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

Ro'ee Levy*

November 20, 2019

Click here for the latest version

**Abstract**

Does social media increase the consumption of ideologically congruent news and 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 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 (a "filter bubble"). Together, these results suggest that social media algorithms are increasing polarization by limiting exposure to counter-attitudinal news.

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 throughout this project. I thank Eran Amsalem, Dirk Bergemann, Leonardo Bursztyn, Alex Coppock, Oren Danieli, Eduardo Fraga, Matthew Gentzkow, Matthew Graham, Alexey Makarin, Martin Mattsson, David Rand, Oren Sarig, David Schönholzer, Katherine Wagner and Mor Zoran for helpful comments and advice. I also thank seminar participants at Berkeley, Northwestern, 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. 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 ask two questions to study whether these concerns are warranted. First, how does social media affect news consumption habits? Second, what is the effect of social media news consumption on political opinions and polarization? I study these questions by conducting a large online field experiment randomizing exposure to news outlets on social media, and by collecting survey, browsing, and social media data.

To motivate the experiment, I first analyze the association between social media and online news consumption. I merge data on browsing behavior, voting, and news outlets to show that news sites visited through Facebook tend to be more extreme and to better match the consumer's ideology, compared to other news sites visited. While this association suggests that Facebook is affecting the news individuals consume, exogenous variation is required to identify a causal effect.

I recruited American Facebook users to the experiment using Facebook ads. After completing a baseline survey, participants were randomly assigned to either a liberal treatment, a conservative treatment, or a control group. Participants in the liberal treatment were asked to subscribe to four liberal outlets on Facebook, such as MSNBC. Participants in the conservative treatment were asked to subscribe to four conservative outlets, such as Fox News. Remarkably, half of the participants complied with the treatments by subscribing to at least one outlet. When individuals subscribe to an outlet by "liking" its Facebook page,[1] posts shared by the outlet are likely to subsequently appear in their Facebook feed. Individuals exposed to the posts can view the headlines directly in their feed, and they can click on links in the posts to consume the full news stories in the outlets' websites.

I designed the experiment to have high external validity. A nudge offering subscriptions to outlets is very common on social media. News outlets often promote their Facebook pages with similar nudges, and participants could have subscribed to any of these outlets, at no cost, without the intervention. Besides this 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 subscribed outlets appeared in the participants' Facebook feeds. Finally, participants decided whether to read, skip, or share specific posts. As a result, the treatment is almost identical to the experience of millions of Americans who subscribe to news outlets on Facebook.

[1] 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.

I estimate the effect of the intervention on news consumption, political opinions, and affective polarization, defined as negative attitudes toward the opposing political party. I focus on news consumption because social media is suspected to increase exposure to pro-attitudinal news and since an effect on news consumption is a key mechanism that can affect opinions and attitudes. Affective polarization is a primary outcome of interest since this measure of polarization has been increasing (Gentzkow, 2016; Iyengar and Krupenkin, 2018; Lelkes, 2016), and there are concerns over its implications for governance, accountability of elected officials and even labor markets Iyengar et al. (2019).

To measure participants' subscriptions to outlets on Facebook, I asked participants to log in to the survey using their Facebook account. I include in the sample 37,492 participants who completed the baseline survey and provided explicit permissions to observe their subscriptions. To measure exposure to news on the Facebook feed and visits to news sites, I developed a Google Chrome extension that collected this data for a subset of participants who were offered the extension and installed it. To estimate the effect on opinions and attitudes, I invited participants to an endline survey approximately two months after the intervention.

This paper has four main findings. First, exposure to news on social media substantially affects online news consumption. The intervention increased exposure to posts from the offered outlets in the participants' Facebook feeds. As a result, participants visited the news sites of the outlets they were randomly offered, even when the outlets did not match their ideology. Although social media is typically associated with pro-attitudinal news, individuals are willing to engage with counter-attitudinal news when it is made more accessible on social media.

Both the liberal and conservative treatments had a significant effect on the mean slant of news sites visited. The difference between the treatment-on-the-treated (TOT) effects of the liberal and conservative treatments is similar to the difference between sites visited in New York, a blue state, and South Carolina, a red state. Various economic theories explain why individuals optimally choose to consume news that matches their ideology.[2] However, I find that news consumption 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. This implies that algorithms determining the news supplied in social media feeds can substantially alter news consumption habits.

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 measure includes questions such as how participants feel toward their own party and the opposing party (i.e., a "feeling thermometer") and how they would feel if their son or daughter married a Democrat or Republican. The TOT effect of the counter-attitudinal

[2]This could occur since outlets sharing the consumers' ideologies 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.

treatment decreased the index by 0.06 standard deviations compared to the pro-attitudinal treatment.[3] I compare the result 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 party and the opposing party by 0.58 and 0.97 degrees on a 0-100 scale, respectively. For comparison, this measure of affective polarization increased by 3.83 degrees between 1996 and 2016.

I conduct back-of-the-envelope calculations estimating how counterfactual social media platforms would change affective polarization. I find that if individuals were exposed to an equal share of pro- and counter-attitudinal news on Facebook, the difference in their feelings toward parties would decrease by 3.76 degrees, almost the entire increase over the past two decades. This estimate should be interpreted cautiously since it assumes that the treatment affected attitudes only through its effect on the share of counter-attitudinal news on Facebook, it does not take into account general equilibrium effects, and it is based on the effect found over a two-month period.

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 opinion index, focusing on issues covered in the news during the study period, such as the March for Our Lives Movement or the Mueller investigation, is economically small, precisely estimated, and is not statistically significant. I do not find evidence for substantial heterogeneity in this effect.

Why did the treatments affect attitudes toward political parties but not political opinions? Participants may have learned to rationalize the opinion of the opposing party. I propose a model in which political opinions are a weighted average of multiple beliefs, and each party uses different weights according to its priorities. An attitude of an individual toward a party depends on the distance between the party's actual political opinion and the opinion the party would form based on the consumer's beliefs, but still using the party's weights. I show that the model is consistent with the results if the intervention affected beliefs on which the participants place low weights and the opposing party places high weights. Intuitively, participants may have learned the logic behind some of the arguments made by a party and thus developed a more positive attitude toward the party, even if they continued to disagree with its political opinion.

The paper's fourth finding is that Facebook's algorithm limits exposure to counter-attitudinal news. I decompose the gap between exposure to posts from the pro- and counter-attitudinal outlets offered in the experiment into three main explanations: (1) participants are less likely to subscribe to counter-attitudinal outlets; (2) Facebook's algorithm is less likely to supply posts from counter-attitudinal outlets, conditional on subscription; (3) participants decrease their Facebook usage in the counter-attitudinal treatment. While I find evidence for all three forces, the most important explanation for the gap in exposure is Facebook's algorithm.

[3] 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.

Combining the results paints a complicated picture. On the one hand, Facebook's algorithm filters counter-attitudinal news. While it is not possible to estimate the effect of specific posts filtered by the algorithm, I show that decreased exposure to counter-attitudinal news increases 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.

This paper contributes to the literature on social media and news consumption by providing the first experimental evidence that algorithms are increasing exposure to pro-attitudinal news. Papers in this literature typically estimate segregation in online news based on cross-sectional analysis of browsing behavior.[4] These papers usually lack social media data and cannot measure segregation *within* one's social media feed. Other studies focus on the effect of algorithms (e.g., Tufekci 2015). 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).[5] I advance the literature by generating experimental variation in subscriptions to outlets and collecting data on exposure to news in the Facebook feed. This allows me to decompose the mechanisms limiting exposure to counter-attitudinal news and show that a filter bubble does exist, i.e., that conditional on subscription, Facebook is more likely to expose individuals to news matching their ideology.

My findings contribute to the literature on social media, pro-attitudinal news consumption, and polarization by testing the effect of varying 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. Other papers on the topic have focused on the reduced-form effect on polarization and have shown that the Internet and Facebook may increase polarization (Allcott et al., 2019; Lelkes et al., 2015), but are probably not a primary driver in the rise of polarization (Boxell et al., 2018).[6] Since these papers focus on social media generally, they do not identify the causal effect of pro-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.

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; Martin and Yurukoglu

[4]See Barberá (2015), Flaxman et al. (2016), Gentzkow and Shapiro (2011), Guess (2018), and Peterson et al. (2018). Levy and Razin (2019) provide a broad review of the literature on segregation and echo chambers.

[5]See also Zuiderveen Borgesius et al. (2016). The conclusion is partly based on the analysis of Facebook data by Bakshy et al. (2015). They show that exposure to counter-attitudinal news is mostly limited by individual choices and not by algorithmic ranking. Their study focuses on posts shared by individuals' social networks, while I focus on posts shared by outlets individuals subscribe to, which are more likely to be associated with increased segregation.

[6]Other studies estimating the effect of social media on political behavior include Enikolopov et al. (2019), Miner (2015), Müller and Schwarz (2018), and Müller and Schwarz (2019). See Zhuravskaya et al. (2019) for a review.

2017) have found that individuals are persuaded by the news they consume. However, the results depend on the setting studied, and other papers have not found evidence that the political leaning of outlets has persuasive effects (e.g., Gentzkow 2006).[7] In many contexts, the "gold standard" for measuring causal effects is field experiments, since they combine the strong identification of lab experiments with higher external validity. However, with the notable exception of Gerber et al. (2009), there have been almost no field experiments randomly varying subscriptions to news outlets.[8]

Methodologically, this paper contributes to a growing literature conducting online media-related experiments (Allcott et al., 2019; Bail et al., 2018; Chen and Yang, 2019; Jo, 2018) by demonstrating how an experiment can exploit social media's existing infrastructure to gradually distribute news to participants in a natural setting. In contrast to similar experiments, participants were not asked to consume any content, they did not receive any notifications reminding them of the intervention besides the invitations to the endline survey, and they were not asked to continue complying with the treatment over time. Since the treatment occurs organically, the treatment effects in this paper were expected to be relatively small. Therefore, I collect a sample size that is an order of magnitude larger than most other related experiments, to precisely detect the primary effects studied.

The rest of the paper is organized as follows. In Sections 1 and 2, I provide background on Facebook and present exploratory analyses showing that social media is associated with pro-attitudinal news. Section 3 describes the experimental design, the datasets used, and the empirical strategy. Section 4 analyzes the effects of the experiment on news exposure, consumption, and sharing behavior, and Section 5 analyzes the effects on political opinions and affective polarization. Section 6 decomposes the factors increasing exposure to pro-attitudinal news on social media. Section 7 suggests a theoretical framework explaining the effects found on opinions and attitudes. The final section concludes.

[7] Papers estimating the effects of the media on political opinions and behavior also include Adena et al. (2015), Bursztyn and Cantoni (2016), Chiang and Knight (2011), DellaVigna et al. (2014), Durante et al. (2019), Enikolopov et al. (2011), Gentzkow et al. (2011), Larreguy et al. (2019), Okuyama (2019), and Snyder and Strömberg (2010). Other studies focus on the persuasive effects of fake news (e.g., Pennycook and Rand 2019), political advertisements (e.g., Broockman and Green 2014; Spenkuch and Toniatti 2018), or providing voters with information (Pande, 2011).

[8] Gerber et al. (2009) offer individuals random subscriptions to the print editions of the Washington Post or Washington Times before the 2005 Virginia gubernatorial election. They find no effect on knowledge, opinions, or turnout, but both newspapers increase the Democratic vote share. As the authors note, the most important limitation of the study is its relatively small sample size. In addition to recruiting a larger sample and focusing on social media, this paper collects data on news exposure allowing me to estimate the effect of exposure to news on opinion and attitudes.

Bail et al. (2018) randomize exposure to content from liberal and conservative bots on Twitter. The bots retweeted messages from various political twitter accounts, including elected officials, opinion leaders, non-profit groups, and media organizations. In contrast to Bail et al. (2018), this experiment is designed to be natural as possible. Therefore, participants were not encouraged to consume news and were offered subscriptions to major news outlets. I further discuss the studies in Section 5.1.

# 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,[9] 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, probably because Twitter data is more easily accessible (Guess et al., 2018; Tucker et al., 2018).

The most distinctive feature of Facebook is the news feed, where users scroll through a list of posts curated by Facebook. Posts in a user's feed are typically shared by the user's Facebook friends, shared by Facebook pages the user subscribes to, such as media outlets, or are sponsored posts (advertisements shared by pages to promote content). Facebook's algorithm determines which posts appear in the user's feed. 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.[10] 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). According to a recent report by the Reuters Institute, in 37 out of 38 middle and high-income countries surveyed, more than 20% of the population consumed news through Facebook weekly. In 25 countries, at least 40% consumed news through the platform weekly (Reuters Institute, 2019). Facebook probably directly affects the news exposure of more individuals than any other company.[11]

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.

# 2 Exploratory Analysis: Social Media and Pro-Attitudinal News

The increase in social media news consumption is more likely to affect public opinion if news consumed on social media is different from news consumed through other means. In this section, I show that Facebook is associated with greater consumption of pro-attitudinal news and that it is associated with more extreme news.

To estimate the association between Facebook and news consumption, I rely on three datasets. First, the 2017 Comscore WRDS Web Behavior Database Panel provides a sample of the browsing

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

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

[11]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).

behavior of approximately 93,000 US internet users.[12] Second, to determine the slant of each website, I rely on a dataset of news domains constructed by Bakshy et al. (2015).[13] The dataset defines the slant of 500 news sites according to the self-reported ideology of Facebook users sharing articles from these websites. The dataset correlates well with other datasets measuring the slant of outlets (e.g., Gentzkow and Shapiro, 2010). Throughout the paper, I refer to outlets in this dataset as *leading news outlets*. Third, as a proxy for individuals' ideology, I use 2008 zip code level voting data (Mummolo and Nall, 2016).[14] For more details on the processing of the browser and outlet datasets, see Appendices A.1 and A.2, respectively.

The data confirms that Facebook is an important source of news consumption. Overall, 7% of all visits to leading news sites in the sample are referred to by Facebook, and the share increases to 16% among individuals who visited at least one site through Facebook. Facebook is the second most common referral source for online news sites after Google.

Figure 1 shows a clear correlation between the consumers' ideology and the slant of the news they consume. More importantly, the slope of news consumed through Facebook (the solid blue line) is steeper than the slope of news consumed through other means (the dashed black line). This indicates that news consumed through Facebook tends to better match the consumers' ideology.[15] To construct this binned scatter plot, I calculate the mean slant of news sites visited for each individual in the sample when the websites were accessed through Facebook, i.e., the referring domain was facebook.com, and when the sites were accessed through all other means, e.g., through a search engine or by accessing the site directly. In the figure and throughout the paper, news slant is measured at the outlet level since this allows me to classify the news slant of any visit to a major news outlet and since this is the standard measure used in the literature.[16] To keep the sample constant across the news referral sources, I include in the sample only individuals who visit multiple news sites through Facebook and through other means.

While the mean slant of news consumed through Facebook is not extreme even in the most liberal and conservative zip codes,[17] the share of pro-attitudinal news substantially increases when news

[12] Comscore's data has been used to compare offline news to online news (Gentzkow and Shapiro, 2011), but to the best of my knowledge, it has not been used to analyze news consumed through social media.

[13] Guess (2018) and Peterson et al. (2018) also use this data to measure online segregation. This section differs from their work in the sample analyzed and in the in-depth analysis of news consumed through Facebook. The results of this section complement the conclusion by Peterson et al. (2018), who show that news consumed through social media is generally more segregated. In Section 6.1, I use data collected from the experiment to analyze the channels increasing segregation in social media news consumption.

[14] I use 2008 data since the data is available at the precinct level (Ansolabehere and Rodden, 2012), and thus is relatively precise when aggregated at the zip code level (Mummolo and Nall, 2016). The results in this section are robust to using 2016 county-level election data and 2017-2018 donation data.

[15] 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. Such a comparison may underestimate the association between social media and news consumption since the figure shows that individuals living in more moderate Republican and Democratic zip codes consume similar news through Facebook and through other means.

[16] In a separate work in progress, I compare segregation when it is measured at the outlet and article-level.

[17] The mean slant in the most conservative zip code decile is more liberal than the Wall Street Journal and the mean slant in the most liberal zip code decile is more conservative than the New York Times.

is consumed through Facebook. When individuals living in the most conservative zip code decile visit news sites through any means besides a link in Facebook, 10% of their visits are to very conservative sites, such as nationalreview.com, while when they visit news sites through Facebook, the figure goes up to 24%. Similarly, among individuals living in the most liberal zip code decile, 18% of news sites not visited through Facebook are very liberal, such as newyorker.com, compared to 27% of news sites visited through Facebook.

One possible implication of individuals consuming more pro-attitudinal news is that news consumption becomes more extreme. Indeed, Figure 2 presents the density of the mean slant of news consumption at the individual level and shows that Facebook is associated with substantially more extreme news sites. The difference is notable. When visiting news sites through Facebook, 27% of individuals consume news that is on average more conservative than the Wall Street Journal or more liberal than the New York Times. While among all other news sites visited, only 11% of individuals consume such partisan news.

Appendix D.1 tests the results presented in this section in a regression framework and shows that the results are robust to measuring slant at the site, individual or individual by month levels, to controlling for fixed effects and to different measures of Facebook usage.

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.[18] This section helps explain this public concern. While the Comscore data provides a large and diverse panel which is useful in understanding news consumption habits, the analysis of the data does not provide clean identification, nor can it shed light on the implications of increased pro-attitudinal news consumption. Therefore, an experiment generating random variation in exposure to news on social media is required.

## 3 Design and Data

### 3.1 Experimental Design

I recruited American adults to the experiment in February-March 2018 using Facebook ads.[19] 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, after completing baseline questions on media habits and political beliefs, participants were randomly

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

[19] In total, 978,628 people were shown the ads and 87,648 people clicked the link in the ads. The ads are presented in Appendix Figure A.1.

assigned to a liberal treatment, a conservative treatment or a control group, with the randomization blocked by participants' self-reported baseline ideology.[20] 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.

While I encouraged 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 the 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.[21] Facebook users often encounter the Like Page Button nudging them to subscribe to pages when the platform suggests pages they may be interested in or when outlets purchase ads promoting their Facebook page.

After participants subscribed to an outlet, posts from the outlet appeared in their Facebook feeds, according to Facebook's algorithm. The posts were observed among many other posts that usually appear in the feed. Since leading news outlets were chosen, the posts observed by the participants were also observed by other subscribers to these outlets (for example, more than 15 million people subscribe to the Facebook page of the New York Times). Participants were not asked to engage with any posts or read any news content. They were free to make their own media choices and decide 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 reminded of the intervention, experimenter effects are unlikely to play a large role in explaining the effects, at least compared to similar studies.

### 3.2 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 offered are Fox News, The National Review,

[20]Randomization was blocked to ensure balance across the main covariate expected to have prediction power when analyzing political outcomes. At the beginning of the survey, respondents were asked where they position themselves ideologically on a 7-point ideological scale from very liberal to very conservative, with an additional option of "I haven't thought about it much." Participants were blocked based on how they position themselves on the scale and when they answered the question. Each block is composed of three sequential participants who chose the exact same answer among the eight ideological scale options. The first participant in each block was randomly assigned to one of the three groups (the liberal treatment, conservative treatment or control group), the second participant was randomly assigned to one of the two remaining groups, and the third participant was assigned to the remaining group.

[21]The button was generated using Facebook's Page Plugin. It is a standard button that appears across Facebook and on many websites. It was not modified in any way for the experiment. An example survey page with the intervention is presented in Appendix Figure A.2.

The Wall Street Journal, and The Washington Times. The news outlets were chosen according to several criteria. First, they have a relatively clear political slant. Second, these outlets are relatively popular. Specifically, Fox News and the New York Times have the second and third most subscribers among all Facebook news pages. Finally, outlets of varying quality and extremity are included to provide participants with a diverse choice set and thus increase the likelihood that participants engage with at least one of the outlets offered.

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.[22] Appendix Table A.1 displays the full list of outlets, along with the number of times they were offered and the number of new subscriptions among participants who completed the endline survey.

Figure 3 shows that the most prominent men and women mentioned in posts shared by the primary outlets are political figures. This figure is based on all posts shared by the outlets on Facebook during the study period. I process the text in the posts and identify individuals using the Spacy Natural Language Processing algorithm. The figure presents the share of mentions of prominent individuals among all individuals mentioned by the outlets. Unsurprisingly, President Trump is the dominant figure mentioned. Some of the most important political stories during the study period can be observed in the figure: President Trump's alleged affair with Stormy Daniels, Robert Mueller's investigation into the Russian government's efforts to interfere in the 2016 presidential election, Scott Pruitt's ethics scandals, the March for Our Lives Movement led by Parkland Student David Hogg, and the negotiation with North Korea's leader, Kim Jong Un. The figure also demonstrates the difference between conservative and liberal outlets. 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.

### 3.3 Data Collection and Samples

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. 17,629 participants took the endline survey and constitute the *endline survey subsample*.

**Facebook Data** 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

[22] 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.

the list of outlets they subscribe to and the posts they share. Providing permissions was voluntary. Permissions could be revoked at any time and were revoked automatically approximately two months after participants logged in to a survey. Since data on baseline subscriptions was required to define the potential outlets for each participant, only participants who provided permissions to access their subscriptions are included in the baseline sample.[23]

Data on posts shared 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 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 or benefit to the reputation of the participant. Analyzing shared posts provides two additional advantages: their analysis does not depend on participants completing the endline survey, and there is no interaction between the experimenter and the participants when a post is shared. Approximately 92% of baseline participants provided access to the posts they shared for at least two full weeks following the intervention, and this subsample (the *access posts subsample)* is analyzed when estimating the effect on sharing behavior.

**Browser Data** Participants who completed the baseline survey using Google Chrome on a computer were asked to install a browser extension collecting data on the Facebook feed and news-related browsing behavior, in exchange for a small reward.[24] The offer was made toward the end of the survey, but before the intervention, to ensure take-up is not affected by the treatment. The extension was created for the unique requirements of this study. To protect participants' privacy, the extension only collects the URLs of news sites visited. 2,447 of the 8,080 participants who were offered the extension, installed it. In most of the analysis of this data, I focus on the 1,838 participants who kept the extension installed for at least two weeks (the *extension subsample*).[25]

The browser data is used to analyze news exposure by estimating how often posts from specific outlets appeared in the participants' Facebook feeds. I attribute a post to a news outlet if it was created by the outlet's Facebook page or contains links to the outlet's domain.[26] While the variation generated by the experiment is in subscriptions to the outlets' Facebook pages, I do not exclude news articles shared by the participants' Facebook friends, to accurately measure total exposure to news outlets on Facebook. The browser data is also used to estimate the effect on the news

[23]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 provided access to their subscriptions and revoked these subscriptions later are included in the baseline sample.

[24]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.

[25]Participants were only required to keep the extension installed for two days in order to receive the reward, but most participants kept the extension installed longer. Focusing on this subsample decreases the likelihood that behavior was affected by the installation of the extension. If the participants who kept the extension installed were bothered by the extension, they probably would have uninstalled it.

[26]In order to match URLs with news outlets, I first convert over ten million URLs to their final endpoint, allowing redirects along the way. This is required since many links on Facebook are based on URL-shortening services such as tinyurl.com.

