.035%	0.066%	0.013%

These tables show the expression with 1, 2, 3, or 4 words with the highest differential usage between treatment arms. Differential usage is calculated using the following formula: $\chi^2 = \frac{(f_1 f_{-2} * f_2 f_{-1})^2}{(f_1+f_2)(f_1+f_{-1})(f_2+f_{-2})(f_{-1}+f_{-2})}$ where f_1, f_2 are the occurrence of the expression in the first and second groups, and f_{-1}, f_{-2} are the occurrence of all other expressions in the first and second groups. I first process the text by removing non-ascii characters, converting all characters to lowercase, remove common stop words and stemming words to their roots.

Table A.21: Most Common 2-Words Phrases Appearing in Posts

(a) Post Participants were Exposed to in their Feed

Exposed in Feed Conservative Outlets		Exposed in Feed Liberal Outlets	
Pro	Counter	Pro	Counter
donald trump (11.0%)	donald trump (5.2%)	presid trump (8.4%)	presid trump (7.5%)
presid donald (9.2%)	presid trump (5.1%)	donald trump (4.0%)	donald trump (4.8%)
presid trump (3.7%)	presid donald (3.0%)	white hous (3.2%)	white hous (2.7%)
white hous (2.9%)	white hous (2.6%)	stormi daniel (1.9%)	presid donald (2.2%)
high school (2.3%)	high school (1.6%)	presid donald (1.6%)	stormi daniel (2.1%)
hillari clinton (1.6%)	trump administr (1.4%)	high school (1.1%)	high school (1.2%)
gun control (1.6%)	gun control (1.2%)	special counsel (1.0%)	michael cohen (1.2%)
school shoot (1.4%)	school shoot (1.1%)	unit state (1.0%)	unit state (1.0%)
trump administr (1.3%)	special counsel (0.9%)	school shoot (1.0%)	special counsel (0.9%)
attorney general (1.2%)	hillari clinton (0.9%)	michael cohen (0.9%)	gun violenc (0.9%)

(b) Post With Links Visited by Participants

Posts Visited Conservative Outlets		Posts Visited Liberal Outlets	
presid trump (4.85%)	presid trump (5.25%)	presid trump (5.26%)	donald trump (3.04%)
donald trump (4.22%)	donald trump (3.22%)	donald trump (4.35%)	presid trump (3.04%)
white hous (2.32%)	white hous (3.22%)	white hous (2.09%)	day befor (0.91%)
presid donald (2.11%)	gun control (2.04%)	high school (1.19%)	former fbi (0.91%)
gun control (1.69%)	hillari clinton (1.61%)	presid donald (1.07%)	high school (0.91%)
high school (1.69%)	second amend (1.61%)	school shoot (0.79%)	someon els (0.91%)
attorney general (1.27%)	presid donald (1.29%)	special counsel (0.73%)	white hous (0.91%)
hillari clinton (1.27%)	robert mueller (1.29%)	unit state (0.73%)	anoth child (0.61%)
justic depart (1.27%)	special counsel (1.29%)	michael cohen (0.68%)	anyon els (0.61%)
north korea (1.27%)	trump administr (1.18%)	robert mueller (0.68%)	black student (0.61%)

(c) Posts Shared by Participants

Shared Posts Conservative Outlets		Shared Posts Liberal Outlets	
donald trump (6.37%)	presid trump (4.43%)	presid trump (9.94%)	presid trump (3.93%)
presid donald (4.51%)	donald trump (4.33%)	donald trump (4.91%)	donald trump (3.59%)
high school (4.25%)	white hous (3.75%)	white hous (3.17%)	presid donald (2.05%)
illeg immigr (4.19%)	high school (2.31%)	presid donald (1.75%)	unit state (1.20%)
hillari clinton (3.21%)	gun control (2.02%)	trump administr (1.66%)	attorney general (1.03%)
presid trump (3.00%)	presid donald (1.92%)	school shoot (1.65%)	break presid (1.03%)
trump administr (2.38%)	trump administr (1.73%)	high school (1.58%)	cambridg analytica (1.03%)
gun control (2.23%)	special counsel (1.64%)	mass shoot (1.54%)	gun violenc (1.03%)
second amend (2.02%)	gun violenc (1.44%)	stormi daniel (1.54%)	high school (1.03%)
white hous (1.61%)	robert mueller (1.44%)	robert mueller (1.51%)	school shoot (1.03%)

These tables show the most common two-word phrases mentioned in posts from outlets participants subscribed to. Stop word, punctuation and additional media-related words are removed and the words are then stemmed. Posts from the pages of the four primary and first alternative outlets (excluding suspected ads) in the first eight weeks following the intervention are included.