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 browser data was only collected when participants used Facebook or browsed news sites on a computer while being signed into their Chrome account. In practice, individuals often use Facebook and browse the web on a mobile device or at work, where they may use a different browser. Therefore, all estimates of the number of posts individuals were exposed to in their feed or the number of news sites they visited are lower bounds for the actual intention to treat effect.[27]

**Subsamples** The datasets define three separate subsamples. To maximize power, throughout most of the analysis, I analyze each subsample separately according to the outcome analyzed. When analyzing the effect 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). Table 1 summarizes the subsamples, the datasets used and the main outcomes. For more details on the surveys, Facebook data, and browser extension data, see Appendices A.3, A.4, and A.5, respectively. Appendix Table A.2 presents descriptive statistics on each subsample.

## 3.4 Outcomes

### 3.4.1 Media

I measure subscriptions to outlets on Facebook, exposure to news outlets on Facebook, 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 times participants observed their potential liberal and conservative outlets in their feed, and the number of times they visited the websites of their potential liberal and conservative outlets. Second, I measure the mean slant of all *leading news outlets* participants engaged with. The slant of each outlet is based on Bakshy et al. (2015) and a higher value is associated with a more conservative slant. 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* as an additional measure of segregation in news consumption, calculated as the share of news from counter-attitudinal outlets among all news from pro-attitudinal and counter-attitudinal outlets.

[27]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 the extension subsample, 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.

### 3.4.2 Opinions and Attitudes

I analyze the effects of news exposure on two primary outcomes: political opinions and affective polarization. For both outcomes, the final index is composed by taking an average of all the index components and then standardizing the average with respect to the control group, so all effects are measured in standard deviations.The construction of both primary outcomes is defined in the study's pre-analysis plan, which is discussed in more detail in Appendix B.

**Political Opinions** I construct a political opinions index composed of twenty survey questions. The questions focus on domestic political issues and 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.[28] Each outcome variable is defined such that a higher value is associated with a more conservative opinion and then standardized by subtracting the control group mean and dividing by the control group's standard deviation.

**Affective Polarization** 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 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 Democrat 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 attitude toward the participant's party and attitudes toward the other party, a typical measure of affective polarization (Iyengar and Hahn, 2009). 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*).[29] An advantage of using a social-distance measure is that this measure does not correlate as strongly with other affective polarization measures, such as the feeling thermometer, and thus

[28] The full list of questions is presented in Appendix Figure A.8.

[29] Participants stating 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 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 can potentially bias the result if the treatment affected party affiliation. For example, if some 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. 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 any bias is expected to go against the direction of my findings. In Appendix Table A.10, I show that the results do not change when this measure is excluded from the index.

may be capturing an additional distinct aspect of polarization (Druckman and Levendusky, 2019). Each outcome variable is defined such that a higher value is associated with more polarization and then standardized.

There are concerns over how affective polarization is influencing political accountability, governance, accurate beliefs, and even product and labor markets. While the effects of affective polarization are beyond the scope of this paper, it is worthwhile to discuss at least two negative consequences in more detail. First, studies have shown that political behavior today is more likely to be driven by negative attitudes toward the opposing party rather than positive attitudes toward the voter's party (Iyengar and Krupenkin, 2018). Partly as a result, in recent elections, voters split their vote at record-low levels and were loyal to their party at record-high levels (Abramowitz and Webster, 2016). Consequently, elected officials may not be held accountable since they know voters on their side of the aisle will continue voting for them even if they do not represent them well or take positions violating democratic principles (Graham and Svolik, 2019). Second, studies have shown that affective polarization affects economic relations. For example, in experiments workers demand a higher reservation wage when their employer is from the other party (McConnell et al., 2018), and applicants affiliated with the minority party are less likely to receive a callback when sending their resume (Gift and Gift, 2015). The increase in affective polarization has not escaped the public. In a recent survey, 85% of Americans stated that the tone and nature of political debate have become more negative over the past several years, compared to only 3% who said that the tone has become more positive.[30]

### 3.5 Balance and Attrition

Table 2 presents descriptive statistics for participants in the baseline sample by treatment 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 re-defined treatment arms as well.

Similar to other opt-in panels (Yeager et al., 2011) and studies recruiting using Facebook ads (e.g., Allcott et al., 2019), 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%).[31] The gender composition

[30]Pew Research Center - American Trends Panel Wave 48, April-May, 2019.

[31]There are several likely explanations for why the sample is different from the US population. First, it is common that the samples in opt-in surveys are more liberal. Second, anecdotal evidence suggests that some conservatives who saw the ads did not want to participate in a survey conducted at Yale University. Finally, the ads automatically target people who were likely to complete the survey and not a random sample of the population. Still, the sample does not seem substantially different from samples of Mechanical Turk users, for example (Berinsky et al., 2012).

One advantage of the sample used in this experiment 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).

of participants and their average age is similar to the US population. Participants' self-reported exposure to news on Facebook that is in line with their views is similar to the Facebook population.

Tables 2 and Appendix Table A.3 also test for differential attrition among the three endline subsamples: participants who completed the endline survey, participants who provided access to posts they shared for at least two weeks, and participants who installed the browser extension for at least two weeks. While there are almost no differences in the attrition rates in the access posts and extension subsamples, more participants completed the endline survey in the control group (48%), compared to the liberal (45%) and conservative groups (45%).[32] 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.[33]

Appendix Tables A.4 and A.5 include only participants who completed the endline survey and show that despite the differential attrition, the treatment arms and control group are similar on observables. Most importantly, there is no differential attrition between the conservative and liberal treatments and no differential attrition between the pro-attitudinal 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 any concern over differential attrition.

### 3.6 Empirical Strategy

Throughout the paper, I use two main empirical strategies. 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, 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. I also estimate the effect of these treatments on engagement with the pro-attitudinal and counter attitudinal outlets, on the share of counter-attitudinal news participants engaged with, and on the congruence scale of news they engaged with.

[32] Table 2 also shows that 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 participants who initially provided access, before the intervention.

[33] 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.

### 3.6.1 Liberal and Conservative Treatments

I estimate the effects of the liberal and conservative treatments using the following ITT regression:

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

where $T_i^L \in \{0,1\}$ is whether participant $i$ is assigned to the liberal treatment, $T_i^C \in \{0,1\}$ is whether participant $i$ is assigned to the conservative treatment, and $X$ is a set of control variables.

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).[34] To increase power, when estimating the effect on political opinions, I 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. All regressions use robust standard errors unless noted otherwise. Appendix C.1 describes each control variable in detail.

### 3.6.2 Pro-Attitudinal and Counter-Attitudinal Treatments

I estimate the effects of the pro-attitudinal and counter-attitudinal treatments using the following ITT regression:

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

where $T^A \in \{0,1\}$ measures whether the participant was assigned to the counter-attitudinal treatment, defined as liberal treatment assigned to a participant with a conservative ideological leaning or a conservative treatment assigned to a participant with a liberal ideological leaning. $T^P \in \{0,1\}$ measures whether the participant was assigned to the pro-attitudinal treatment, defined as a liberal treatment assigned to a participant with a liberal ideological leaning or a conservative treatment assigned to a participant with a conservative ideological leaning.

I define the ideological leaning of participants according to the party they identify with or lean toward. If participants do not lean toward either party, the ideological leaning is defined according to their self reported ideology, and if the ideological leaning still cannot be determined, it is defined according to the candidate the participants preferred in the 2016 elections. Throughout

[34] In addition to mitigating concerns over differential attrition, comparing these treatments to each other, instead of comparing each treatment separately to the control group, leads to cleaner theoretical predictions. While the liberal outlets are clearly more liberal than the conservative outlets, it is not necessarily the case that the assigned liberal outlets are more liberal than news consumed in the control group (and similarly it is not clear if the conservative treatment exposes participants to more conservative content, compared to the control group). Finally, if individuals are persuaded by both treatments, comparing them to each other takes advantage of the experiment's design and provides more power than comparing each treatment to the control group.

the paper, any reference to liberal or conservative participants is based on this definition. I use this definition since it allows me to determine the ideological leaning of the vast majority of participants in the sample.[35] $X$ 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.

### 3.7 Compliance

Throughout the analysis, I first focus on ITT estimates since these estimates apply to the entire sample and do not require any additional assumption. To measure the effect of compliance, I also analyze TOT estimators by regressing the dependent variable on compliance and instrumenting compliance with the random treatment assignment. Any participant who subscribed to at least one of the outlets offered is considered a complier.[36] Since the intervention only offers new outlets to participants, defiers do not exist in this experiment.[37]

In the entire baseline sample, 56% of participants who were offered pro-attitudinal outlets complied with the treatment and subscribed to at least one outlet, compared to 45% of participants who were offered counter-attitudinal outlets.[38] The difference between the share of participants subscribing to pro- and counter-attitudinal outlets is relatively small, compared to other media experiments (e.g., Iyengar and Hahn 2009) and more in line with observational studies arguing that selective exposure is not high (e.g., Guess et al. 2018). One possibility is that moderates are driving these results. However, even among participants who say they are very liberal or very conservative, 44% of participants comply with the counter-attitudinal treatment.

Table 3 shows that the highest compliance is among liberals assigned to the liberal treatment and the lowest compliance is among conservatives assigned to the liberal treatment. Column (3) shows

[35] 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 results are robust to defining ideological leaning first by self-reported ideology, then by party affiliation and then by the supported candidate, which is how I originally defined ideological leaning in the pre-analysis plan. 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), and since the affective polarization questions focus on Republicans and Democrats. The effect on affective polarization is also robust to including only participants who identify with or lean toward the Democratic or Republican party.

[36] Subscriptions are 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.

[37] Defying the experiment would mean unsubscribing from an offered outlet, but participants are only offered outlets they are not already subscribed to, and therefore, a participant cannot defy the experiment. Since compliance is defined as subscribing to an outlet when it is offered, always-takers do not exist. When focusing on the two weeks following the intervention, an always-taker would be defined as a participant who would subscribe to the potential outlets, regardless of the intervention. In the control group, only 0.5% of participants subscribed to their potential liberal outlets, and only 0.2% subscribed to their potential conservative outlets in the two weeks following the intervention.

[38] 51% of participants complied with the liberal treatment and 50% complied with the conservative treatment.

that most participants still subscribed to the outlets two weeks after the intervention. Column (4)-(6) pool all offered outlets and participants and regresses subscription to an outlet on the outlet's perceived ideology. Participants were more likely to subscribe to outlets they are familiar with, to outlets with a perceived ideology similar to the participant's ideology, and to outlets that are perceived as more moderate. Appendix Table A.6 presents descriptive statistics on the compliers by treatment and shows that liberals, women, and participants who subscribe to more outlets on Facebook were generally more likely to comply with the treatments.

# 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 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-attitudinal outlets or four potential counter-attitudinal outlets in the two weeks following the intervention on the treatment. The control group is the reference group.[39] For example, the first row of the first panel shows that the pro-attitudinal treatment increased the number of subscriptions to pro-attitudinal outlets by 1.94 in the two weeks following the intervention, compared to the control group. The effect is significant as the entire confidence interval is greater than zero.

**Subscription to Outlets** The first panel of Figure 4 shows that two weeks after the intervention, participants assigned to the counter-attitudinal treatment subscribed to 1.43 new counter-attitudinal outlets on average. This figure is similar to the number of outlets participants immediately subscribed to when when offered outlets in the intervention (1.52, not shown in the figure) since few participants unsubscribed from these outlets within two weeks.

**Exposure to Posts** The second panel of Figure 4 shows that following the intervention, participants in the pro-attitudinal and counter-attitudinal treatments were exposed to 67 and 31 additional posts from the potential pro and counter-attitudinal outlets, respectively. While this effect is very strong relative to the control group, it is still a small share of the total number of post participants were exposed to in their Facebook feed.[40] To test whether participants noticed the change two months after the intervention, they were asked in the endline survey how often they saw posts

[39] I use linear regressions for ease of interpretation. Since the dependent variables are count data, Appendix Table A.7 shows that the results are robust to estimating the effects with Poisson regressions.

[40] Participants in the control group were exposed on average to approximately 2,166 posts in the two weeks following the intervention and 222 posts associated with leading news outlets.

from various outlets in their Facebook feed. Appendix Figure A.3 shows that participants in both treatments reported seeing more posts 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 confirms that the treatment had a noticeable effect on the subsample of participants who completed the endline survey and not only on participants who installed the extension.

**Browsing Behavior - News Sites Visited** Participants could consume news from the outlets they subscribed to by clicking on links included in the posts (approximately 81% of posts participants observed from the subscribed outlets contained links). The third panel of Figure 4 shows that the counter-attitudinal treatment increased visits to the websites of the counter-attitudinal outlets by 82%, an ITT effect of 1.37 additional visits over a baseline of 1.67. The pro-attitudinal treatment increased the number of visits to the websites of pro-attitudinal outlets by 22%, an ITT effect of 2.91 additional visits over a baseline of 12.94.[41]

**Sharing Behavior** The fourth panel of Figure 4 shows that participants not only consumed news from counter-attitudinal outlets when they started appearing in their feeds, they also shared the posts with their social network. To increase power, I also analyze the effect on posts shared using the entire subsample of participants who provided access to their posts. Appendix Figure A.4 confirms that both treatments had a significant effect on the number of posts shared. Complementing previous studies focusing on Twitter (Halberstam and Knight, 2016; Gorodnichenko et al., 2018), participants are much more likely to share posts from pro-attitudinal outlets. However, the relative effect on sharing counter-attitudinal posts is stronger. The fact that participants chose to share the posts suggests that they considered the posts important. Sharing the posts also implies that participants expanded the treatment to their social network.

One possibility is that participants shared posts while commenting negatively on their content. The second panel of Appendix Figure A.4 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 their overall news diet will not change. For example, individuals randomly of-

[41] The effects on pro-attitudinal outlets have wider confidence intervals than the effects on counter-attitudinal outlets due to the larger variation in engagement with pro-attitudinal content.

fered the New York Times may start consuming more articles from the outlet's website, but consequently decide to consume less news from the Washington Post, which offers a similar perspective. In this section, I focus on the conservative and liberal treatments since there are clear predictions on how these treatments would affect the mean news slant.[42]

**Exposure to Posts** The first panel of Figure 5 shows that when participants were randomly offered liberal or conservative outlets, their feed became substantially more liberal or conservative, respectively. The combined TOT effect of the liberal and conservative treatments equals approximately half of the gap between the slant of the feed of liberals and conservatives. The change in slant is important for two reasons. First, it provides a strong first stage, which is useful when analyzing the effect on political beliefs. Second, it provides an opportunity 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.

**Browsing Behavior - 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. Appendix Table A.8 shows that the effect is robust across various subsamples (e.g., when excluding participants who did not complete the endline survey).

The difference between the TOT effects of the liberal and conservative treatments is substantial and equals 19% of the difference between the slant of the browsing behavior of conservatives and liberals in the control group. Another way to understand the magnitude of this effect is to use the large Comscore panel to estimate the mean slant of the news individuals consume online in different states. 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, while the TOT effect of the conservative treatment would have shifted the online news consumption of a consumer in Pennsylvania to a news diet similar to an individual in South Carolina.[43]

I exploit the variation generated by the treatment to estimate the importance of the social media feed in determining the slant of news consumption. Table 4 shows that when the compliers' news feed becomes one standard deviation more conservative, the slant of the sites they visit becomes 0.31 standard deviations more conservative. The figure is calculated by instrumenting the slant of the posts observed in the Facebook feed with the treatment. In column (2), I focus only on news

[42]For example, if a conservative treatment had an effect, the slant of news participants engaged with should become more conservative, while it is not clear if a pro-attitudinal treatment should make the slant more conservative or liberal.

[43]For each individual in Comscore's database, the websites visited are matched with the leading news outlets to determine the individual's mean news consumption slant. Individuals who visited a news site only once 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 at least 700 Comscore panelists who visited news sites more than once.

sites visited through Facebook, rather than all news sites visited, and find that when the feed becomes one standard deviation more conservative, the slant of sites visited through Facebook becomes 0.71 standard deviations more conservative. These regressions rely on the exclusion restriction that the treatment only has an effect on the slant of sites visited through the slant of the Facebook feed. While the treatment is only expected to affect the slant of news sites through the Facebook feed, the treatment could affect the feed in many ways. I am condensing the feed, a complicated object with many dimensions, to a scalar, the mean slant of news an individual was exposed to. This scalar is strongly affected by the treatment and has intuitive economic meaning, but it is possible that other changes in the feed, not captured in this measure, could affect the news sites visited. Since the calculations rely on stronger assumptions than the ITT and TOT estimates, they should be interpreted cautiously.

Appendix Figure A.5a shows that the effects of the treatments on exposure to posts and browsing behavior decline over time but mostly remain significant. The figure displays the difference between the effects of the liberal and conservative treatments on the mean slant of all news participants were exposed to and sites they visited, based on participants who kept the extension installed for at least six weeks. The difference between the effects of the treatments in the sixth week following the intervention declines by 30% compared to the first week.

To test for spillovers across news outlets, I recalculate the effect of the treatments on the slant of exposure and news sites visited, excluding the eight potential outlets defined for each individual. Appendix Figure A.6 shows that the mean slant of news consumption is not substantially affected by the treatments when the potential outlets are excluded, implying that the experiment did not have strong crowd-in or crowd-out effects.

**Sharing Behavior** The third panel of Figure 5 shows that the slant of posts shared was affected by the treatment.[44] Appendix Figure A.5b shows that the effect on posts is persistent, by calculating the weekly effect of the conservative treatment, compared to the liberal treatment among participants who provided access to posts for at least six weeks.

Theoretically, this section shows that when modeling the demand for news, it is important to take into account incidental news consumption and not only active optimization by news consumers. It is possible that news is consumed incidentally due to the search costs of finding a preferred article or since consumers suffer from default bias and often prefer reading one of the first news articles that appear in their feed. In any case, this raises concerns regarding the power of social media companies in shaping news consumption habits. The extremely low effort required in order to subscribe to an outlet on social media, along with the effect of the feed on news consumption, imply that merely suggesting several new outlets can drastically change one's news diet. Such suggestions happen all the time. They can stem from companies attempting to maximize profits

[44] The effect is significant for the liberal treatment and when comparing the treatments to each other. Appendix Figure A.7 confirms the result among the entire subsample of participants who provided access to their posts.

by increasing user engagement, for example, when platforms suggest to users outlets they may be interested in. Subscription suggestions may also originate from entities attempting to maximize political goals, whether they are NGOs or political candidates purchasing ads on social media or even foreign agents promoting specific pages on Facebook in order to influence the American electorate.[45]

# 5 Findings: Opinions and Attitudes

## 5.1 Social Media News Exposure Does Not Strongly Affect Political Opinions

The top panel of Figure 6 shows that the treatment did not affect the political opinions index. The conservative treatment increased the political opinions index by 0.003 standard deviations compared to the liberal treatment. While the point estimate has the expected sign, the effect is minimal and is not statistically significant. The upper bound for the combined liberal and conservative treatment effects, based on a 95% confidence interval, is only 0.7% of the difference in political opinions between liberals and conservatives in the control group. Appendix Figure A.8 shows the effect on each component in the political opinions index. The effects are economically 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? One possibility is that participants consumed a small share of their news on Facebook. This explanation seems unlikely as participants who reported getting most of their news on social media were affected similarly to other participants (see Appendix D.2). A second possibility is that the null effect masks substantial heterogeneity. Perhaps some participants were persuaded by the outlets they consumed, while in other cases, there was a backlash effect and participants' opinions moved in the opposite direction of the treatment. For example, some conservatives may have become more conservative when exposed to liberal outlets due to motivated reasoning (Kunda, 1990; Taber and Lodge, 2006), while liberals became more liberal, and as a result, the average treatment effect is close to zero. In Appendix Figure A.9, I test this hypothesis by estimating the effect of the interaction of ideology and treatment on the political opinions index. I find no evidence for a backlash effect.

Finally, to verify that the magnitude of the estimate is economically small, I use the treatment as an instrument for the slant of participants' Facebook feed to estimate how the feed's slant affects opinions. I find that if the Facebook feed of a liberal became similar to the Facebook feed of a conservative over a two month period (or vice versa), the opinions of the liberal would move only

[45] For example, many ads purchased by Russian organizations in their attempt to influence the 2016 election promoted Facebook pages. US House of Representatives - Permanent Select Committee on Intelligence. Exposing Russia's Effort to Sow Discord Online: The Internet Research Agency and Advertisements. https://intelligence.house.gov/social-media-content/

3% in the direction of the opinions of the conservative, and I can reject an effect larger than 7%.[46]

Interestingly, the results differ from a recent study by Bail et al. (2018), which exposes individuals to different views on Twitter and finds evidence for a backlash effect. Besides possible differences between Facebook and Twitter, two major differences in the design of studies can explain the differing results. First, Bail et al. expose individuals to a bot retweeting counter-attitudinal *views* on Twitter. Individuals plausibly become more upset when they are exposed to tweets of opposing elected officials and opinion leaders, compared to posts from counter-attitudinal news outlets. Second, Bail et al. provided participants with monetary incentives to continuously follow the new information they were exposed to. Social scientists have long criticized the generalizability of forced exposure media experiments since the effects found may be concentrated among individuals who would not consume the content under different circumstances (Arceneaux and Johnson, 2013; Bennett and Iyengar, 2008; Hovland, 1959). In this case, the monetary incentives may have encouraged participants to continue consuming tweets, which upset them and increased their partisan hostility, but that they would have stopped consuming outside the experimental setting.

### 5.2 Exposure to Counter-Attitudinal News Decreases Affective Polarization

The bottom panel of Figure 6 shows that the counter-attitudinal treatment decreased affective polarization 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 increased exposure to pro-attitudinal news are not misguided.

Figure 7 presents the results of regressions estimating the effect for each measure in the affective polarization index separately. The effect is especially pronounced for the question asking participants how difficult they find it to see things from each party's point of view, and the effect is weakest when participants are asked if parties have good ideas. While the experiment is underpowered to detect a statistically significant effect for each coefficient separately, in all cases, the pro-attitudinal treatment is associated with a more polarized outcome, and the coefficients are similar in magnitude to the point estimate of the index measure.

Appendix Tables A.9, A.10, and A.11 show that the result is robust to not controlling for covariates, excluding each of the affective polarization measures, and excluding participants who already subscribed to at least one of the primary outlets before the intervention. Appendix Table A.12 shows that an effect is detected even when focusing only on the subsample of participants who completed the endline survey and installed the extension. The effect is stronger among this group, partially due to higher compliance rates. Appendix D.3 shows that an effect is detected when the

[46] The calculation is based on the following steps: I first regress the political opinion index on the slant of posts in the Facebook feed, where the slant of posts is instrumented with the treatment. I then calculate the difference in the slant of posts in the Facebook feed of liberals and conservatives in the control group. I multiply the difference by the effect of the slant of posts on political opinions to estimate how opinions would have changed if the feeds of liberals and conservatives had the same slant. I compare the final effect to the actual difference in opinions between liberals and conservatives in the control group.

regressions are reweighted to match populations means in self-reported ideology, party affiliation, gender, age, and a baseline measure of affective polarization. Appendix Figure A.10 compares each treatment to the control group and shows that most of the difference between the pro- and counter-attitudinal treatments stems from the effect of the counter-attitudinal treatment.

Appendix D.2 does 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. One exception is that the treatment seems to be stronger for participants who were less polarized in baseline according to the feeling thermometer question. Based on eight pre-registered survey questions, in Appendix D.4, I test if the effect on polarization could be explained by changes in the knowledge of participants. I do not find evidence for strong effects on knowledge.

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 more balanced news exposure. All of the estimates are based on the effects over a two-month period. It is possible that with longer exposure to different news, the effects would have been stronger.

To compare the results to existing benchmarks, I focus on the feeling thermometer question, which is asked regularly in the American National Election Surveys. The ITT effect of the counter-attitudinal treatment decreases the difference between the feeling toward the participant's party and the opposing party by 0.58 degrees (on a 0-100 scale), and the TOT effect reduced the gap by 0.97 degrees. For comparison, in the past 20 years, the feeling thermometer measure increased by 3.83 degrees.[47] An additional point of comparison is a recent experiment which found that disconnecting from Facebook decreases the feeling thermometer measure by 2.09 degrees (Allcott et al., 2019). Hence, one way to interpret these results is that approximately 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] Two recent survey experiments decreased the feeling toward the opposing party by 3-6 degrees by priming participants' identify as Americans (Levendusky, 2017) or showing participants a news story highlighting a warm relationship between Republican and Democratic leaders (Huddy and Yair, 2019). It is unsurprising that the result of a field experiment, where participants choose which news stories to consume, is substantially smaller than survey experiments where participants are assigned to read certain stories and then asked related questions.

[47] To compare responses consistently over time, I only use face to face samples and weight the data accordingly.

[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. (2019) conduct the study over a one-month period 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. (2019) screen respondents who report using Facebook for less than 15 minutes per day or who are not willing to deactivate Facebook for 24 hours.

To estimate the effect of exposure to pro- or counter-attitudinal news on polarization, I focus on participants who both 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, where a higher value in the scale is associated with more conservative news exposure for conservative participants and more liberal exposure for liberal participants. 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.

Appendix Table A.13 shows that an increase of one standard deviation in the share of exposure to counter-attitudinal news decreases affective polarization by 0.12 standard deviations. Similarly, an increase of one standard deviation in the congruence scale decreases affective polarization by 0.10 standard deviations. One challenge in studying affective polarization based on non-experimental survey data (e.g., Kelly Garrett et al., 2014; Tsfati and Nir, 2017) is determining whether the correlation between pro- and counter-attitudinal 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. The effects of news exposure on affective polarization are approximately 24%-31% 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.

Finally, I use two back-of-the-envelope calculations to estimate how affective polarization would have changed if individuals were exposed to more balanced news on Facebook. In Appendix Table A.14, I find that if the Facebook feed had an equal share of pro- and counter-attitudinal news, the difference between the feeling toward one's party and the opposing party would decrease by 3.76 degrees, a change similar to the entire increase in polarization over the past two decades. For this calculation, I rely on the difference between the share of exposure to counter-attitudinal news in the control group, 17%, and an exposure of 50%. I estimate the effect of this difference on affective polarization based on the IV regressions described above. The estimation does not rely on out-of-sample predictions as the share of counter-attitudinal news was greater than 50% for a non-negligible share of participants in the counter-attitudinal treatment.

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 to a similar share of counter-attitudinal outlets in their Facebook feed and when visiting news sites not through Facebook. I find that the feeling thermometer outcome would decrease by 0.46 degrees. I repeat both counterfactuals using the congruence scale as the independent variable and the results are similar. These back-of-the-envelope calculations should be interpreted carefully since they do not take into account general equilibrium effects.[49] Nevertheless, they suggest that the Facebook feed can play an important role in amplifying or mitigating polarization.

# 6 Findings: Why is Social Media Associated with Pro-Attitudinal News?

Since the previous section shows that exposure to pro-attitudinal news affects partisan hostility, it is crucial to understand what influences the news individuals are exposed to on social media.

## 6.1 News Sites Visited Through Social Media

In this section, I analyze data on the browsing behavior and social media feed of participants in the experiment's control group. Figure 8a confirms that news consumed through Facebook is more likely to be extreme and pro-attitudinal (as shown using a different dataset in Section 2).[50]

Next, I focus only on news consumed through Facebook and compare two mechanisms for how Facebook increases segregation in online news consumption: do individuals consume more pro-attitudinal news due to homophily in social networks (an "echo chamber" effect)? Or is there increased consumption of pro-attitudinal news due to the abundance of accessible, free media options on social media allowing consumers to personalize their news feed? I test the first theory in the first row of Figure 8b, which presents the distribution of the slant of news sites visited through links shared by Facebook friends. The second row in the figure tests the latter theory by presenting the slant for news sites visited through links shared by Facebook pages. In the control group, approximately 60% of visits to news sites are through Facebook pages.[51]

I find that sites visited through Facebook pages are driving most of the increased consumption of extreme news matching the consumer's ideology. When news sites are not visited specifically through Facebook, approximately 20% of news sites visited are extreme pro-attitudinal sites. When participants visit news sites through posts shared by their Facebook friends, the share increases to 24%, and when they visit news sites through posts shared by Facebook pages they subscribe to, the share increases to 29%. This suggests that in order to understand why pro-attitudinal

[49] 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 counterfactuals since participants in the counter-attitudinal treatment did use Facebook less often (as discussed in Section 6.2). However, with network effects, the decrease in Facebook use could be greater. In addition, in the second counterfactual, indirect effect of Facebook on browsing behavior are not taken into account. I equate the share of counter-attitudinal news in the Facebook feed to the share in all news sites not visited directly through Facebook. However, it is likely that the feed affects some of these visits. For example, an individual may click on a Facebook link, visit an outlet and then click on an additional link in the outlet's website. I am not excluding the second link from my calculation even though the individual may not have visited it if the Facebook feed was more balanced. The results of the second counterfactual would probably be slightly larger if indirect visits to websites through Facebook were taken into account.

[50] Extreme pro-attitudinal news outlets are defined as outlets in the most liberal decile for liberal participants or outlets in the most conservative decile for conservative participants.

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

news is consumed through social media, it is important to study the forces determining which pages appear in the social media feed.

### 6.2 Exposure to News on Social Media

This section decomposes the gap in exposure to posts shared by the pages of 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 (“selective exposure”); Facebook’s algorithm supplies fewer posts from counter-attitudinal outlets, conditional on participants subscribing to them (the “filter bubble”); and participants use Facebook less often when offered counter-attitudinal outlets. The decomposition exercise is based on the following framework:

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

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 ($P_{ij}$), and the total number of posts individual $i$ observed on Facebook ($U_i$). I decompose the gap in exposure using the following formula:

$$\Delta E = \underbrace{S_\Delta * P_C * U_C}_{\text{Subscriptions}} + \underbrace{S_C * P_\Delta * U_C}_{\text{Platform Algorithm}} + \underbrace{S_C * P_C * U_\Delta}_{\text{Platform Usage}} + \underbrace{S_\Delta * P_\Delta * U_C + S_C * P_\Delta * U_\Delta + S_\Delta * P_\Delta * U_C + S_\Delta * P_\Delta * U_\Delta}_{\text{Combinations}} \quad (3)$$

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

I calculate $S_C$ and $U_C$ according to the mean number of new subscriptions and the total number of posts participants were exposed to, respectively, in the counter-attitudinal treatment. I include only posts shared directly by the outlets, to isolate the effects of subscriptions, the algorithm, and platform usage, from any effect of Facebook friends sharing specific articles. I estimate $S_\Delta$ and $U_\Delta$ by regressing the number of subscriptions and total exposure to posts on whether participants were assigned to a pro-attitudinal treatment.

To estimate $P_\Delta$ and $P_C$, 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 supplied from a 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 calculation is discussed in more detail in Appendix D.5. The appendix also discusses alternative estimations and exposure to outlets not included in the experiment.

Figure 9 shows that the strongest force driving exposure to pro-attitudinal news in the experiment is the algorithm. This provides evidence that a filter bubble exists in social media. 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. 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 I interpret it as suggestive evidence that participants use social media less often when they are exposed to more news they disagree with. This could explain why personalization is leading to segregation online–when consumers are exposed to more counter-attitudinal news, they seem to decrease their Facebook usage, and therefore, platforms may have an incentive to filter counter-attitudinal news in order to maximize engagement. This result raises the question of whether the algorithm also personalizes content within an outlet, and show conservatives relatively conservative posts shared by an outlet and liberals relatively liberal posts shared by the same outlet. In Appendix D.5.4, I find no evidence for within-outlet personalization.

This section does *not* suggest that Facebook's algorithm intentionally increases segregation by ranking posts according to whether they share the user's beliefs, or that the interaction of the slant of an outlet and ideology of a user has a causal effect on whether the algorithm places a post in the feed. The platform ranks a post based on many signals. These signals likely include the consumer's past behavior and engagement with the page, her social network and possibly other pages she subscribes to. Based on these factors, the algorithm determines that posts from specific outlets are less likely to interest a user, and these outlets tend to be counter-attitudinal. Still, the effect of personalization on news exposure is an important departure from how news was supplied and consumed in the past. Until recently, the engagement of an individual or her social network with news (e.g., the articles she read in the newspaper or news channels her friends choose to watch) did not affect the news supplied to the individual.

While I focus on Facebook, the logic likely applies to other platforms that personalize content as well. For example, since 2016, Twitter has been ranking tweets according to how interesting and engaging they would be for a specific user, and the highest-scoring tweets are shown at the top of a user's timeline. If Twitter's ranking algorithm is similar to Facebook, this may increase exposure to pro-attitudinal news.[52] Furthermore, major news outlets have also started to personalize their

[52]Factors taken into account when determining the ranking of 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.

websites and the articles they suggest to their customers.[53]

## 7 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 $\theta_{ik}^0$ 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, $\gamma_i$, is a weighted average of $K$ beliefs:

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

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 will 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.[54] 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 $\theta_k^*$ is the true state of the world and $h_{jk}$ is the precision of the signal received. Media outlets act as delegates for their consumers by covering issues according to the weights their consumers place on them.[55] 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 3 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

[53] 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.

[54] 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.

[55] Delegation has long been suggested as an explanation for why consumers prefer like-minded news (Chan and Suen, 2008; Downs, 1957; Suen, 2004).

differential coverage: for every post from a conservative outlet mentioning the word "environment," 2.73 posts mentioned the word "economy," while for liberal outlets, the ratio was 0.83.[56]

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).[57]

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 $i$:

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

where $A_{ip}$ is the attitude of individual $i$ toward party $p$, $\gamma_p$ 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:** $A_{ip} = g(\gamma_p - \sum_k w_{ik}\theta_{ik})$

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

$$\Delta A_{ip} = g(\gamma_p - \gamma_i^1) - g(\gamma_p - \gamma_i^0) = g(\sum_k w_{ik}(\theta_{ik}^1 - \theta_{ik}^0)) \tag{4}$$

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.

[56] 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.

[57] 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.

**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.[58] The change in affective polarization following an update to the consumer's beliefs is:

$$\Delta A_i = g(\gamma_p - \sum_k w_{pk}\theta^1_{ik}) - g(\gamma_p - \sum_k w_{pk}\theta^0_{ik}) = g(\sum_k w_{pk}(\theta^0_{ik} - \theta^1_{ik})) \tag{5}$$

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.[59]

This theory is consistent with the results of the experiment if the consumers updated beliefs on which they place zero weights, but at least one of the parties places positive weights.[60] This would result in consumers' political opinions remaining constant, but attitudes toward parties changing.[61]

[58] 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.

[59] Stone (2019) 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.

[60] 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.3 shows that participants assigned to the counter-attitudinal treatment were more likely to say that they modified their views in the past two months about a political or social issue because of something they saw on social media, compared to participants assigned to the pro-attitudinal treatment.

[61] 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

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 4). 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 5). Therefore, pro-attitudinal outlets are more likely to affect an individual's attitudes toward her own party, while counter-attitudinal outlets are more likely to affect attitudes toward the opposing party.

Table 5 shows that attitudes toward the opposing party are indeed more likely to be affected by exposure to counter-attitudinal outlets, supporting the theory that affective polarization is due to perceived unreasonable opinions. This result also contradicts a third hypothesis, which argues that affective polarization is increasing because outlets are covering the opposing party more negatively over time (Iyengar et al., 2019).

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. This section provides evidence ruling out several theories: it is unlikely that affective polarization simply increases due to a growing difference in political opinions or that affective polarization is only explained by increased negative media coverage. 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,[62] and more research is needed to pinpoint the precise mechanisms explaining how affective polarization evolves.

---

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).

[62]For example, Mason (2015) explains that partisan bias may increase without changes in position extremity as a result of stronger partisan identity. The counter-attitudinal treatment may have mitigated tribalism, which could have decreased affective polarization. Field experiments have found that strengthening partisan behavior affects political beliefs (Gerber et al., 2010). However, this is unlikely to explain the experiment's results as Appendix Figure A.3 shows that the treatments did not significantly affect party affiliation (the point estimate of the effect on Democratic party affiliation has the predicted sign, and I cannot reject that this treatment had a small effect on affiliation with the party).

There is evidence that Americans perceive members of the opposing party as more extreme than they are (Yudkin et al., 2019). It is possible that attitudes changed because participants learned the opposing party is not as extreme as they thought. This theory is consistent with a study by Orr and Huber (2018) who find that negative feelings toward individuals from the opposing party decrease when information is provided about the individuals' policy position. In Appendix Figure A.3, 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.

# 8 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.

The study shows that individuals are willing to engage with new viewpoints. Participants in the experiment not only subscribed to counter-attitudinal news outlets, they also consumed and shared news from those outlets. However, Facebook's algorithm limited exposure to counter-attitudinal news. This "filter bubble" effect did not exist until recent technological developments enabled automated personalization of news content. Personalization may have stronger impacts in the future due to growth in online news consumption and advances in machine learning algorithms customizing news exposure.

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, and the Internet generally, increase polarization (Allcott et al., 2019; Lelkes et al., 2015). On the other hand, it seems that individuals are not so 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 (Gentzkow, 2016; Lelkes, 2016; Mason, 2015). This suggests that a more segregated news environment may partially explain the increase in affective polarization over the past several decades.[63] In any case, affective polarization could influence policy even when political opinions remain stable, by decreasing trust in governance, impeding bipartisanship, and increasing voters' party loyalty.

The experiment has high external validity when it comes to analyzing actual behavior on Facebook in 2018, as supply and consumption of news occurred just as they do when individuals subscribe to any other outlet on Facebook. Still, as with other randomized control trials, one should be careful when extrapolating the results of the study. 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. Future studies can test whether the results hold in a different context. Specifically, I can only estimate the effects on media outcomes when participants use their computer, and it would be interesting to test whether they hold when news is consumed on a smartphone. In addition, I focus on outlets with an ideological slant. Future studies could estimate the effect of popular moderate outlets, such as USA Today, on opinions and attitudes. Finally, lab experiments

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

measuring individuals' beliefs, the weights they place on issues, and their attitudes toward parties can directly test affective polarization theories.

This study has important policy implications. Even though social media platforms are associated with news matching the consumer's ideology, they also provide an opportunity to expose individuals to more counter-attitudinal news. Suggestions include making algorithms more transparent,[64] nudging users to diversify their feed, and modifying algorithms such that they 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 since individuals are willing to engage with other viewpoints. Social media platforms have recently started rolling out features that could potentially diversity the users' feeds, and thus may have positive externalities by decreasing polarization.[65]

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

[64]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).

[65]For example, in 2017, Facebook implemented a feature that shows users articles related to a post in their feed from additional outlets. In August 2018, Twitter announced that it will allow users to follow topics in addition to specific accounts. In October 2019, Facebook started piloting Facebook News, a dedicated place for news curated by both algorithms and journalists. Facebook highlighted that news will be personalized. Hence, there is a risk that this new feature will expose consumers to mostly pro-attitudinal news. However, the company also mentioned that it will include diverse voices in Facebook News, and therefore, the feature can potentially decrease polarization.

## References


Abramowitz, A. I. and S. Webster (2016). The Rise of Negative Partisanship and the Nationalization of U. S. Elections in the 21st Century. *Electoral Studies 41*, 12–22.

Adena, M., R. Enikolopov, M. Petrova, V. Santarosa, and E. Zhuravskaya (2015). Radio and the Rise of the Nazis in Prewar Germany. *The Quarterly Journal of Economics 130*(4), 1885–1939.

Allcott, H., L. Braghieri, S. Eichmeyer, and M. Gentzkow (2019). The Welfare Effects of Social Media. *NBER Working Paper*.

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. In D. Acemoglu, M. Arellano, and E. Dekel (Eds.), *Advances in Economics and Econometrics - Tenth World Congress*, pp. 401–433.

Ansolabehere, S. and J. Rodden (2012). Harvard Election Data Archive.

Arceneaux, K. and M. Johnson (2013). *Changing Minds or Changing Channels? Media Effects in the Era of Expanded Choice*. University of Chicago Press.

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.

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.

Barberá, P. (2015). How Social Media Reduces Mass Political Polarization. Evidence from Germany, Spain, and the U.S.

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.

Broockman, D. E. and D. P. Green (2014). Do Online Advertisements Increase Political Candidates' Name Recognition or Favorability? Evidence from Randomized Field Experiments. *Political Behavior 36*(2), 263–289.

Bursztyn, L. and D. Cantoni (2016). A Tear in the Iron Curtain - The Impact of Western Television on Consumption Behavior. *The Review of Economics and Statistics 98*(1), 25–41.

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., R. Enikolopov, V. Mironova, M. Petrova, and E. Zhuravskaya (2014). Cross-Border Media and Nationalism: Evidence from Serbian Radio in Croatia. *American Economic Journal: Applied Economics 6*(3), 103–132.

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

Downs, A. (1957). *An Economic Theory of Democracy*. New York: Harper and Row.

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

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.

Enikolopov, R., M. Petrova, and E. Zhuravskaya (2011). Media and Political Persuasion: Evidence from Russia. *American Economic Review 101*(7), 3253–3285.

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

Gentzkow, M. (2006). Television and Voter Turnout. *The Quarterly Journal of Economics 121*(3), 931–972.

Gentzkow, M. (2016). Polarization in 2016. *Toulouse Network for Information Technology Whitepaper*, 1–22.

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

Gentzkow, M. and J. M. Shapiro (2010). What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica 78*(1), 35–71.

Gentzkow, M. and J. M. Shapiro (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. In *Handbook of Media Economics, 1B*, Volume 1, pp. 623–645. Elsevier B.V.

Gerber, A. S., G. A. Huber, D. Doherty, and C. M. Dowling (2012). Disagreement and the Avoidance of Political Discussion: Aggregate Relationships and Differences across Personality Traits. *American Journal of Political Science 56*(4), 849–874.

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.

Gift, K. and T. Gift (2015). Does Politics Influence Hiring? Evidence from a Randomized Experiment. *Political Behavior 37*(3), 653–675.

GlobalWebIndex (2018). *Flagship Report 2018*.

Gorodnichenko, Y., T. Pham, and O. Talavera (2018). Social Media, Sentiment and Public Opinions: Evidence from #BREXIT and #USELECTION. *NBER Working Paper*.

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.

Graham, M. and M. Svolik (2019). Democracy in America? Partisanship, Polarization, and the Robustness of Support for Democracy in the United States.

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.

Guess, A. M. (2018). (Almost) Everything in Moderation: New Evidence on Americans' Online Media Diets.

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.

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.

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*(August), 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.

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

Huddy, L. and O. Yair (2019). Reducing Affective Partisan Polarization: Warm Group Relations or Policy Compromise?

Iyengar, S. and K. S. Hahn (2009). Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use. *Journal of Communication 59*(1), 19–39.

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.

Kelly Garrett, R., 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.

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

Kunda, Z. (1990). The Case for Motivated Reasoning. *Psychological Bulletin 108*(3), 480–498.

Larreguy, H. A., J. Marshall, and J. M. J. Snyder (2019). Publicizing Malfeasance: When the Local Media Structure Facilitates Electoral Accountability in Mexico.

Lelkes, Y. (2016). The Polls-Review: Mass Polarization: Manifestations and Measurements. *Public Opinion Quarterly 80*, 392–410.

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. S. (2017). Americans, Not Partisans: Can Priming American National Identity Reduce Affective Polarization? *Journal of Politics 80*(1), 59–70.

Levy, G. and R. Razin (2019). Echo Chambers and Their Effects on Economic and Political Outcomes. *Annual Review of Economics 11*(1), 303–328.

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.

McConnell, C., Y. Margalit, N. Malhotra, and M. Levendusky (2018). The Economic Consequences of Partisanship in a Polarized Era. *American Journal of Political Science 62*(1), 5–18.

Miner, L. (2015). The Unintended Consequences of Internet Diffusion: Evidence from Malaysia. *Journal of Public Economics 132*, 66–78.

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

Müller, K. and C. Schwarz (2018). Fanning the Flames of Hate: Social Media and Hate Crime.

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

Mummolo, J. and C. Nall (2016). Why Partisans Do Not Sort: The Constraints on Political Segregation. *The Journal of Politics 79*(1), 45–59.

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.

Pande, R. (2011). Can Informed Voters Enforce Better Governance? Experiments in Low-Income Democracies. *Annual Review of Economics 3*(1), 215–237.

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

Pennycook, G. and D. G. Rand (2019). Lazy, Not Biased: Susceptibility to Partisan Fake News is Better Explained by Lack of Reasoning than by Motivated Reasoning. *Cognition 188*, 39–50.

Peterson, E., G. Shared, and S. Iyengar (2018). Echo Chambers and Partisan Polarization: Evidence from the 2016 Presidential Campaign. *Working Paper*.

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.

Reuters Institute (2019). *Digital News Report 2019*. University of Oxford.

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.

Snyder, J. and D. Strömberg (2010). Press Coverage and Political Accountability. *Journal of Political Economy 118*(2), 355–408.

Spenkuch, J. L. and D. Toniatti (2018). Political Advertising and Election Results. *Quarterly Journal of Economics 133*(4), 1981–2036.

Stone, D. F. (2019). 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.

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

Taber, C. S. and M. Lodge (2006). Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science 50*(3), 755–769.

Tsfati, Y. and L. Nir (2017). Frames and Reasoning: Two Pathways from Selective Exposure to Affective Polarization. *International Journal of Communication 11*(1), 301–322.

Tucker, J. A., A. Guess, P. Barbera, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan (2018). *Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature*. Hewlett Foundation.

Tufekci, Z. (2015). Algorithmic Harms beyond Facebook and Google: Emergent Challenges of Computational Agency. *Journal on Telecommunications & High Tech Law 13*(23), 203–216.