Table A.22: Effect of the Treatments on Primary Outcomes, Reweighted to Match the US Population

	News Exposure		Browsing Behavior		Shared Posts	
	(1)	(2)	(3)	(4)	(5)	(6)
Liberal Treatment	-0.237^{***}	-0.337^{***}	-0.092^{**}	-0.059	-0.021^{*}	-0.011
	(0.060)	(0.095)	(0.037)	(0.052)	(0.012)	(0.019)
Conservative Treatment	0.355^{***}	0.419^{***}	0.102^{**}	0.148^{**}	0.046^{***}	0.067^{***}
	(0.067)	(0.100)	(0.040)	(0.068)	(0.013)	(0.019)
Cons. Treat. - Lib. Treat.	0.59^{***}	0.76^{***}	0.19^{***}	0.21^{***}	0.07^{***}	0.08^{***}
	(0.06)	(0.09)	(0.04)	(0.07)	(0.01)	(0.02)
Reweigted		X		X		X
Observations	1,556	1,556	1,785	1,785	18,328	18,328

This table estimates the effect of the treatments on the slant of posts observed in the Facebook feed, websites visited and posts shared. Columns (1), (3), and (5) show the estimates in the extension or access posts subsamples using equal weights. These columns are the same as columns (1), (4), and (7) in Appendix Table A.10. Columns (2), (4), and (6) reweight the subsamples to match the population based on the following covariates: self-reported ideology, the share of participants identifying as Democrats, Republicans, and Independents, the difference between the participants' feelings toward their party and the opposing party, age, and the share of females. This analysis is discussed in Appendix C.4. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.23: Effect of the Treatments on Primary Outcomes, Reweighted to Match the US Population

(a) Political Opinions

	(1)	(2)
Liberal Treatment	−0.006	−0.005
	(0.005)	(0.007)
Conservative Treatment	−0.001	−0.0003
	(0.005)	(0.008)
Cons. Treat - Lib. Treat	0.005	0.005
	(0.005)	(0.008)
Controls	X	X
Reweighted		X
Observations	17,635	17,635

(b) Affective Polarization

	(1)	(2)
Pro-Att. Treatment	0.005	0.019
	(0.012)	(0.020)
Counter-Att. Treatment	-0.028^{**}	−0.014
	(0.012)	(0.022)
Pro-Att. Treat. - Counter-Att. Treat	0.033^{***}	0.033
	(0.012)	(0.020)
Controls	X	X
Reweighted		X
Observations	16,896	16,896

These tables estimate the effect of the treatments on the polarization and political opinions indices after reweighting the endline participants. Column (1) uses equal weights for all participants. Column (2) reweights the participants to match the population means based on the following covariates: self-reported ideology, the share of participants identifying as Democrats, Republicans, and Independents, the difference between the participants' feelings toward their own party and the opposing party, age, and the share of females. This analysis is discussed in Appendix C.4. The specification and controls are described in Section 2.5. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.24: Predicted Effect in Full Baseline Sample

Outcome	Treatment	Predicted Effect in Subsample	Predicted Effect in Baseline Sample
News exposure, posts slant	Conservative treatment, compared to liberal treatment	0.542	0.567
Browsing behavior, news sites slant	Conservative treatment, compared to liberal treatment	0.197	0.211
Political opinions index	Conservative treatment, compared to liberal treatment	0.004	0.004
Affective polarization index	Pro-Attitudinal treatment, compared to counter-attitudinal treatment	0.029	0.030

This table predicts the main effects estimated in the paper for the baseline sample. I first estimate heterogeneous effects in the endline survey and extension subsamples using causal forests with many survey and Facebook covariates as explained in Section C.5. Column (3) predicts the treatment effect within the subsample using out-of-bag prediction. Column (4) predicts the effect for the entire baseline sample.