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

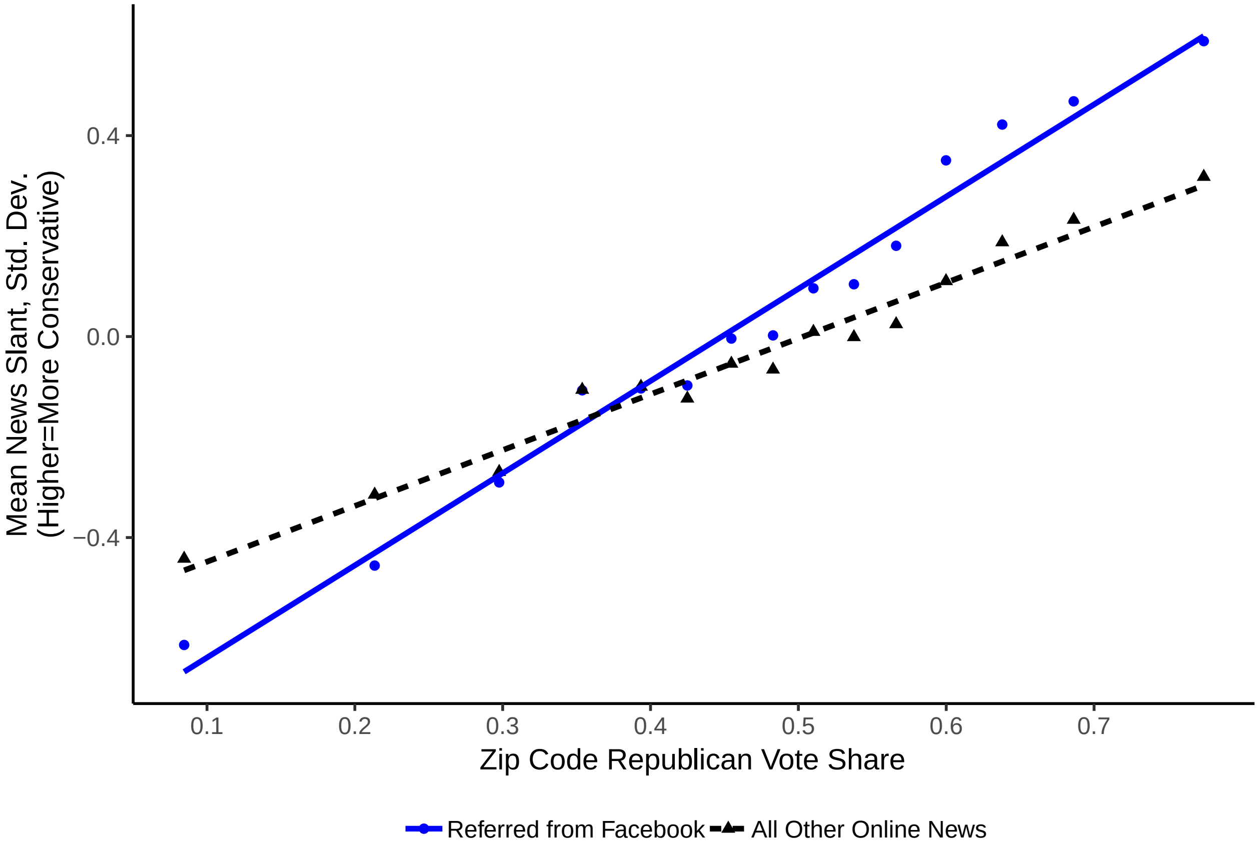
Yeager, D. S., J. A. Krosnick, L. Chang, H. S. Javitz, M. S. Levendusky, A. Simpser, and R. Wang (2011). Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples. *Public Opinion Quarterly 75*(4), 709–747.

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 (2019). Political Effects of the Internet and Social Media. *Annual Review of Economics*.

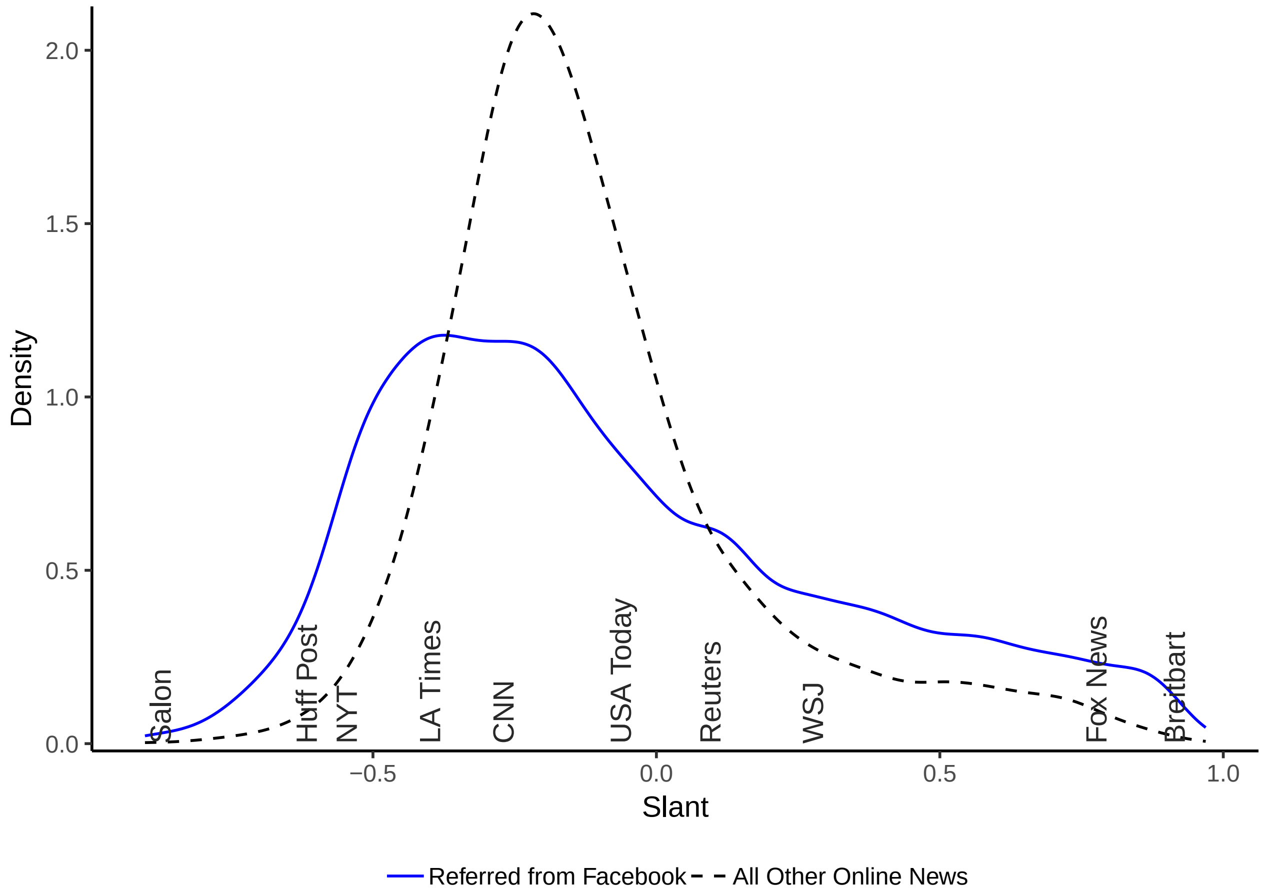
Zuiderveen Borgesius, F., D. Trilling, J. Möller, B. Bodó, C. De Vreese, and N. Helberger (2016). Should We Worry about Filter Bubbles? *Internet Policy Review 5*(1), 1–16.

Figure 1: Ideology and Slant of News Consumption



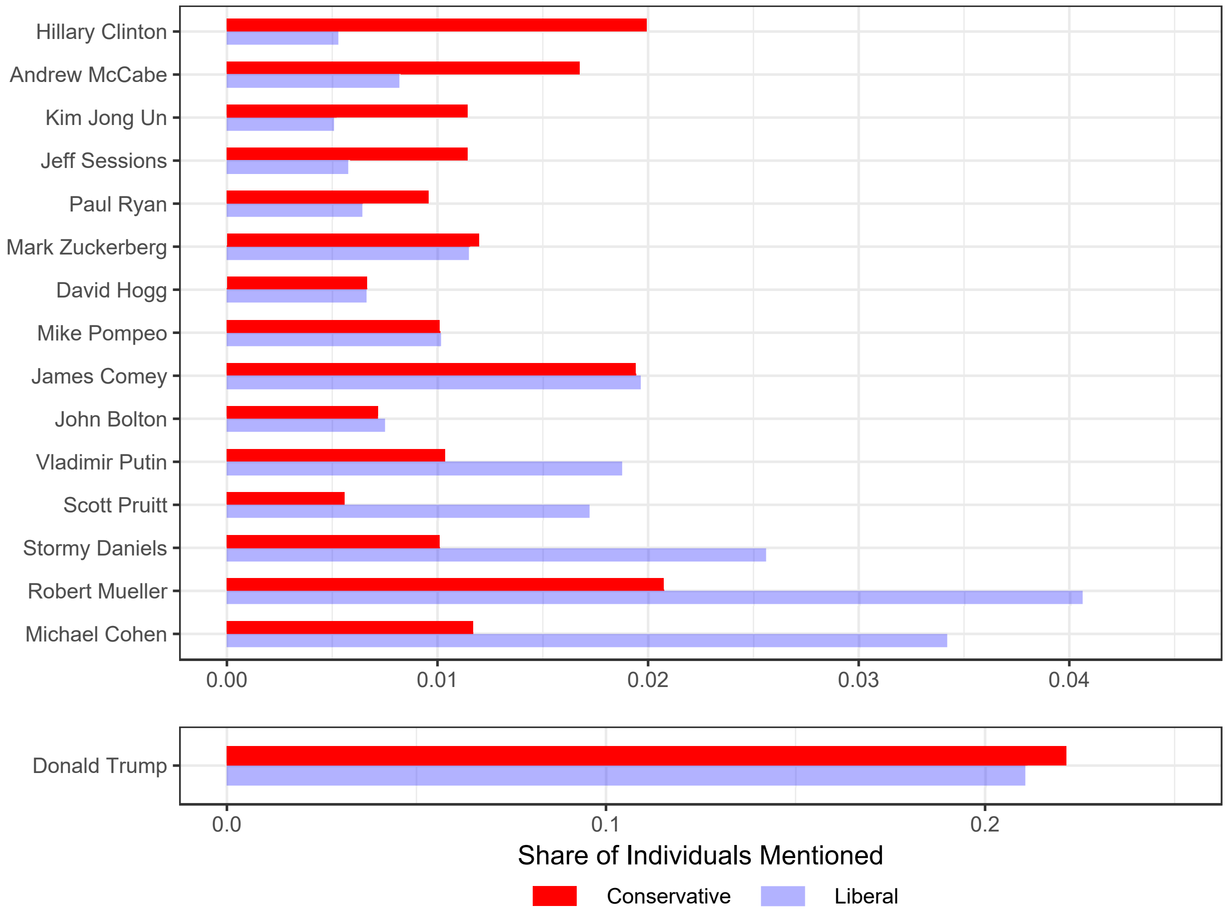

This binned scatter plot shows the correlation between consumers' ideology and online news consumption. The Republican vote share in the x-axis is based on 2008 zip code level voting data. The slant in the y-axis is based on the mean slant of all news sites visited by an individual, where the slant of each domain is determined according to Bakshy et al. (2015). A visit to a news site is referred from Facebook if the referring domain is "facebook.com." The sample includes all individuals in the 2017 Comscore Web Behavior Dataset Panel, who visited at least two news sites through Facebook and two news sites through other means.

Figure 2: Distribution of Mean News Slant



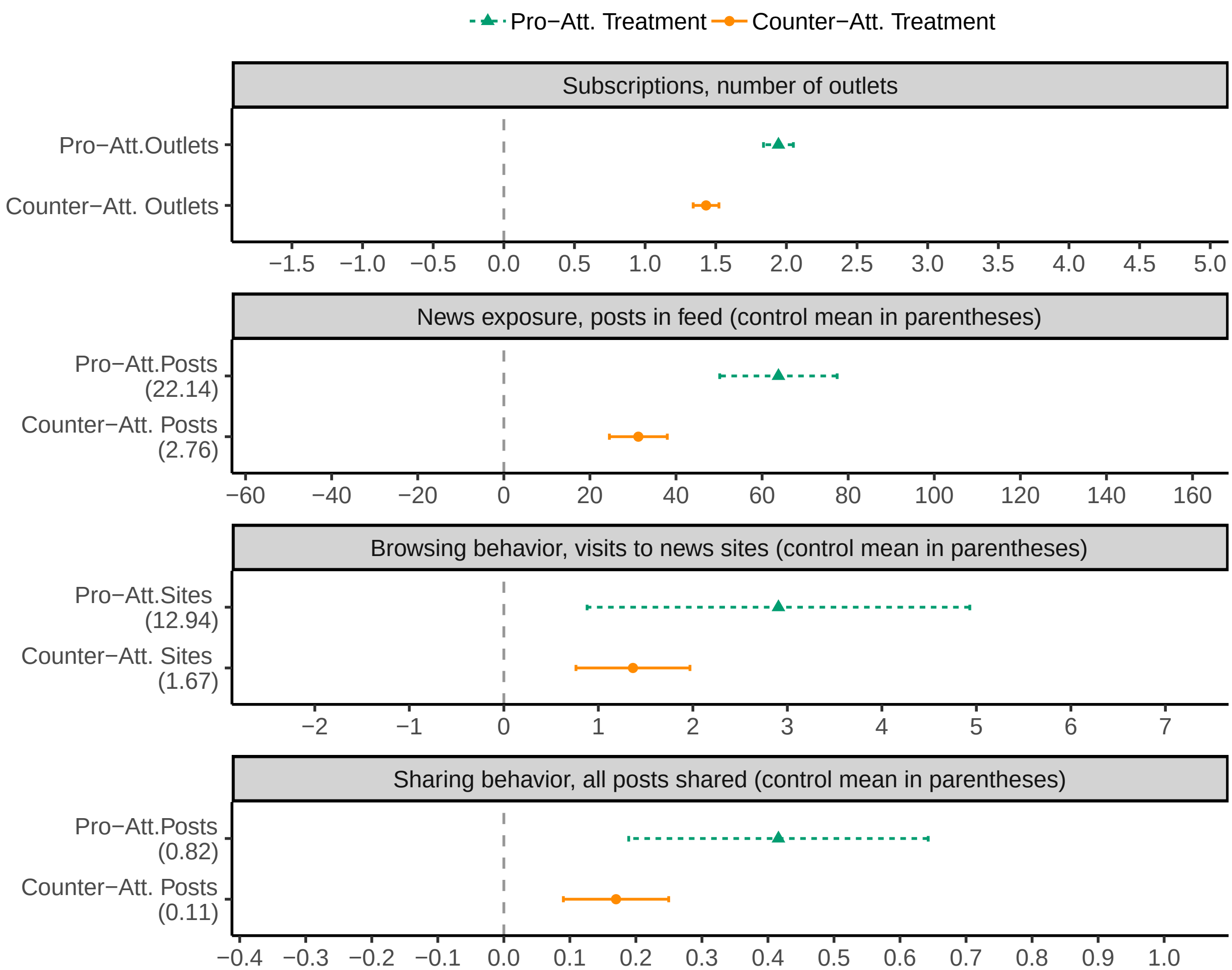

This figure shows the distribution of the mean slant of all news sites visited by individuals. The slant of each domain is based on Bakshy et al. (2015). A visit to a news site is referred from Facebook if the referring domain is "facebook.com." The sample includes all individuals in the 2017 Comscore Web Behavior Dataset Panel, who visited at least two news sites through Facebook and two news sites through other means. Major news outlets are added to the x-axis for reference. Smoothing bandwidth = 0.05.

Figure 3: Figures Discussed in the News During the Study Period



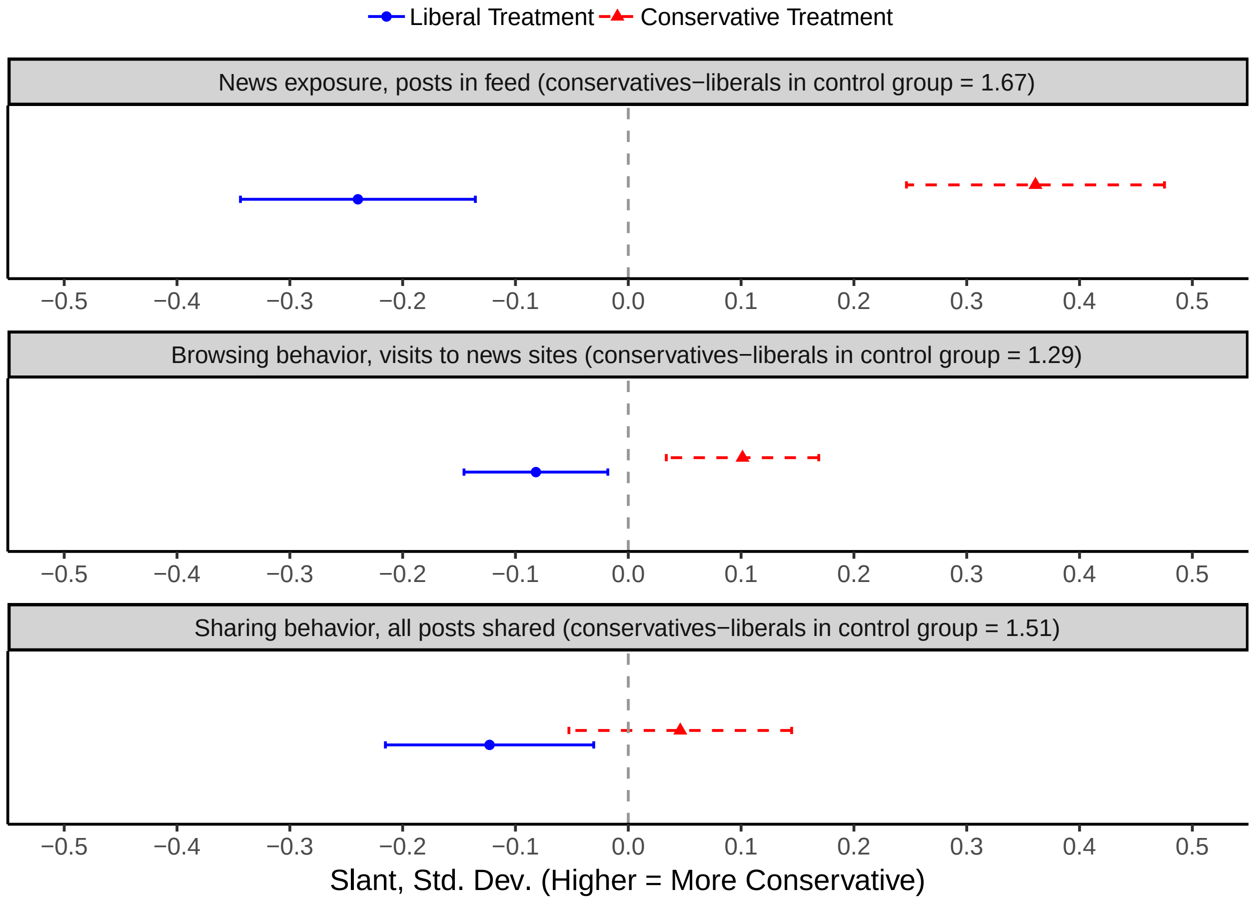

This figure shows the prominent men and women mentioned in posts shared by the primary outlets between February 28 and April 25, 2018, the median dates the baseline survey and endline survey were taken. The x-axis is the share of times an individual was mentioned in a post by one of the four primary conservative outlets (top bars) and by one of the four primary liberal outlets (bottom bars), of all individuals mentioned. 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 in President Trump's family.

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



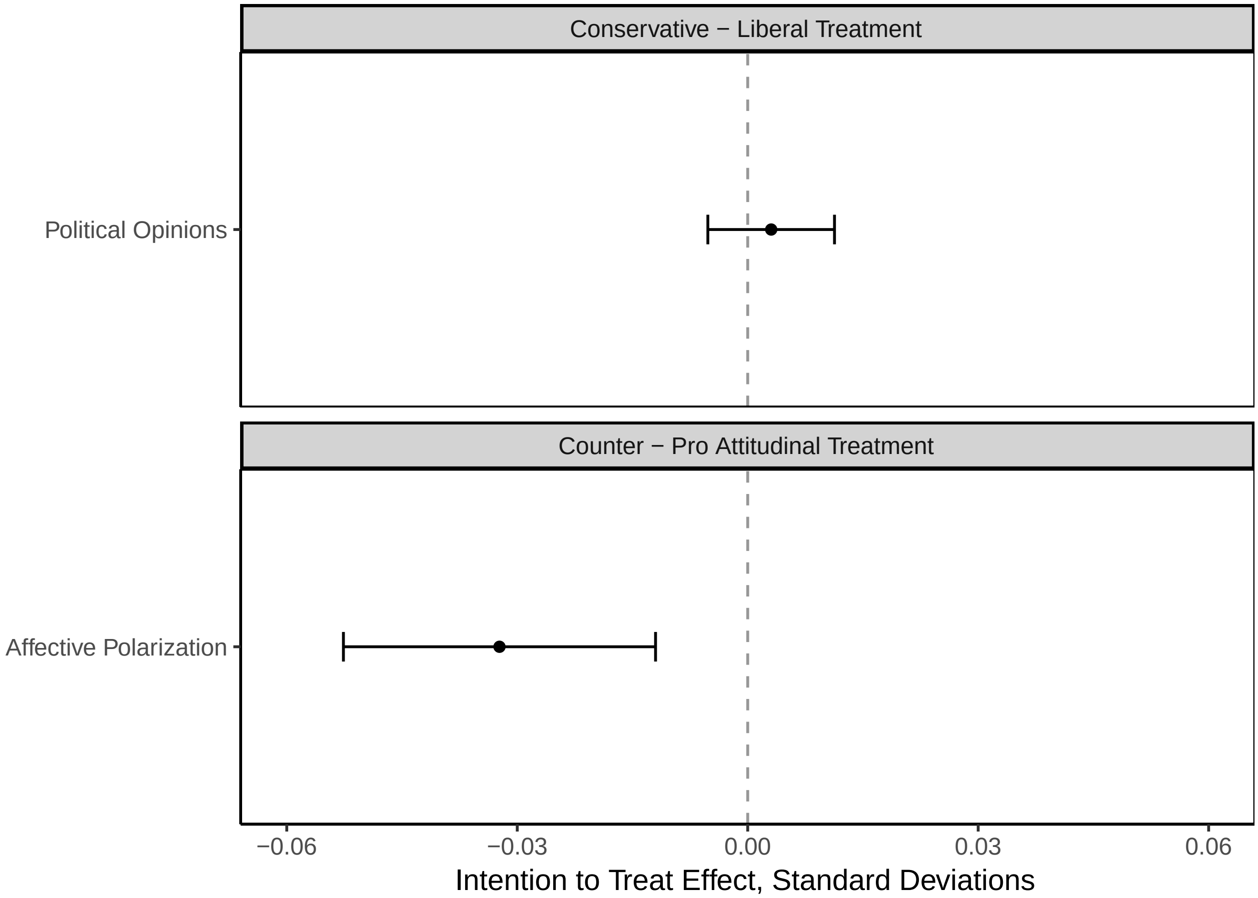

This figure shows the effect of the pro-attitudinal and counter-attitudinal treatments on total engagement with each participant's potential outlets in the two weeks following the intervention. In each row, the dependent variable is engagement with the four potential pro-attitudinal outlets or four potential counter-attitudinal outlets and the independent variable is the treatment. The outcomes are the number of outlets participants subscribed to, the number of posts from the outlets that appeared in their Facebook feed, the number of times they visited the outlets' websites, and the number of posts shared from the outlets by the participants. For example, in the third panel, the triangle and dashed lines 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 in the two weeks following the intervention. The regressions control for the outcome measure in baseline if it exists. The sample includes 1,651 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 Treatment on Slant of News Consumption



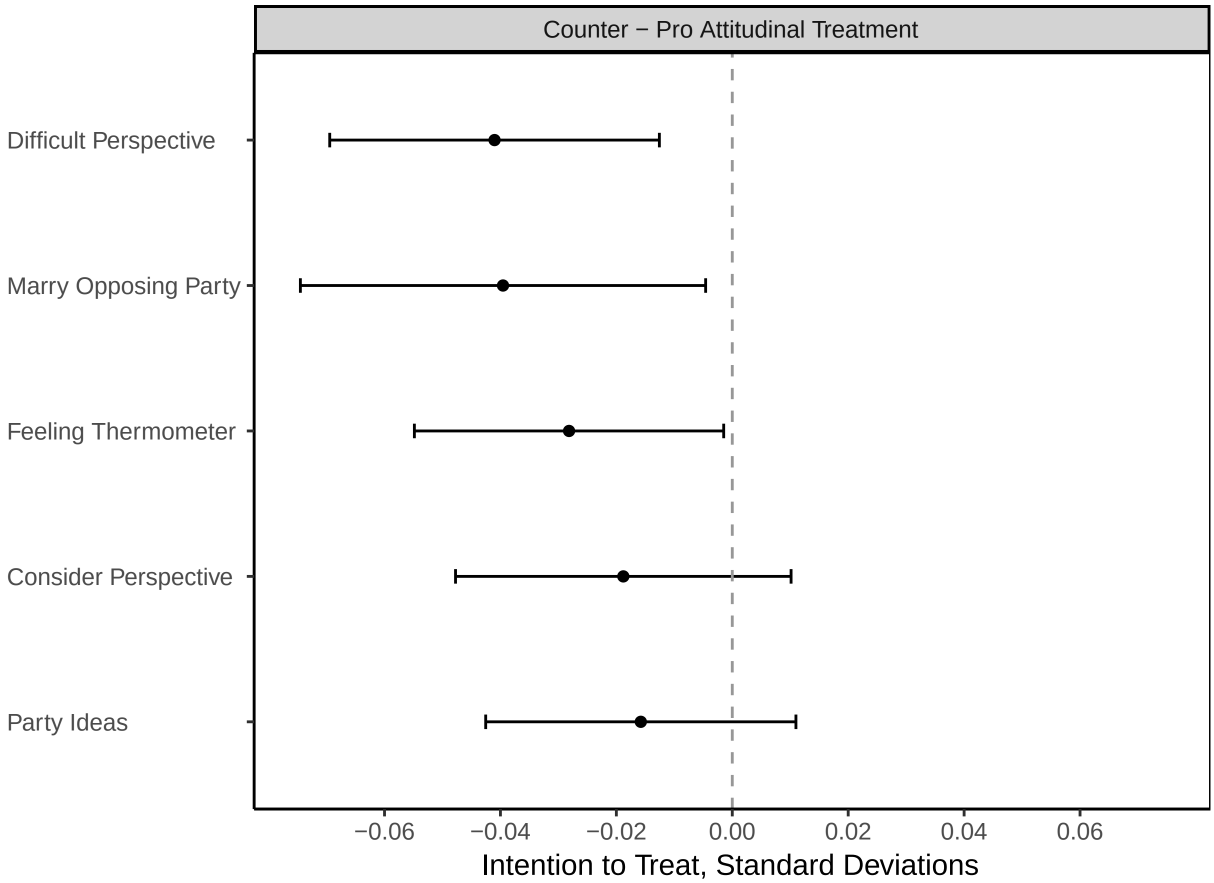

This figure shows the effect of the liberal and conservative treatments on the mean slant, in standard deviations, of all news individuals engaged with. In each panel, the dependent variable is the mean slant of outlets and the independent variable is the treatment. The regressions control for the outcome in baseline if it exists. The figure displays the slant for three outcomes: exposure to posts on Facebook (panel 1), news sites visited (panel 2), and posts shared (panel 3). The sample includes 1,702 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: Effect of the Treatment on Political Opinions and Polarization




This figure shows the effect of the treatments on the primary 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 3.4.2 and the regressions specifications are detailed in Section 3.6. The panels are based on 17,629 participants who took the endline survey. Error bars reflect 90 percent confidence intervals.

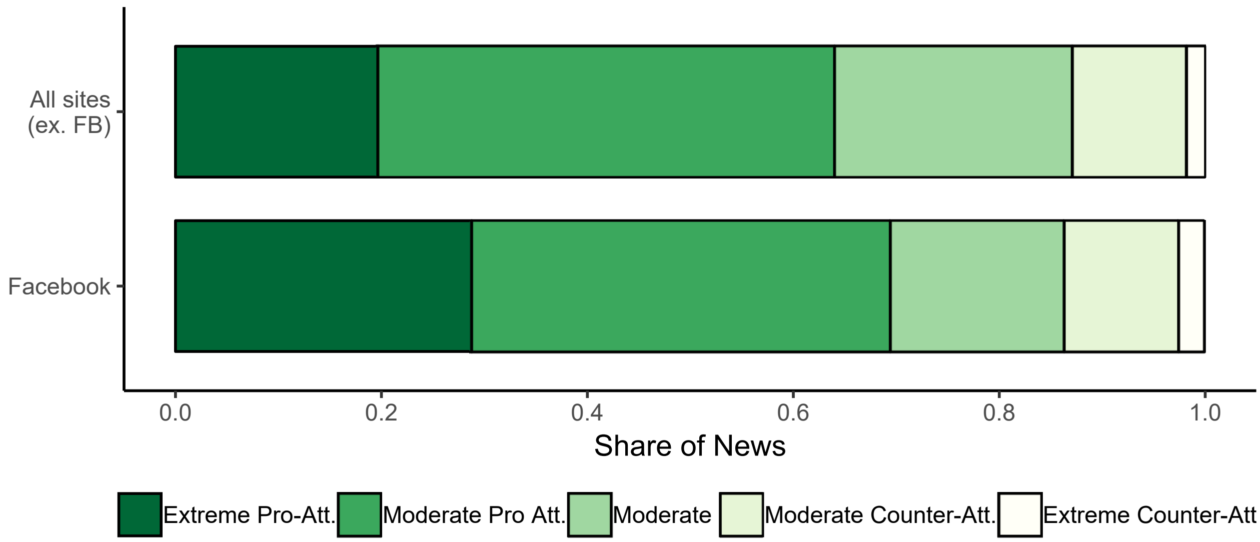
Figure 7: Effect of the Treatment on Affective Polarization - Individual Measures



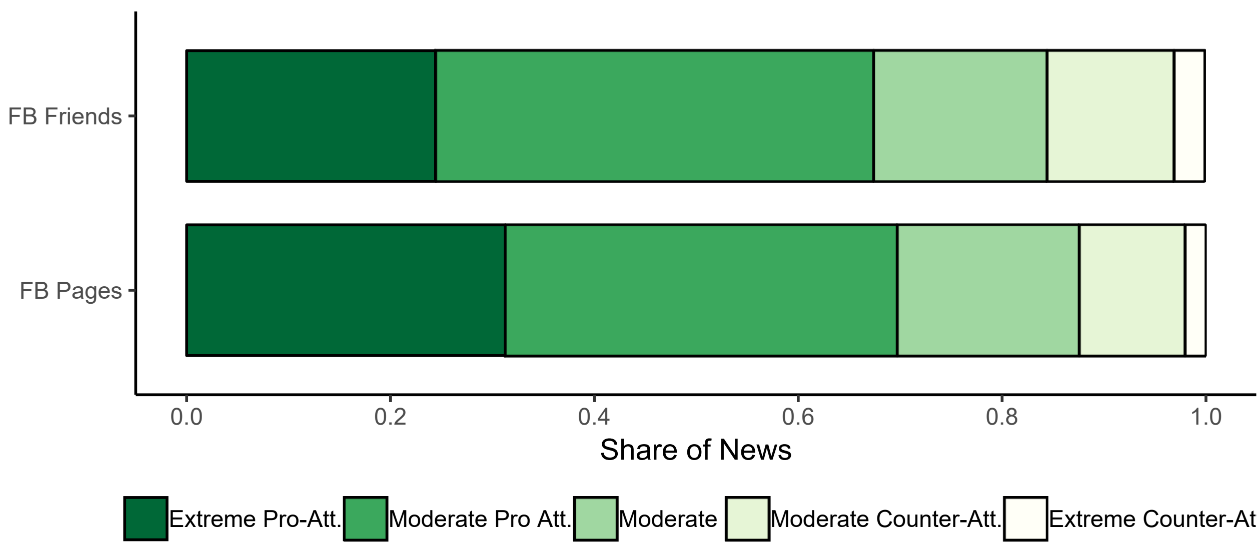

This figure shows the effect of the treatment on the measures composing the affective polarization index. Each row presents the result of a regression estimating the effect of the treatment on one dependent variable where a higher value is associated with a more polarized outcome. *Difficult Perspective* and *Consider Perspective* measures 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 variable 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 a feeling thermometer question asking participants how warm they 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 3.4.2. The regressions specifications are detailed in Section 3.6. Error bars reflect 90 percent confidence intervals.

Figure 8: Referral to News Sites, Control Group Browser Extension Data

(a) News Referred Through Facebook and Through All Other Means

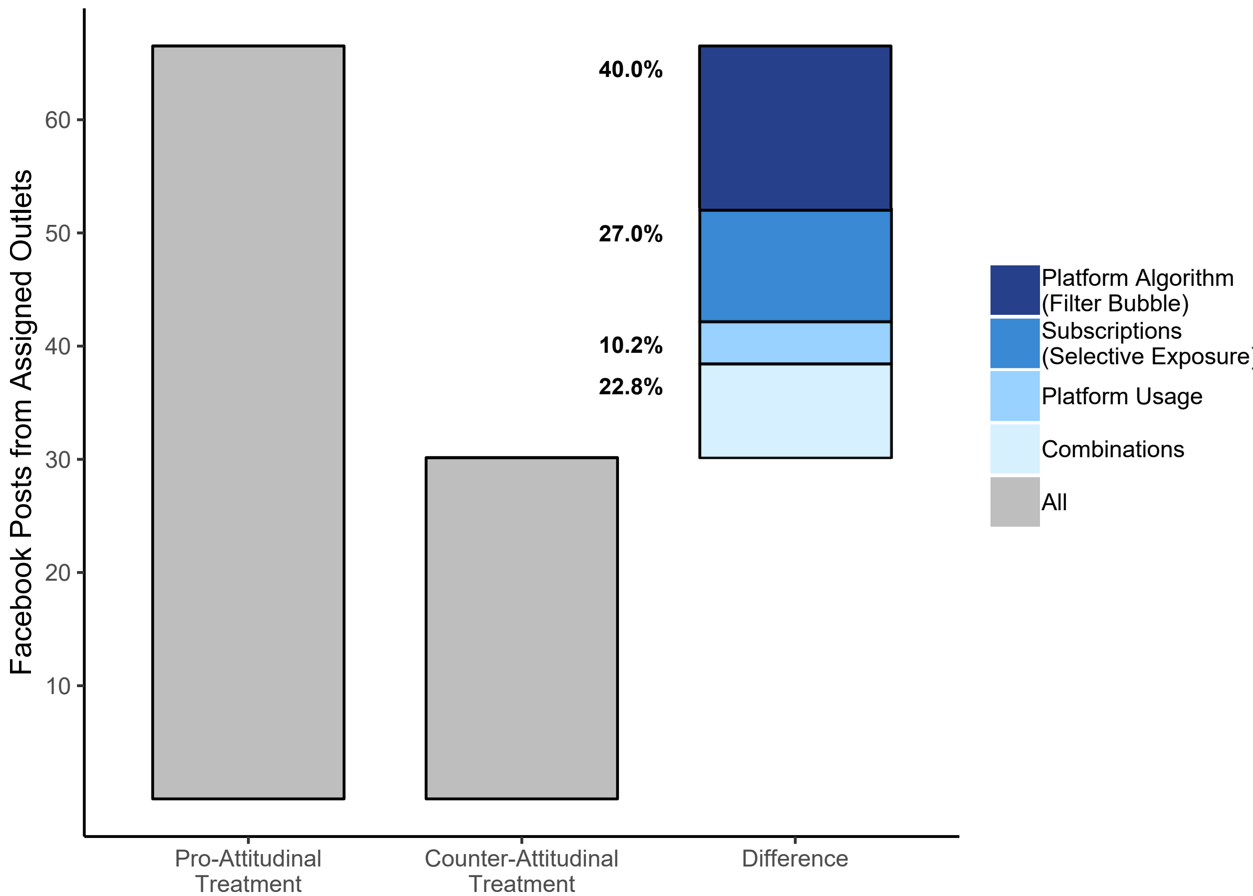



(b) News Referred Through Facebook: Comparison of Friends and Pages Referrals




These figures show the distribution of news consumed online from leading outlets according to the consumer's ideology. The share of news consumed for each category is first calculated at the individual level and then the mean is calculated for each group. For more details on processing leading outlets, see Appendix A.3. The first sub-figure shows all news consumed when a user clicked any link on Facebook compared to news consumed through any other means. The second sub-figure focuses on news consumed through Facebook and compared news sites accessed by clicking links shared by Facebook friends and news sites accessed by clicking links shared by Facebook pages. Data is from 30 days after the baseline survey and is based on 321 participants in the control group who installed the Chrome extension and who visited at least one news site by clicking a link shared by a Facebook page, one news site by clicking a link shared by a friend on Facebook and one news site through other means.

Figure 9: 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. *Platform 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. *Platform 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 1,058 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 Table A.22.

Table 1: Samples, Data Sources and Outcomes

| **Sample / subsample** | **Data sources** | **Number of Participants** | **Main Outcomes Measured** |
|---|---|---|---|
| Baseline sample | Baseline survey; Facebook data on participants' subscriptions to outlets. | 37,483 (all participants) | Subscriptions to outlets in the intervention (compliance) |
| Extension subsample | Browser data from participants who installed the chrome extension for at least two weeks | 1,838 | Posts observed on the Facebook feed (exposure); news sites visited |
| Access posts subsample | Facebook data on posts shared by participants who provided permissions to access their posts for at least two weeks | 34,568 | Sharing behavior; subscription to outlets over time |
| Endline survey subsample | Baseline and endline surveys responses for participants who completed both surveys | 17,629 | Political opinions; affective polarization |

This table describes the main sample and the subsamples analyzed along with the data source, the number of participants, and the main outcomes analyzed. The subsamples and data are described in Section 3.3. The outcomes are described in Section 3.4.

Table 2: Balance Table by Assignment to the Liberal and Conservative Treatments

| Variable | Mean | | | Difference Between Treatments | | |
|---|---|---|---|---|---|---|
| | Sample | US (2016) | FB Users (2018) | 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. | 29.1 | 43.1 | | 0.12 | 0.25 | -0.14 |
| Feeling Therm., Dem. | 47.0 | 48.7 | | 0.38 | 0.45 | -0.07 |
| 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 | | 0.00 | 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 | 47.7 | 47.3 | 42.9 | 0.21 | -0.14 | 0.34 |
| Total Subscriptions | 474 | | | 5.33 | 9.10 | -3.77 |
| News Outlets Slant (-1, 1) | -0.20 | | | 0.00 | 0.00 | 0.00 |
| Access Posts, Pre-Treatment | 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 | | | | 1.24 | 0.88 | 1.05 |
| P-Value | | | | [0.18] | [0.65] | [0.39] |

This table presents descriptive statistics, along with the difference between participants assigned to each treatment. *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), 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. Social media usage is based on a similar question in the Pew American Trends Panel Wave 23 among people who pay attention to posts about government and politics on Facebook. All other US data is based on the 2016 ANES. Opinions seen on Facebook is based on the Pew American Trends Panel, Wave 1. All other data on Facebook users is based on the 2018 Pew Core Trends Survey. *p<0.1 **p<0.05 ***p<0.01.

Table 3: Compliance with the Treatments

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Cons. Treat., Cons. Ideology | 0.471*** | 0.486*** | 0.503*** | | | |
| | (0.008) | (0.008) | (0.008) | | | |
| Lib. Treat., Cons. Ideology | 0.320*** | 0.334*** | 0.286*** | | | |
| | (0.007) | (0.008) | (0.007) | | | |
| Cons. Treat., Lib. Ideology | 0.526*** | 0.515*** | 0.507*** | | | |
| | (0.006) | (0.006) | (0.006) | | | |
| Lib. Treat., Lib. Ideology | 0.603*** | 0.591*** | 0.600*** | | | |
| | (0.006) | (0.006) | (0.006) | | | |
| Know Slant | | | | 0.251*** | 0.221*** | 0.227*** |
| | | | | (0.005) | (0.006) | (0.006) |
| Outlet Ideology, Abs. Value (Std. Dev.) | | | | −0.069*** | −0.046*** | −0.044*** |
| | | | | (0.002) | (0.003) | (0.003) |
| Ideological Distance (Std. Dev.) | | | | −0.080*** | −0.080*** | −0.083*** |
| | | | | (0.002) | (0.002) | (0.002) |
| Controls | | X | X | | X | X |
| Immediate Compliance | X | X | | X | X | |
| Compliance After 2 Weeks | | | X | | | X |
| Observations | 36,717 | 36,717 | 34,611 | 97,909 | 97,909 | 92,212 |

This table estimates the association between participants' characteristics and compliance with the treatment. In columns (1-3), the dependent variable is subscription to at least one offered outlet and the independent variable is the interaction of participant's ideological leaning and her treatment assignment. The reference group is the control group, where there is no compliance. In columns (4)-(6) each observation is a participant and an outlet offered (the control group is excluded). The dependent variable is whether a participant subscribed to a specific outlet. The independent variables are based on the outlet's perceived ideology according to the participant (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), (3), (5), (6) control for the participant's age, age squared, gender and the potential outlets defined for the participant. Column (5) also controls for outlet fixed effects. Column (3), (6) measure compliance two weeks after the intervention and exclude participants who did not provide permission to observe pages they subscribe to for at least two weeks. Data is from the entire sample of participants who completed the baseline survey. Columns (1)-(3) use robust standard errors and columns (4)-(6) use clustered standard errors at the individual level . *p<0.1 **p<0.05 ***p<0.01

Table 4: Effect of the Social Media Feed on News Sites Visited

| | IV | |
|---|---|---|
| | Slant of News Sites Visited | Slant of News Sites Visited through FB |
| | (1) | (2) |
| Slant of FB Feed | 0.311$^{***}$ | 0.707$^{***}$ |
| | (0.068) | (0.086) |
| Controls | X | X |
| First Stage F-Stat | 61.03 | 72.36 |
| Observations | 1,525 | 1,204 |

This table shows two IV regressions estimating the effect of the Facebook feed on news sites visited. The dependent variable is the mean slant of news sites visited and the independent variable is the mean slant of the participant's Facebook feed, instrumented by the treatment. In column (1) the dependent variable is based on all news sites visited and column (2) is based on news sites visited through Facebook. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table 5: Effect of the Treatments on Attitudes Toward Each Party

| | Attitude Own Party | Attitude Opposing Party |
|---|---|---|
| | (1) | (2) |
| Counter-Att. Treatment | 0.001 | 0.031** |
| | (0.014) | (0.014) |
| Pro-Att. Treatment | 0.009 | −0.003 |
| | (0.013) | (0.014) |
| Pro - Counter | 0.007 | -0.034** |
| | (0.014) | (0.014) |
| Observations | 16,889 | 16,889 |

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 (based on self-reported ideology, party affiliation, and candidate preferred) 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 *Marry Opposing Party* question is not included since participants were only asked how they would feel if their son/daughter married someone from the opposing party. The controls are specified in Section 3.6. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

# Appendix For Online Publication

## A Data Collection and Processing

### A.1 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). Each observation in the dataset is a domain visited by a computer and includes details on the referring domain, the time visited, the number of pages visited, and online purchases. I assume that each id represents a unique individual. While Comscore attempts to identify each unique individual, it is still possible that multiple individuals use the same machine.

I include in the sample only individuals who visited at least two news sites through Facebook and two news sites through other means. This ensures that for every individual in the sample, I can calculate the slant of news sites visited through Facebook and the slant of other news sites visited, and thus any difference in the slant does not stem from differences in the individuals who visited the news sites.[66]

### A.2 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.

I exclude from the dataset the following popular websites which are not related to news outlets: Amazon, The White House, Twitter, Vimeo, Wikipedia, and YouTube. I also exclude the 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., 2018). I then remove the web reference ("www.") from all outlets' websites, so all outlets only contain the domain used.[67] After processing the data, the list of leading outlets contains 488 websites. I determine the Facebook page for outlets on this list by searching for all pages with names similar

[66] I require at least two news sites for each category to slightly decrease measurement error. The results are similar when including individuals who visit at least one news site through Facebook and one news site through other means.

[67] Websites which appear twice in the dataset, with and without the web reference, are merged 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.

to each outlet's domain and manually checking the pages. Overall, Facebook pages were found for 374 out of 488 outlets.

Similarly to Bakshy et al. (2015), I define very liberal outlets as outlets in the bottom quintile of the distribution of news slant outlets, liberal outlets as outlets in the second quintile, moderate outlets as outlets in the middle quintile, conservative outlets as outlets in the fourth quintile and very conservative outlets are defined as outlets in the top quintile of the news slant distribution.

## A.3 Surveys and Sample

### A.3.1 Baseline Survey

The baseline survey took place from early February to mid-March 2018 and participants in the survey compose the baseline sample. The recruitment ads either emphasized that a survey is being conducted and participants will take part in a gift card raffle or that the survey may be of interest to people who follow politics.[68] The ads targeted Facebook users living in the US who are over 18 years old, and who are likely to click the ad and begin the survey. 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 browser extension. 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.[69]

40,514 responders took the survey and reached the screen where the intervention occurs. Of those, 37,492 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 to define for the responder four potential liberal outlets and four potential conservative outlets (3.64%); technical issues with the Qualtrics survey which prevented some data from being collected (0.28%); taking the survey a second time (0.01%).; responding carelessly (0.03%). 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

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

[69] 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.

minutes) and responders who did not answer at least half of the closed-ended, non-required questions. 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. This removes fewer than 0.1% of participants from the baseline sample.

### A.3.2 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.[70] 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.

17,629 participants are included in the endline survey subsample. Respondents are included in the subsample, even if they did not complete the endline survey. 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.[71] The data allowed me to customize the survey by ensuring participants are not offered outlets they 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

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

[71] 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.

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. This means that the effect I find on the number of posts shared as a result of the experiment are 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 redirected URL before being matched to the list of domains.

### A.5 Extension Data on Facebook Feed and 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 actually 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-related posts appearing in participants' Facebook feeds are matched to outlets using the same method explained in the previous section. 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.[72] All URLs are first converted to the final redirected URL before matching posts or news sites visited.

## B Pre-Analysis Plan

The main outcome and hypothesis tested in this study were pre-registered in the AEA RCT Registry.[73] 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

[72] 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%.

[73] AEA RCT Registry Trial 0002713.

index variables will be determined by the inverse of the covariance between variables at endline (Anderson, 2008). This method is not used since it generates negative weights. When using 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.

Appendix Table A.15 repeats the analysis with the inverse-covariance weights. Column (2) shows that the difference between the effects of the pro-attitudinal and counter-attitudinal treatments on affective polarization remains almost exactly the same when using these 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, in order to create the index for all participants who have at least one non-missing outcome. The results remain essentially the same. Appendix Table A.15a shows the results of a similar analysis with the political opinions index. Since the inverse-covariance method generates negative weights, columns (4) and (5) 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, and 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.[74]

Column (1) of Appendix Table A.16 shows that the primary estimate is still significant when using all eight variables in the polarization index.[75] Column (3) measures the effect only on the behav-

[74]The correlation between the behavioral polarization measures and partisanship is 0.03-0.06, while the correlation between the affective polarization measures and partisanship is 0.17-0.36.

[75]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,

ioral 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.

# C Additional Details on Empirical Strategy

## C.1 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. Ideological leaning is a binary variable, defined according to party affiliation, self-reported ideology and the presidential candidate supported.[76]

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 3.4.2).