Table A.25: Effect of the Treatments on Self-reported Familiarity and Accurate Political Knowledge Outcomes

	Heard Michael Cohen	Heard Clark Shooting	Heard Louis Farrakhan	Heard Clinton Speech	Correct Russian Influence	Correct Wall Built	Correct Trump Target	Correct Tax Cut
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Liberal Treatment	−0.004	0.007	−0.004	0.008	0.002	0.016*	−0.003	−0.001
	(0.006)	(0.007)	(0.006)	(0.008)	(0.005)	(0.009)	(0.009)	(0.006)
Conservative Treatment	−0.002	0.002	−0.002	0.019**	0.010*	0.0001	−0.007	0.0004
	(0.006)	(0.007)	(0.006)	(0.008)	(0.005)	(0.009)	(0.009)	(0.006)
Cons. Treat - Lib. Treat	0.00	-0.01	0.00	0.01	0.01	-0.02*	-0.00	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Controls	X	X	X	X	X	X	X	X
Expected Effect	Lib Treat	Lib Treat	Cons Treat	Cons Treat	Lib Treat	Lib Treat	Cons Treat	Cons Treat
Observations	17,635	17,431	17,635	17,464	16,167	13,872	12,141	15,655

This table estimates the effect of the treatments on eight knowledge outcomes. All the outcomes are binary. *Michael Cohen* and *Louis Farrakhan* are whether the participant did not mark "Never heard of" when asked for their favorability ratings of the individuals. *Clark Shooting* is whether the participant heard that Stephon Clark was shot and killed by police officers in Sacramento. *Clinton Speech* is whether the participant heard that Hillary Clinton suggested many white women voted for Trump since they took their voting cues from their husbands. *Russian Influence* is agreement with "the Russian government tried to influence the 2016 presidential election". *Wall Built* is disagreement with "the US has recently started building a new border wall at the US-Mexico border." *Trump Target* is disagreement with "President Trump is a criminal target of Robert Mueller's investigation." *Tax Cut* is agreement with "most people will receive an income tax cut, salary increase or bonus under the new tax reform law." All regressions control for party affiliation, ideology, vote, age, age squared, whether the participant follows the news and whether the participant stated they know the name of their representative in congress. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.26: Effect of the Treatments on Exposure to Words in the Facebook Feed

	Michael Cohen	Clark Shooting	Louis Farrakhan	Clinton Speech
	(1)	(2)	(3)	(4)
Liberal Treatment	1.828***	1.187***	0.156	0.044
	(0.573)	(0.347)	(0.116)	(0.042)
Conservative Treatment	0.634	0.146	0.368***	0.076**
	(0.424)	(0.260)	(0.101)	(0.032)
Cons. Treat - Lib. Treat	-1.19**	-1.04***	0.21*	0.03
	(0.58)	(0.31)	(0.12)	(0.04)
Controls	X	X	X	X
Expected Effect	Lib. Treat	Lib. Treat	Cons. Treat	Cons. Treat
Observations	1,720	1,720	1,720	1,720

This table estimates the effect of the treatments on topics appearing in participants' Facebook feeds. *Michael Cohen*, *Clark Shooting*, and *Louis Farrakhan* are the number of times the expressions "Michael Cohen", "Stephon Clark", and "Louis Farrakhan" appeared, respectively. *Clinton Speech* is the number of times the word Clinton appeared along with the word vote and either the word India or the word husband. All regressions control for party affiliation, ideology, vote, age, age squared, whether the participant follows the news and whether the participant stated they know the name of their representative in congress. Data is from the extension subsample and all posts until April 15, 2018 are included in the analysis. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.27: Estimations Decomposing the Segregation in News Exposure

	Subscriptions	FB Usage: Total Posts Observed	Platform Algorithm: Share of Posts
	OLS	OLS	IV
	(1)	(2)	(3)
Pro-Att. Treatment	0.505^{***}	17.769^{*}	
	(0.086)	(10.762)	
Subscriptions			0.966^{***}
			(0.093)
Subscriptions * Pro-Att.			0.460^{***}
			(0.162)
Unit	Participant	Participant	Participant* Outlet Group
Baseline Controls		X	
Mean in Counter-Att. Treatment	1.535	145.93	0.851
Observations	1,059	1,059	2,117