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

the effects are similar since the index created when using all eight variables has less variation in the control group.

[76]Ideological leaning was not explicitly mentioned as a control variable in the pre-analysis plan. This covariate is included since it is used to determine if a participant was assigned to the pro-attitudinal or counter-attitudinal treatment. This does not affect the results. The pre-analysis plan also stated that I would control for strata. I do not control for strata since due to attrition, some strata have only one or two respondents instead of the original three respondents defined for each stratum. The results are robust to controlling for strata.

since individuals who were assigned the alternative outlet may have different characteristics than individuals who were assigned the primary outlets.

# D Additional Analysis

## D.1 Social Media and News Consumption

Appendix Table A.17 estimates the association between the slant of news consumed and the interaction of the consumer's ideology and Facebook usage in a regression framework. Column (1) includes all domains visited and shows that an increase of 10% in the Republican vote share is associated with an increase of 0.09 standard deviations in the slant of news consumed (where a higher value is more conservative) when news is consumed through Facebook compared to other news consumed. Column (2) shows that an effect is found when adding individual and month fixed effects. These regressions do not take into account spillovers. Individuals could merely be switching the medium through which they consume specific domains without being affected by social media. For example, if a conservative visits Fox News by clicking Facebook links, she may, as a result, consume less news by directly entering the Fox News home page. Columns (3)-(5) overcome this issue by measuring the association between the share of news consumed through Facebook and the mean slant of *all* news an individual consumed. Column (3) confirms that individuals visiting a greater share of news sites through Facebook tend to consume news better matching their ideology. Column (4), my preferred specification, shows that months when an individual consumes more news through Facebook are associated with news consumption better matching the individual's ideology. Column (5) shows that this result is robust to using total Facebook usage instead of the share of news consumed through Facebook as the independent variable.

Appendix Table A.18 presents a similar table where the dependent variable is the absolute value of the slant of news consumed and the independent value is whether news is consumed through Facebook. The table confirms that Facebook is associated with more extreme news.

## D.2 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 define that participants are in an *Echo Chamber* if their 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 median. *Potential Outlets Echo Chamber* is whether the difference between how often participants reported seeing the potential pro-attitudinal and counter-outlets in their feed is above the median.

I measure whether a participant has an *Open Personality* according to whether her average answer to the following questions 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. Participants are defined to get *Most News Social Media* if they get most of their news about government and politics through social networking sites (such as Facebook or Twitter). 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 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 equal 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 it could be due to the fact that Fox News is more persuasive than New York Times.

The panel on the left side of Appendix Figure A.11 shows that the effect on political opinions is mostly homogeneous (i.e., most people were not persuaded by the treatment). In the panel, each row represents a separate regression estimating the effect of interacting the conservative treatment with the specified moderator, where the reference group is the liberal treatment. A higher value means that this group of participants was more likely to be persuaded by the treatment (they were more likely to become conservative, compared to other participants, when exposed to the conservative treatment, compared to the liberate treatment). Only two marginally significant effects are detected. Participants who self-report following the news more frequently were slightly more likely to be affected by the treatment. Somewhat surprisingly, participants who scale higher on the openness index were *less* likely to be persuaded by the intervention.[77]

The panel on the right side of Appendix Figure A.11 does not find strong heterogeneous effects on affective polarization according to most covariates tested. In the panel, each row represents a separate regression estimating the effect of interacting the pro-attitudinal treatment with the specified variable, where the reference group is the counter-attitudinal treatment. 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 strongest heterogeneous effect found is based on the baseline feeling thermometer measure for affective polarization. The effect of the pro- and counter-attitudinal treatments on affective polarization is weaker among people who were more polarized in baseline (had a larger difference between their feeling toward their party and the opposing party). However, this result is significant at the 10% level and the results are not adjusted for multiple hypothesis testing. More research is needed to test whether people who already have more negative opinions toward the opposing party are less likely to be affected by counter-attitudinal outlets.[78]

### D.3 Reweighting for National Representativeness

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 variables missing variables are first replaced with the mean value (less than 5% of observations are missing for each of these variables). The estimates for the national population are based on similar questions in the 2016 American National Election Survey (ANES).[79] 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 for reweighting the sample.[80]

[77] Using the same measure of openness, Gerber et al. (2012) find surprising evidence that more open individuals discuss politics more frequently the more they agree about politics. These two results suggest that this definition of openness might not predict being open to discussion of or persuasion by counter-attitudinal opinions.

[78] 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.

[79] All estimates are based on the pre-election survey, besides vote or support for a presidential candidate, which is based on the post-election survey.

[80] 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

Appendix Table A.19 shows that reweighting the variables does not change the main conclusions of the study. The effect on political opinion remains very close to zero, and the effect on affective polarization remains essentially the same. These tables should be interpreted with caution. It is likely that even after reweighting the sample, the sample is still different than the national population on various characteristics. Still, the tables show that it is unlikely that an effect on affective polarization is found since the survey sample is more liberal or more polarized than the rest of the population.

### D.4 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.20 presents the effect of the treatment 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 treatment seems to have little to no effect on the knowledge outcomes.

Appendix Table A.21 uses the browser extension data to show that the intervention affected news exposure. The regressions measure the effect of the treatment on the number of posts that appeared in the participants' social media feeds and referred to relevant topics.[81] For all four topics, the treatment 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.[82]

The results presented in this section suggest that while the slant of one's social media news feed

survey. Overall, 94% of respondents in the ANES survey are included.

[81] 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."

[82] 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.

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 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).

## D.5 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, discuss the gap in exposure to pro and counter-attitudinal outlets among outlets not included in the experiment and test whether there is a gap in exposure to pro and counter-attitudinal articles within outlets.

### D.5.1 Decomposition Calculations

I include in the data only participants in the pro-attitudinal 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 520 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:

$$S_C * P_C * U_C$$

where $S_C$ is the mean number of new subscriptions to the offered counter-attitudinal outlets. $P_C$ 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:

$$S_P * P_P * U_P = (S_C + S_\Delta) * (P_C + P_\Delta) * (U_C + U_\Delta)$$

I then decompose the difference in exposure to four separate expressions as described in Equation 3. To calculate $S_\Delta$ and $P_\Delta$, 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 she observed, respectively. These regressions are presented in Appendix

Table A.22, 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 how many outlets the participant subscribed to in each group and how many posts from each group of outlets appeared in her Facebook feed. I only include posts shared by the outlet to isolate any effect of friends sharing specific posts. Finally, I calculate the share of posts the participants observed from these outlets among the total number of posts from all sources the participant observed in her feed in the two weeks following the intervention. 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. $P_C$ and $P_\Delta$ are estimated using the following regression:

$$SharePosts_{ij} = P_C * Sub_{ij} + P_\Delta * Sub_{ij} \times Pro_{ij} + \delta * Pro_{ij} + \nu_{ij} \tag{6}$$

where $SharePosts_{ij}$ is the number 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 $Offer_{ij} \times Pro_{ij}$. This regression is presented in column (3) of Appendix Table A.22. 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 ($P_C$). 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. $P_\Delta$ is the difference between the coefficients.

### D.5.2 Alternative Decompositions

Appendix Figure A.12 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 9. 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 selection, 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.12, I do not define cases where participants canceled a subscription in the first two weeks following the intervention as a subscription. In this estimation, I only include participants for which I could observe their subscriptions to outlets in the two weeks following the intervention. Taking into account unsubscriptions does not substantially change the results.

**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-attitudinal and counter-attitudinal treatments might be due to treatment heterogeneity and selection into compliance, and not due to different treatment effects. In the fourth row panel of Appendix Figure A.12, I re-weight the IV estimators, such that the compliers resemble the entire sample. 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 (3 dummy variable 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 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 average 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.

**Reweighting for Population** In the fifth 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 algorithms tendency to increase exposure to matching news outlets is driven by the liberals in my sample. However, I am underpowered to estimate these results precisely. Furthermore, this difference can be due to the ideology of the participants or to the properties of the outlets offered to the participants.

**Excluding Facebook Usage** The effect on Facebook usage is only marginally significant. In the sixth row of Appendix Figure A.12, 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 P in equation 6. Instead of estimating the effect on the share of posts in the feed, I estimate the effect on the number of posts observed in the feed by participant $i$ from outlets in group $j$.

#### D.5.3 Exposure Among Outlets Not Included in the Experiment

Interestingly, when estimating exposure cross-sectionally among outlets not included in the intervention, there is a very large gap in subscriptions to pro and counter-attitudinal outlets, but participants are not exposed more often to posts from pro-attitudinal outlets, conditional on subscription. For example, among the 20 most popular liberal and conservative outlets not included in the experiment, participants subscribed to 21% of outlets matching their ideology, compared to only 3% of outlets not matching their ideology and observed 1.68 posts from the pro-attitudinal outlets and 0.34 posts from the counter-attitudinal outlets. This gap in exposure is much larger than the gap among outlets offered in the experiment and seems to be driven by selection and not algorithms. However, this type of analysis cannot cleanly identify the effect of offering or subscribing to outlets since the offer to subscribe is not randomly assigned. It is plausible that the nudges participants received are not random. If, for example, participants are usually nudged (either by Facebook or by ad) to subscribe to pro-attitudinal outlets, one would expect to find a large gap in subscriptions. In such a scenario, the subscribers to counter-attitudinal outlets may be the consumers who actively sought out these outlets without a nudge, and thus are more likely to engage with these outlets as this may mitigate the effect of the algorithm.

While I cannot observe why an individual subscribed to an outlet, it is clear that participants subscribing to pro- and counter-attitudinal outlets are substantially different from each. For example, among the 20 most popular liberal and conservative outlets, there is a difference of 0.4 standard deviations in the absolute value of ideology of participant subscribing to at least one pro- and counter-attitudinal outlet.[83] Moreover, the actual subscriptions occurred separately. On average, subscriptions to counter-attitudinal outlets occurred nine months later than subscriptions to pro-attitudinal outlets, and it is likely that posts from more recent subscriptions are more likely to appear in the feed. The experiment assures that all subscriptions occur at the same time, and all subscriptions occur due to a random offer. While there could still be differences in compliers, the LATE effect of subscriptions on exposure can be estimated for each treatment arm cleanly. Furthermore, the differences between the compliers in each treatment arm are relatively small (for example, there is less than 0.1 standard deviation difference in the absolute value of ideology between compliers in the pro- and counter-attitudinal treatments).

[83] The difference is similar when focusing on the *feeling thermometer* and *difficult perspective* measures.

This cross-sectional analysis not only demonstrates why an experiment is necessary, it 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, if policymakers or platforms want to increase diversity in news exposure, they should also aim for balanced nudges encouraging participants to subscribe to outlets.

#### D.5.4 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 Congress Members who shared the article.[84] Using this measure, I 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.

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

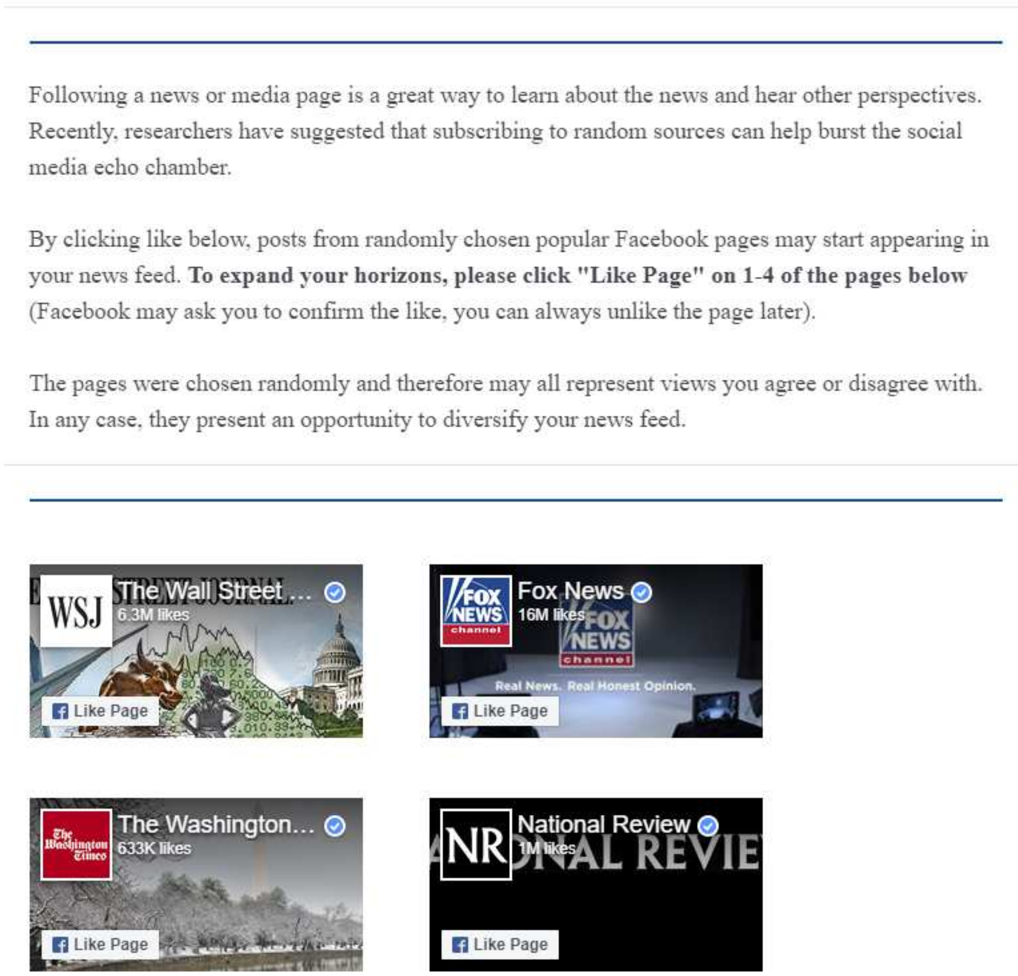
# E Additional Figures and Tables

Figure A.1: Recruitment Ads


(a) Political Ad

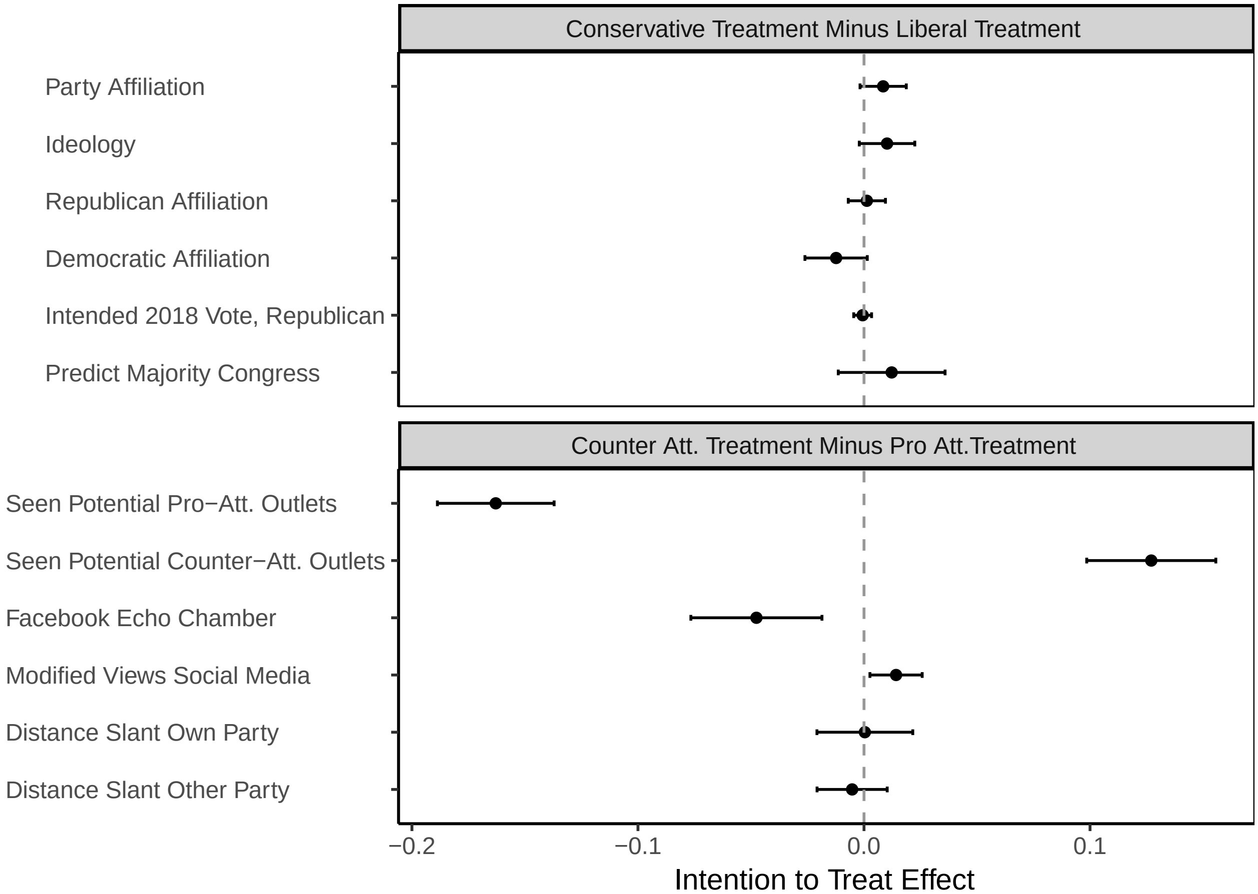

(b) General Ad

Figure A.2: The Conservative Treatment Intervention



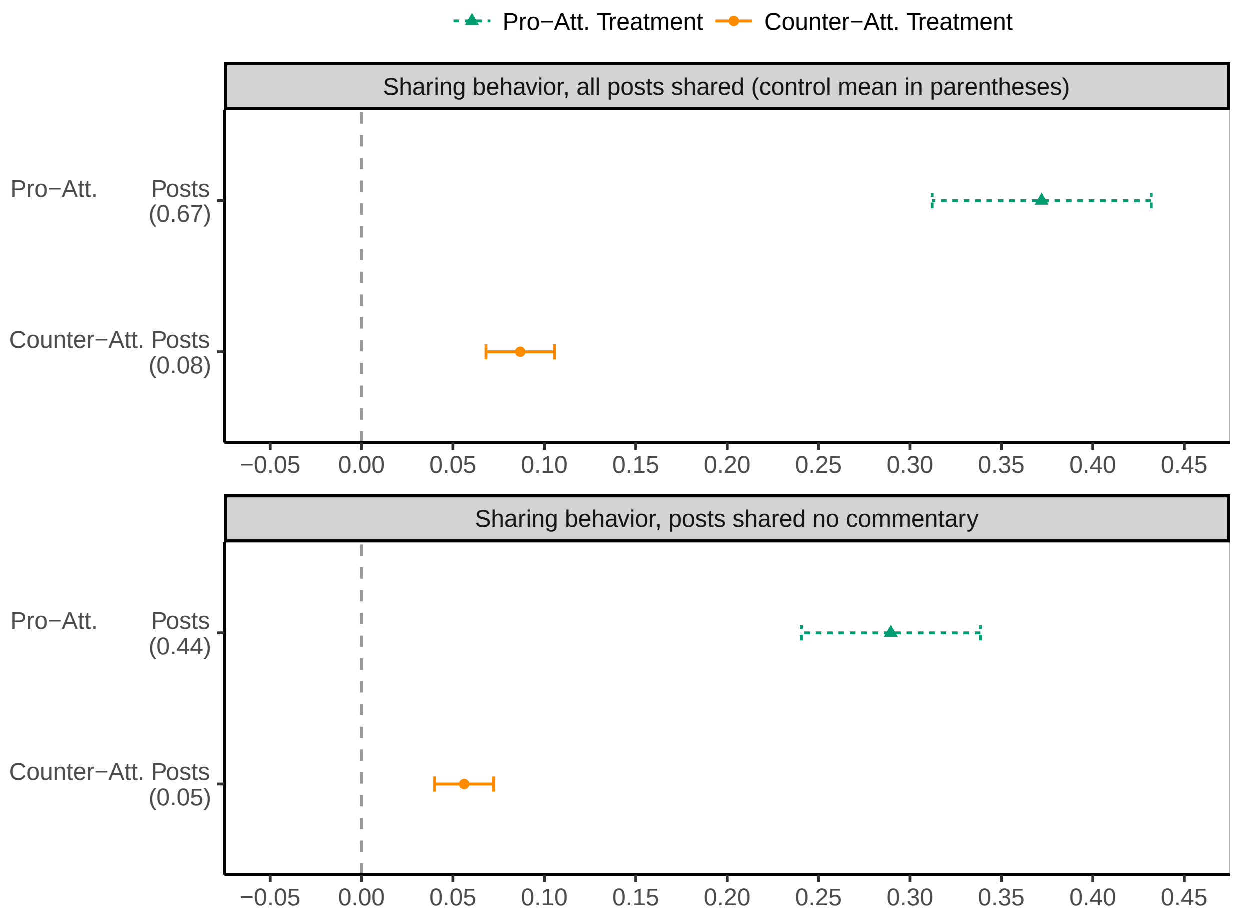

This figure shows the survey page asking participants to subscribe to four conservative outlets. Participants randomly assigned to the conservative treatment, who have not already subscribed to the four primary outlets, were shown a page similar to this figure. The image in the background of each outlet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

Figure A.3: Effect of the Treatment on Additional Outcomes




This figure shows the effect of the experiment on additional outcomes. *Party Affiliation* is the party the participant identifies with on a 7-point scale. *Ideology* is self-reported on a 7-point scale. *Republican/Democrat Affiliation* is coded as 1 if the participant leans toward the Republican/Democrat party, 2 if the participant is a Republican/Democrat, 3 if the participant is a strong Republican/Democrat and 0 otherwise. *Intended 2018 Vote, Republican* is whether the participant would have voted for the Republican Party candidate in their district if the election was held the day the survey was taken, among participants intending to vote for the Republican or Democratic Party candidates. *Predict Majority Congress* is the party participants predicted will hold the majority of seats in Congress after the 2018 vote where the Republican Party is coded as 1, note sure is coded as 0, and the Democratic Party is coded as -1. *Seen Pro/Counter Attitudinal* is how often the participant reported seeing news from their potential pro- or counter-attitudinal outlets in their Facebook feed over the past week where the possible answers are more than 5 times (3), 3-5 times (2), 1-2 times (1), have not seen (0). *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), not too often (0). *Modified Views Social Media* is whether consumers self-reported modifying their views in the past two months about a political or social issue because of something they 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 the control group standard deviation. The specification and controls are described in more detail in Section 3.6. The regressions also control for baseline measures of the outcomes when they exist. Error bars reflect 90 percent confidence intervals.

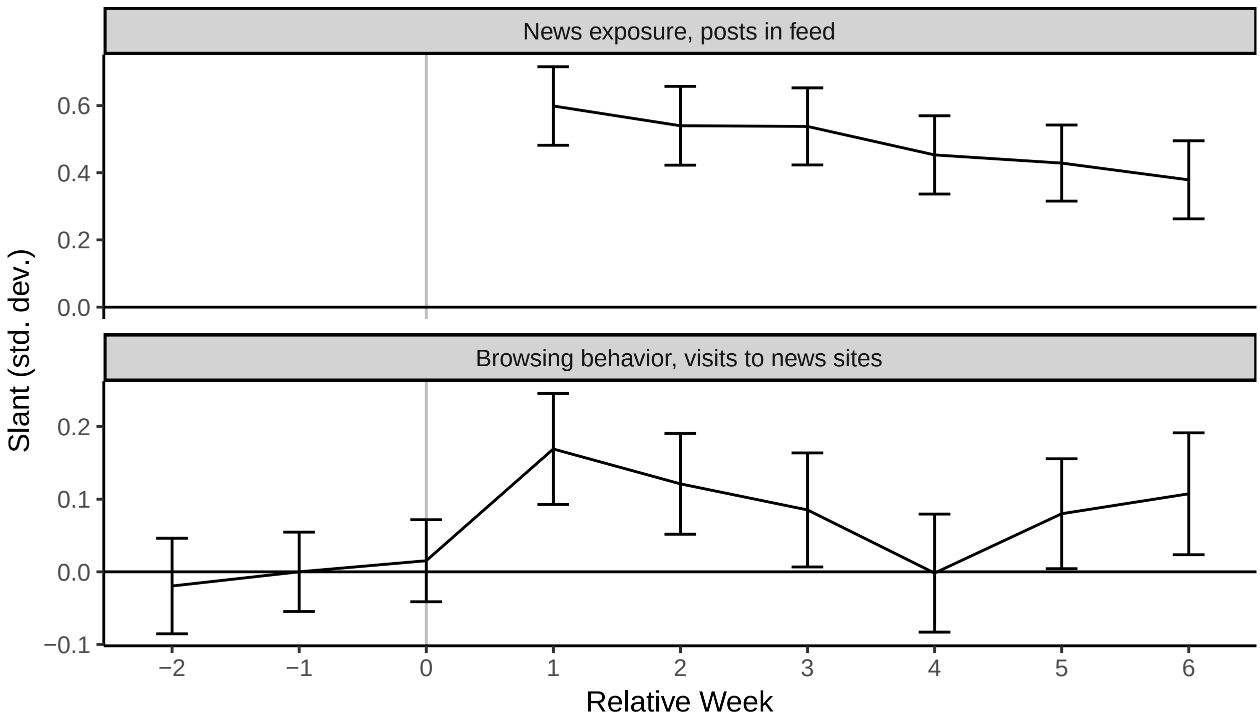
Figure A.4: Effects of the Pro- and Counter-Attitudinal Treatments on Number of Posts Shared, Access Posts subsample



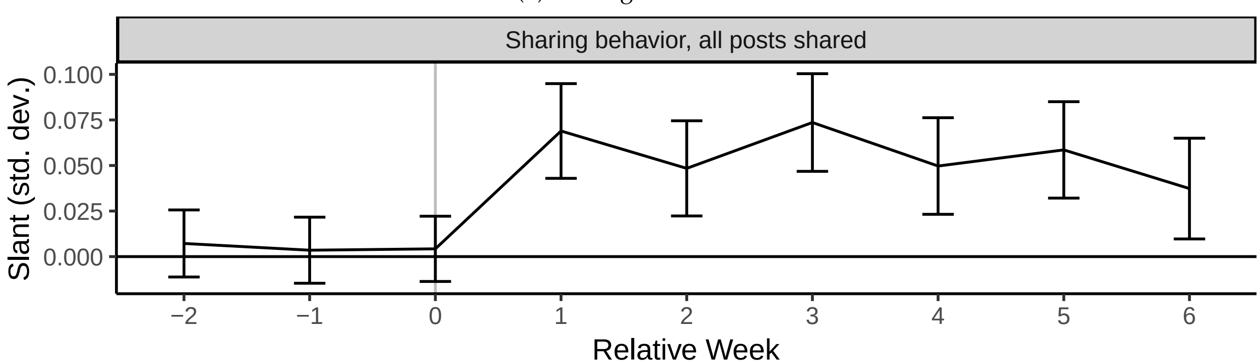

This figure shows the effect of the pro-attitudinal and counter-attitudinal treatments on the number of posts participants shared from the four potential pro-attitudinal outlets and four potential counter-attitudinal outlets in the two weeks following the intervention. The first panel includes all posts 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: 33,509 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.5: Mean Slant over Time, Conservative Treatment, Compared to Liberal Treatment

(a) News Exposure and Browsing Behavior

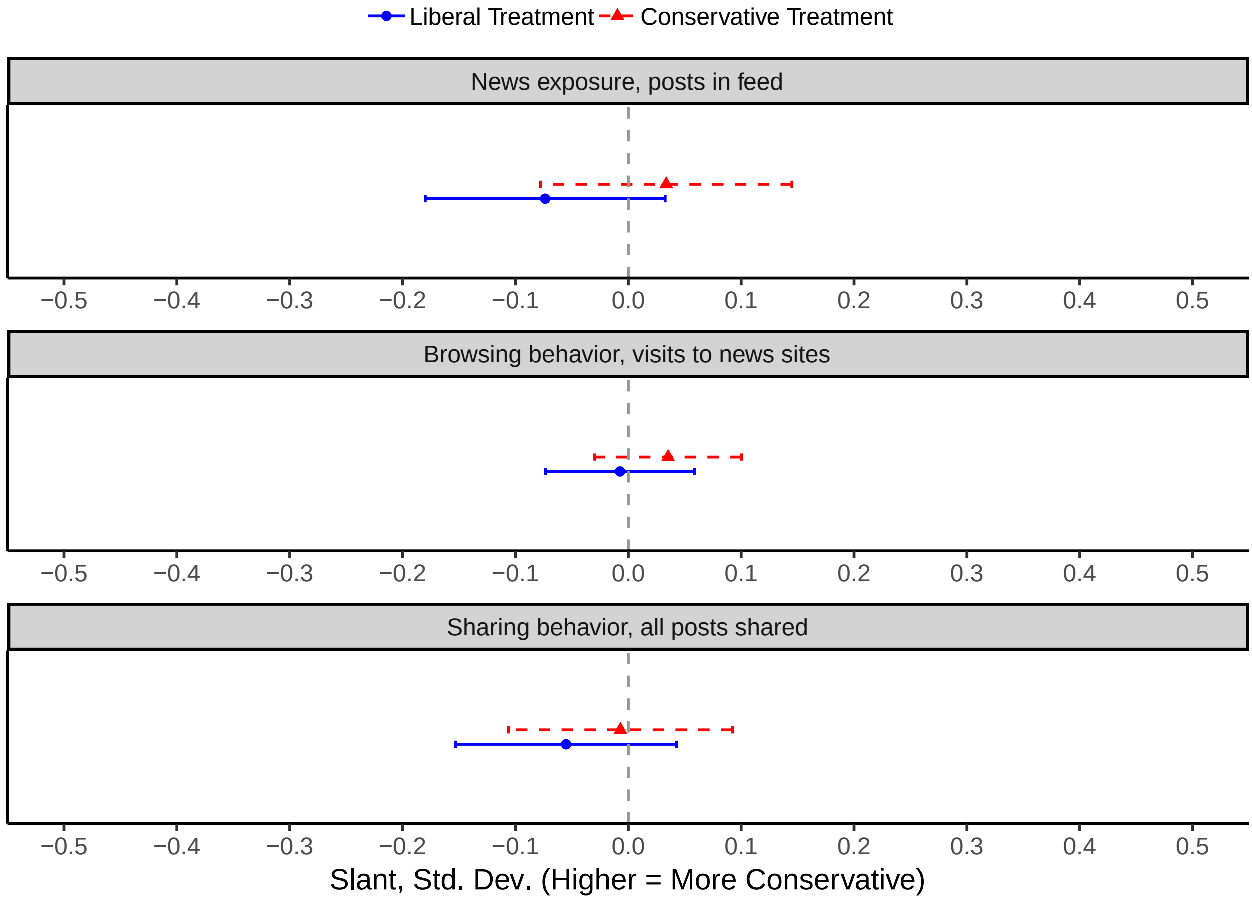



(b) Sharing Behavior



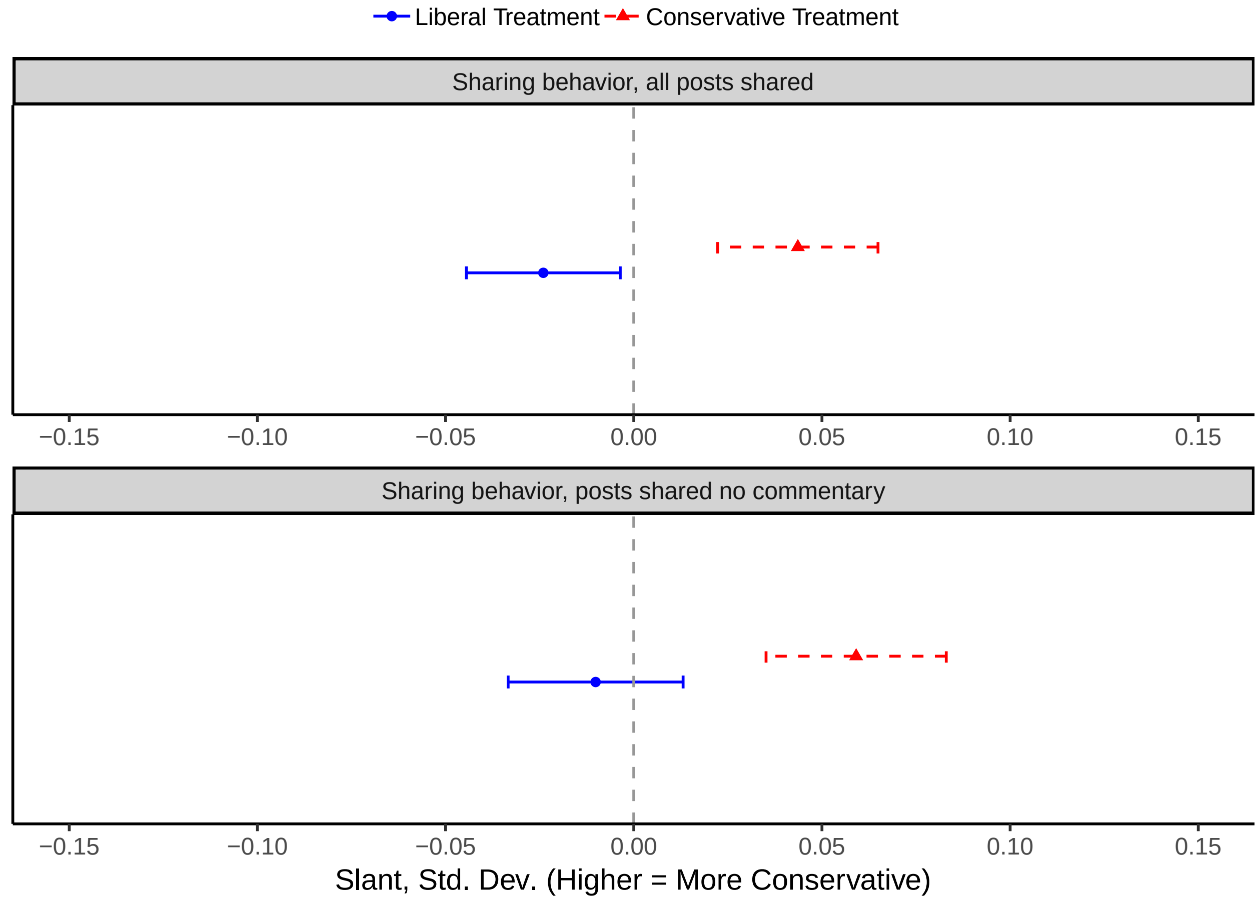

These figures show the difference between the effect of the liberal and conservative treatments on the mean slant over time. Each panel presents a series of regressions, where the dependent variable is the slant of outlets in a specific week. The regressions control for the outcome in baseline when it exists. In the x-axis, relative week 1 is a full week immediately following the intervention and relative week 0 is a full week immediately preceding the intervention. In the first figure, the data is based on 1,596 participants who kept the extension installed for at least six weeks following the intervention and the outcomes are the slant of posts observed in the Facebook feed and the slant of news sites visited. In the second figure, the data is based on 29,108 participants who provided access to posts they shared for at least six weeks following the intervention and the outcome is the slant of posts shared. Error bars reflect 90 percent confidence intervals.

Figure A.6: Effect of the Liberal and Conservative Treatments on Slant of News Consumption, Excluding each Participant's Eight Potential Experimental Outlets



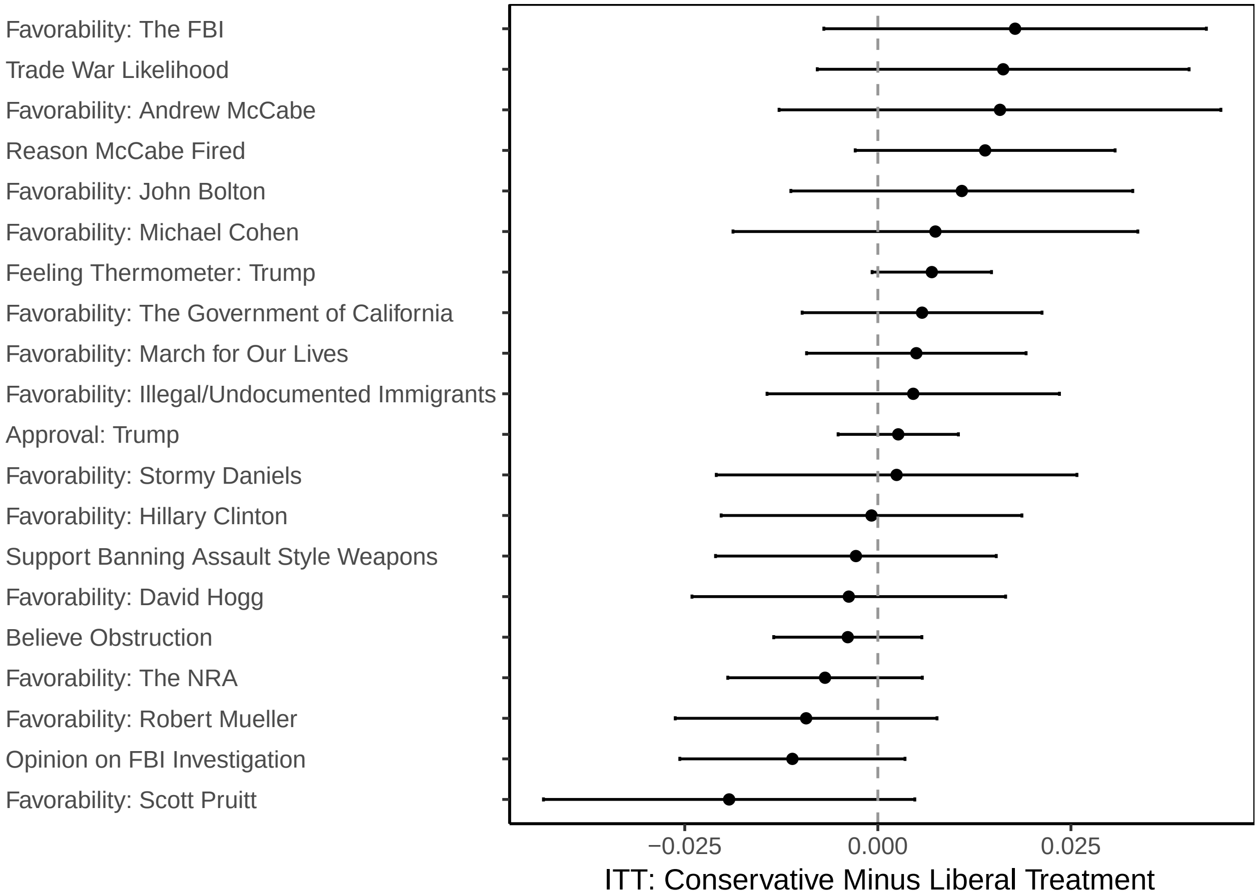

This figure shows the effect of the liberal and conservative treatments on the mean slant, in standard deviations, of all news participants engaged with, excluding the four potential liberal outlets and the four potential conservative outlets defined for each participant. Each row in the figure is estimated by regressing engagement with the four potential conservative outlets or four potential liberal outlets on the treatment. The regressions control for the outcome in baseline if it exists. The figure displays the slant for three outcomes: exposure to posts on Facebook (panel 1), news sites visited (panel 2), and posts shared (panel 3). The sample includes 1,702 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.7: Effects of the Liberal and Conservative Treatments on Slant of Posts Shared, Access Posts Subsample



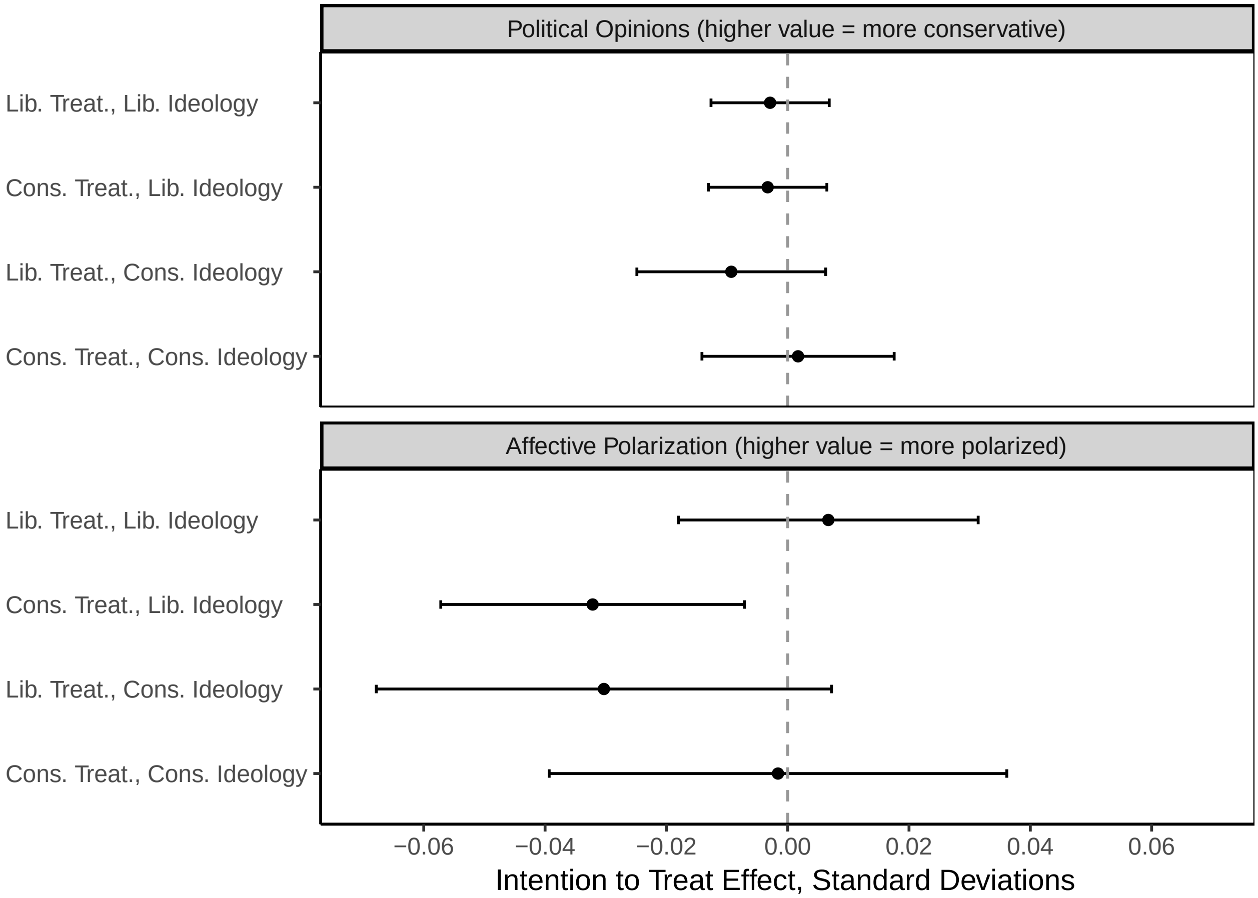

This figure shows the effect of the liberal and conservative treatments on the mean slant, in standard deviations, of all news participants shared. Each panel in the figure is estimated by regressing the slant of posts shared on the treatment. The first panel includes all posts and the second panel includes only posts that were shared without any commentary by the participant. The regressions control for the outcome in baseline. Data based on the access posts subsample: 34,568 participants who provided access to their posts for at least two weeks following the intervention. Error bars reflect 90 percent confidence intervals.