This table displays the regressions used to decompose the gap in exposure to posts from the offered pro- and counter-attitudinal outlets. In column (1), the dependent variable is the number of outlets the participant subscribed to. In column (2), the dependent variable is the total number of posts observed by the participant on Facebook per day. The regression controls for Facebook visits before the intervention. In column (3), the two groups of outlets and participants are pooled in an IV regression. Each observation is a participant and the group of pro-attitudinal or counter-attitudinal outlets. The dependent variable is the share of posts (in percentage points) from the group of outlets that the participant was exposed to among all posts in the participant's Facebook feed and the independent variable is the full interaction of the number of outlets the participant subscribed to among this group of outlets and whether the outlets in the group are pro-attitudinal. Subscriptions are instrumented with whether this group of outlets was offered in the experiment. The first two columns use robust standard errors and in the third column standard errors are clustered at the participant level. The sample is composed of participants who were assigned to the pro- and counter-attitudinal treatments, for which the Facebook feed is observed in the two weeks following the intervention and where at least one post is observed. *p<0.1 **p<0.05 ***p<0.01

References

Allcott, H. and M. Gentzkow (2017). “Social Media and Fake News in the 2016 Election”. *Journal of Economic Perspectives* 31(2), 211–236.

Anderson, M. L. (2008). “Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects”. *Journal of the American Statistical Association* 103(484), 1481–1495.

Angrist, J. D. and I. Fernandez-Val (2013). “ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework”. *Advances in Economics and Econometrics - Tenth World Congress*. Ed. by D. Acemoglu, M. Arellano, and E. Dekel, 401–433.

Aronow, P. M. and A. Carnegie (2013). “Beyond LATE: Estimation of the Average Treatment Effect with an Instrumental Variable”. *Political Analysis* 21(04), 492–506.

Bakshy, E., S. Messing, and L. A. Adamic (2015). “Exposure to Ideologically Diverse News and Opinion on Facebook”. *Science* 348(6239), 1130–1132.

Chan, J. and W. Suen (2008). “A Spatial Theory of News Consumption and Electoral Competition”. *Review of Economic Studies* 75(3), 699–728.

DellaVigna, S. and E. Kaplan (2007). “The Fox News Effect: Media Bias and Voting”. *The Quarterly Journal of Economics* 122(3), 1187–1234.

Druckman, J. N. and M. S. Levendusky (2019). “What Do We Measure When We Measure Affective Polarization?” *Public Opinion Quarterly* 83(1), 114–122.

Flaxman, S. R., G. Sharad, and J. M. Rao (2016). “Filter Bubbles, Echo Chambers, and Online News Consumption”. *Public Opinion Quarterly* 80, 298–320.

Gosling, S. D., P. J. Rentfrow, and W. B. Swann (2003). “A Very Brief Measure of the Big-Five Personality Domains”. *Journal of Research in Personality* 37(6), 504–528.

Hainmueller, J. (2012). “Entropy Balancing for Causal Effects: a Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies”. *Political Analysis* 20(1), 25–46.

Heckman, J. J., S. Urzua, and E. J. Vytlacil (2006). “Understanding Instrumental Variables in Models With Essential Heterogeneity”. *The Review of Economics and Statistics* 88(3), 389–432.

Hortacsu, A., M. R. Wildenbeest, and B. De Los Santos (2012). “Testing Models of Consumer Search using Data on Web Browsing and Purchasing Behavior”. *American Economic Review* 102, 2955–2980.

Peterson, E., G. Shared, and S. Iyengar (2019). “Partisan Selective Exposure in Online News Consumption: Evidence from the 2016 Presidential Campaign”. *Political Science Research and Methods*, 1–17.

Rogowski, J. C. and J. L. Sutherland (2016). “How Ideology Fuels Affective Polarization”. *Political Behavior* 38(2), 485–508.

Schroeder, E. and D. F. Stone (2015). “Fox News and Political Knowledge”. *Journal of Public Economics* 126, 52–63.

Shane, F. (2005). “Cognitive Reflection and Decision Making”. *The Journal of Economic Perspectives* 19(4), 25–42.

Stone, D. F. (2020). "Just a Big Misunderstanding? Bias and Affective Polarization". *International Econ Review* (Forthcoming).

Suen, W. (2004). "The Self-Perpetuation of Biased Beliefs". *The Economic Journal* 114(495), 377–396.

Wager, S. and S. Athey (2018). "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests". *Journal of the American Statistical Association* 113(523), 1228–1242.

Yeager, D. S., J. A. Krosnick, L. Chang, H. S. Javitz, M. S. Levendusky, A. Simpser, and R. Wang (2011). "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples". *Public Opinion Quarterly* 75(4), 709–747.

Yuksel, S. (2018). "Specialized Learning and Political Polarization".