Figure A.8: Effect of the Treatment on Components of the Political Opinion Index



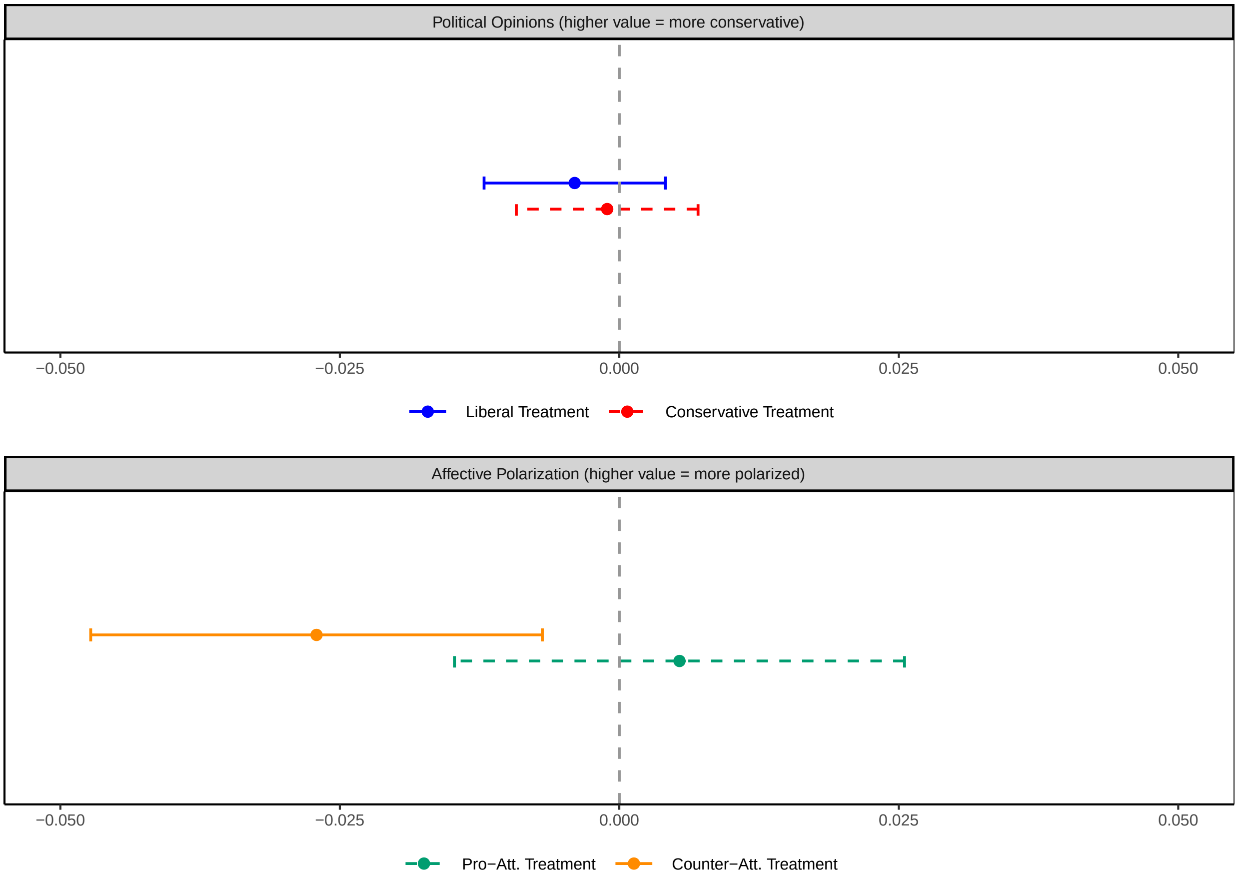

This figure shows the effect of the conservative treatment, compared to the liberal treatment on outcomes composing the political opinions index. Each row represents a separate regression as specified in Section 3.6. Outcomes are defined such that a higher value is associated with a more conservative 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.9: Effect of the Treatment on Political Opinions, by Baseline Ideology



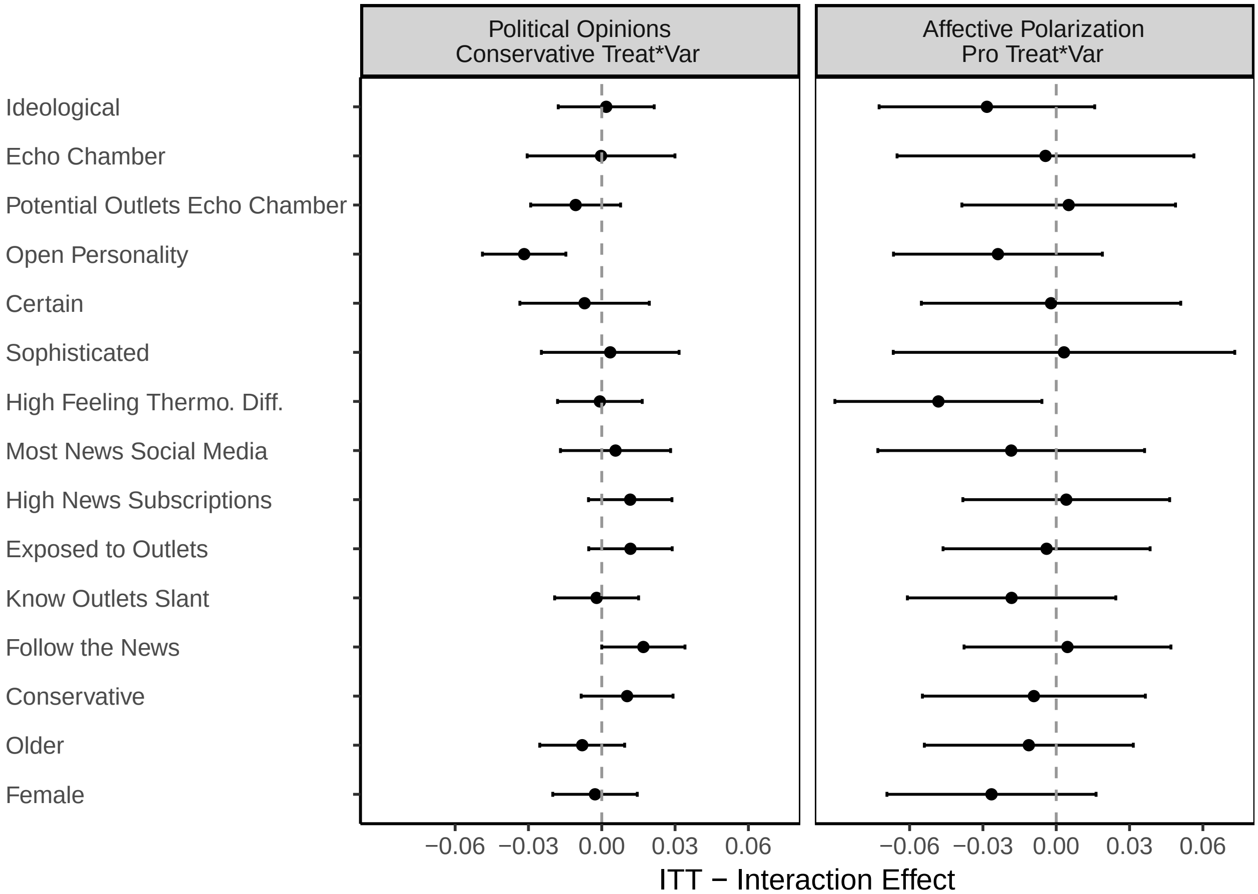

This figure shows the effect of the treatment among ideological subgroups based on the following model: $Y_i = \beta_1 T_i^L I_i^L + \beta_2 T_i^L I_i^C + \beta_3 T_i^C I_i^L + \beta_4 T_i^C I_i^C + I_\text{i} + \alpha X_i + \varepsilon_i$
where: $T_i^C, T_i^L$ are binary indicators for the conservative and liberal treatments, $I_i^C$,$I_i^L$ are binary indicators for whether the participants are conservative or liberal according to the baseline survey. The reference group is the control group. The controls are specified in Section 3.6. In the first panel, the x-axis is the intention to treat effect on the political opinions index, where a higher value is a more conservative outcome. In the second panel, the x-axis is the intention to treat effect on the affective polarization index, where a higher value is a more polarized outcome. Error bars reflect 90 percent confidence intervals.

Figure A.10: Effect of the Treatment by Treatment Arm



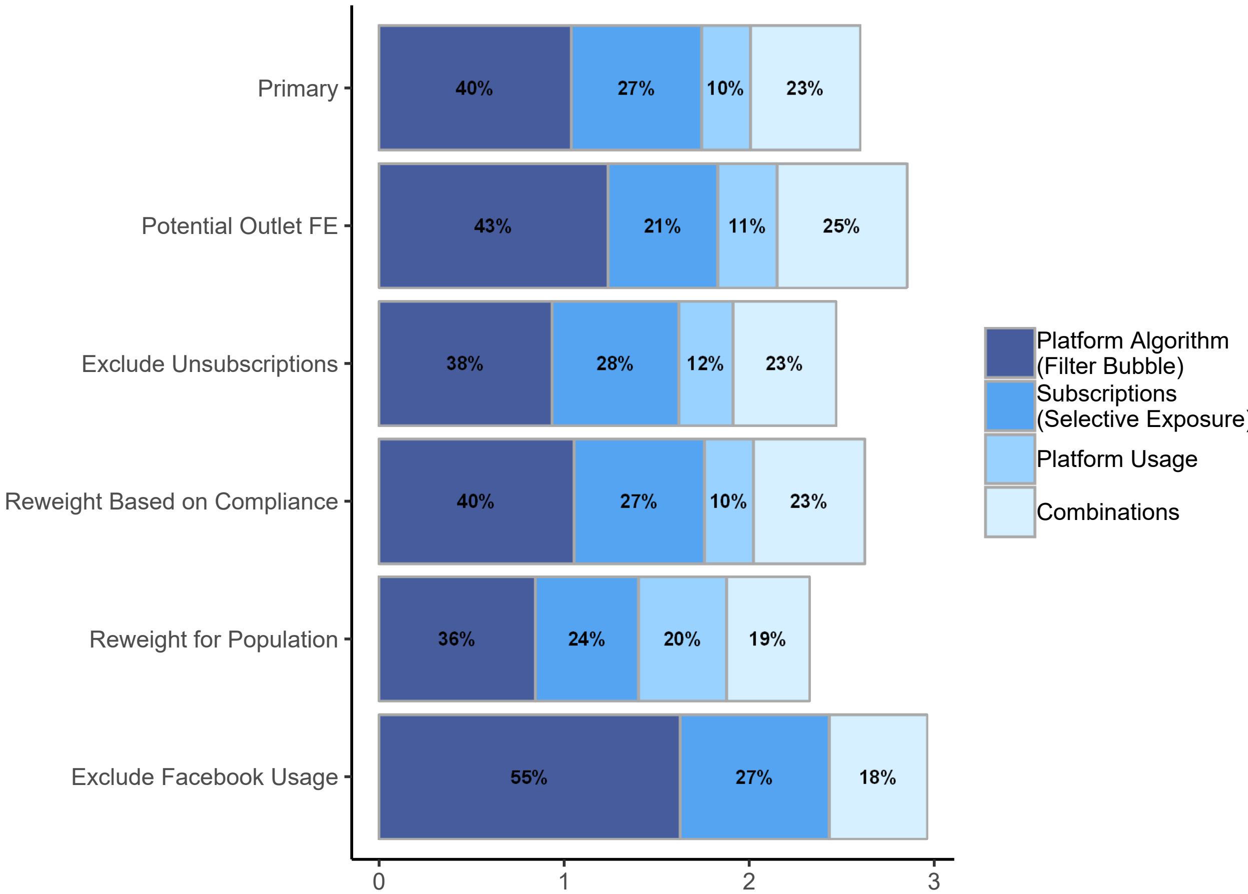

This figure shows the effect of each treatment arm on the political opinions index and affective polarization index. The indices are described in Section 3.4.2. The specification and controls are described in more detail in Section 3.6. Error bars reflect 90 percent confidence intervals.

Figure A.11: Heterogeneous Effects on Political Opinions and Affective Polarization




In the left panel, each row represents the $\beta$ 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. The control group is excluded so the reference category is the liberal treatment. A higher value means individuals were more likely to be persuaded by the treatment (they became more conservative as a result of the conservative treatment, compared to the liberal treatment).
In the right panel, each row presents the $\beta$ 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. The control group is excluded so the reference category is the counter-attitudinal treatment. 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 3.6. The definitions of the variables analyzed are described in Section D.2. Error bars reflect 90 percent confidence intervals.

Figure A.12: 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 9. The second row controls for the potential outlets defined for each participant. The third row excludes from the subscriptions cases where the participants unsubscribed from an outlet within two weeks. The fourth row reweights the participants in each treatment such that the compliers resemble the entire sample. The fifth row reweights the participants such that the entire sample resembles the US population. The sixth row excludes differences in platform usage between the groups. Each row is described in more detail in Section D.5.2.

Table A.1: List of Outlets Offered and Subscriptions

| Outlet | Group | Slant | Potential | Offered | Subscribed | Share |
|---|---|---|---|---|---|---|
| Fox News | Conservative | 0.78 | 32,559 | 10,839 | 1,387 | 0.13 |
| The Wall Street Journal | Conservative | 0.28 | 35,399 | 11,803 | 3,913 | 0.33 |
| The National Review | Conservative | 0.90 | 36,160 | 12,054 | 2,868 | 0.24 |
| The Washington Times | Conservative | 0.70 | 37,112 | 12,363 | 3,165 | 0.26 |
| The Daily Caller | Conservative | 0.87 | 4,519 | 1,470 | 309 | 0.21 |
| The Western Journal | Conservative | 0.90 | 1,529 | 509 | 146 | 0.29 |
| Washington Examiner | Conservative | 0.81 | 1,719 | 607 | 131 | 0.22 |
| Townhall | Conservative | 0.93 | 397 | 135 | 37 | 0.27 |
| The Conservative Tribune | Conservative | 0.89 | 203 | 72 | 31 | 0.43 |
| The Blaze | Conservative | 0.89 | 221 | 80 | 24 | 0.30 |
| Newsmax | Conservative | 0.77 | 114 | 32 | 13 | 0.41 |
| The New York Times | Liberal | -0.55 | 30,334 | 10,142 | 3,285 | 0.32 |
| HuffPost | Liberal | -0.62 | 31,920 | 10,639 | 2,303 | 0.22 |
| MSNBC | Liberal | -0.81 | 35,087 | 11,684 | 2,715 | 0.23 |
| Slate | Liberal | -0.68 | 35,201 | 11,734 | 2,937 | 0.25 |
| Washington Post | Liberal | -0.26 | 8,230 | 2,823 | 1,295 | 0.46 |
| Salon | Liberal | -0.88 | 5,117 | 1,668 | 572 | 0.34 |
| Daily Kos | Liberal | -0.90 | 2,013 | 661 | 222 | 0.34 |
| NPR | Liberal | -0.61 | 431 | 119 | 66 | 0.55 |
| Mother Jones | Liberal | -0.87 | 513 | 150 | 58 | 0.39 |
| The Atlantic | Liberal | -0.54 | 635 | 203 | 111 | 0.55 |
| 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 all 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. *Offered* is the number of participants in the baseline sample who were offered to subscribe to the outlet. *Subscribed* is the number of participants who subscribed to each outlet. *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.

Table A.2: Descriptive Statistics by Sample

| | Baseline Sample | Access Posts Subsample | Endline Survey Subsample | Extension Subsample |
|---|---|---|---|---|
| Ideology (-3, 3) | -0.61 | -0.61 | -0.70 | -0.95 |
| Ideology, Abs. Value (0, 3) | 1.75 | 1.75 | 1.80 | 1.81 |
| Feeling Therm., Rep. | 29.1 | 29.2 | 27.5 | 22.8 |
| Feeling Therm., Dem. | 47.0 | 47.0 | 47.8 | 51.2 |
| Feeling Therm., Difference | 50.2 | 50.3 | 50.3 | 51.0 |
| Difficult Pers., Difference | 1.92 | 1.92 | 1.96 | 1.92 |
| Most News Social Media | 0.18 | 0.18 | 0.17 | 0.16 |
| Took Survey Mobile | 0.67 | 0.67 | 0.63 | 0.00 |
| Total Subscriptions | 474 | 474 | 472 | 477 |
| News Outlets Subscriptions | 8.41 | 8.42 | 8.59 | 8.91 |
| Compliance | 0.50 | 0.51 | 0.57 | 0.76 |
| N | 37,483 | 34,568 | 17,629 | 1,838 |

This table presents descriptive statistics by subsample. *Baseline Sample* includes all participants. *Access-Posts Subsample* includes participants who provided access to posts they shared for at least two weeks. *Endline Survey Subsample* includes all participants who completed the baseline survey. *Extension Subsample* includes all 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 the feeling 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 point of view. For all other variables, see Table 2.

Table A.3: Balance Table by Assignment to the Pro-Attitudinal and Counter-Attitudinal Treatments

| Variable | Mean | | Difference Between Treatments | | |
|---|---|---|---|---|---|
| | Sample | US (2016), Ex. No Ideo. Leaning | Control - Pro. | Control - Counter. | Pro. - Counter. |
| **Baseline Survey** | | | | | |
| Ideology, Abs. Value (0, 3) | 1.75 | 1.31 | 0.00 | -0.00 | -0.00 |
| Democrat | 0.38 | 0.37 | 0.01 | 0.00 | -0.01 |
| Republican | 0.17 | 0.30 | 0.00 | -0.01 | -0.01 |
| Independent | 0.37 | 0.29 | -0.01* | 0.00 | 0.01** |
| Vote Support Clinton | 0.53 | | -0.00 | -0.00 | 0.00 |
| Vote Support Trump | 0.26 | | 0.00 | 0.00 | 0.00 |
| Feeling Therm., Difference | 50.2 | 38.4 | 0.35 | 0.40 | 0.05 |
| Difficult Pers., Difference | 1.92 | | 0.03 | 0.01 | -0.02 |
| Facebook Echo Chamber | 1.18 | | 0.00 | -0.01 | -0.01 |
| Follows News | 3.35 | 2.48 | 0.01 | 0.01 | 0.01 |
| Most News Social Media | 0.18 | 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 | 47.7 | 47.7 | 0.02 | 0.07 | 0.05 |
| Total Subscriptions | 474 | | 8.02 | 2.38 | -5.63 |
| News Outlets Slant, Abs. Value | 0.55 | | -0.00 | -0.00 | -0.00 |
| Access Posts, Pre-Treatment | 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 | | | 1.27 | 0.84 | 1.00 |
| P-value | | | [0.17] | [0.69] | [0.47] |

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 Americans adults for whom an ideological leaning can be defined (individuals who identify or lean toward the Democratic or Republican party, identify as liberal or conservative, or preferred one of the two major presidential candidates according to the pre-election version of the 2016 American National Election Survey). This is the relevant comparison group since participants for which an ideological leaning is not defined are excluded when analyzing the effects of the pro- and counter-attitudinal treatments. *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 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. *News Outlets Slant, Abs. Value* is the absolute value of the mean slant of all outlets participants subscribed to on Facebook in baseline. Slant range from -1 to 1 and is based on Bakshy et al. (2015). For all other variables see Table 2. *p<0.1 **p<0.05 ***p<0.01

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

| Variable | Mean | | | Difference Between Treatments | | |
|---|---|---|---|---|---|---|
| | Sample | US (2016) | FB Users (2018) | Control - Lib. | Control - Cons. | Cons. - Lib. |
| **Baseline Survey** | | | | | | |
| Ideology (-3, 3) | -0.70 | 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. | 27.5 | 43.1 | | 0.22 | -0.05 | 0.27 |
| Feeling Therm., Dem. | 47.8 | 48.7 | | 0.41 | 0.68 | -0.27 |
| Difficult Pers., Rep. (1, 5) | 3.18 | | | 0.04 | 0.01 | 0.03 |
| 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 | 48.8 | 47.3 | 42.9 | 0.56* | -0.32 | 0.87*** |
| Total Subscriptions | 472 | | | 3.00 | 14.93 | -11.93 |
| News Outlets Slant (-1, 1) | -0.22 | | | 0.00 | -0.01 | 0.01 |
| Access Posts, Pre-Treatment | 0.98 | | | -0.00 | 0.00 | -0.00 |
| F-Test | | | | 1.09 | 1.00 | 1.28 |
| P-Value | | | | [0.35] | [0.46] | [0.18] |

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 by Assignment to the Pro-Attitudinal and Counter-Attitudinal Treatment, Among Participants Who Completed the Follow-up Survey

| | Mean | | Difference Between Treatments | | |
|---|---|---|---|---|---|
| Variable | Sample | US (2016), Ex. No Ideo. Leaning | Control - Pro. | Control - Counter. | Pro. - Counter. |
| **Baseline Survey** | | | | | |
| Ideology, Abs. Value (0, 3) | 1.80 | 1.31 | -0.00 | 0.00 | 0.00 |
| Democrat | 0.40 | 0.37 | 0.02* | 0.01 | -0.01 |
| Republican | 0.16 | 0.30 | 0.00 | 0.00 | -0.00 |
| Independent | 0.36 | 0.29 | -0.02** | -0.00 | 0.01 |
| Vote Support Clinton | 0.55 | | -0.00 | 0.00 | 0.00 |
| Vote Support Trump | 0.25 | | 0.00 | 0.01 | 0.01 |
| Feeling Therm., Difference | 50.3 | 38.4 | 0.92* | 1.10** | 0.18 |
| Difficult Pers., Difference | 1.96 | | 0.05 | 0.04 | -0.01 |
| Facebook Echo Chamber | 1.20 | | 0.00 | 0.00 | -0.00 |
| Follows News | 3.38 | 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 | 48.8 | 47.7 | 0.11 | 0.20 | 0.09 |
| Total Subscriptions | 472 | | 6.57 | 2.18 | -4.40 |
| News Outlets Slant, Abs. Value | 0.56 | | -0.00 | 0.00 | 0.00 |
| Access Posts, Pre-Treatment | 0.98 | | -0.00 | 0.00 | 0.00 |
| F-Test | | | 0.59 | 0.71 | 0.56 |
| P-value | | | [0.92] | [0.81] | [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 | Pro-Attitudinal | | Counter-Attitudinal | | Liberal | | Conservative | |
|---|---|---|---|---|---|---|---|---|---|
| | | Comply | Non-Comply | Comply | Non-Comply | Comply | Non-Comply | Comply | Non-Comply |
| Ideology (-3, 3) | -0.62 | -0.87 | -0.34 | -1.04 | -0.29 | -1.13 | -0.10 | -0.72 | -0.52 |
| Ideology, Abs. Value (0, 3) | 1.80 | 1.84 | 1.75 | 1.78 | 1.82 | 1.78 | 1.72 | 1.75 | 1.75 |
| Democrat | 0.40 | 0.44 | 0.33 | 0.45 | 0.35 | 0.47 | 0.29 | 0.40 | 0.37 |
| Republican | 0.17 | 0.15 | 0.20 | 0.13 | 0.22 | 0.11 | 0.24 | 0.16 | 0.18 |
| Independent | 0.35 | 0.35 | 0.38 | 0.36 | 0.35 | 0.36 | 0.38 | 0.37 | 0.36 |
| Vote Support Clinton | 0.54 | 0.60 | 0.48 | 0.63 | 0.47 | 0.64 | 0.41 | 0.55 | 0.50 |
| Vote Support Trump | 0.27 | 0.23 | 0.33 | 0.17 | 0.35 | 0.15 | 0.37 | 0.25 | 0.28 |
| Feeling Therm., Difference | 50.5 | 51.0 | 49.0 | 48.9 | 51.0 | 50.4 | 49.7 | 49.7 | 50.5 |
| Difficult Pers., Difference | 1.93 | 1.97 | 1.82 | 1.88 | 1.95 | 1.93 | 1.89 | 1.92 | 1.88 |
| Facebook Echo Chamber | 1.20 | 1.23 | 1.15 | 1.22 | 1.19 | 1.23 | 1.14 | 1.19 | 1.18 |
| Most News Social Media | 0.17 | 0.17 | 0.17 | 0.19 | 0.17 | 0.18 | 0.17 | 0.18 | 0.17 |
| Took Survey Mobile | 0.67 | 0.66 | 0.70 | 0.66 | 0.67 | 0.67 | 0.68 | 0.65 | 0.69 |
| Female | 0.52 | 0.56 | 0.48 | 0.59 | 0.46 | 0.58 | 0.46 | 0.56 | 0.47 |
| Age | 47.9 | 48.6 | 47.1 | 47.5 | 48.2 | 47.7 | 47.3 | 48.1 | 47.6 |
| Total Subscriptions | 476 | 496 | 433 | 526 | 431 | 516 | 432 | 507 | 433 |
| News Outlets Subscriptions | 8.46 | 9.13 | 7.74 | 9.10 | 8.08 | 9.10 | 7.76 | 8.99 | 7.87 |
| N | 12,101 | 6,739 | 5,356 | 5,505 | 6,617 | 6,278 | 6,214 | 6,311 | 6,180 |

This table presents descriptive statistics on compliance by treatment arm for the entire baseline sample. For an explanation on each variable see Table 2.

Table A.7: Effects of the Treatments on News Exposure, News Sites Visited and Sharing Behavior, Two Weeks Following the Intervention, Poisson Regression

| | Pro-Att. Outlets News Exposure | Pro-Att. Outlets Browsing Behavior | Pro-Att. Outlets Sharing Behavior | Counter-Att. Outlets News Exposure | Counter-Att. Outlets Browsing Behavior | Counter-Att. Outlets Sharing Behavior |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Pro-Att. Treat. | $1.34^{***}$ | $0.30^{**}$ | $0.56^{***}$ | $0.33^{**}$ | 0.20 | 0.30 |
| | (0.13) | (0.14) | (0.21) | (0.16) | (0.25) | (0.31) |
| Counter-Att. Treat. | $-0.06$ | $-0.03$ | 0.25 | $2.50^{***}$ | $0.54^{***}$ | $1.37^{***}$ |
| | (0.13) | (0.14) | (0.21) | (0.16) | (0.19) | (0.31) |
| Pro-Att. exponentiated | 3.83 | 1.35 | 1.76 | 1.39 | 1.22 | 1.34 |
| Counter-Att. exponentiated | 0.94 | 0.97 | 1.28 | 12.12 | 1.71 | 3.92 |
| Observations | 1,651 | 1,651 | 1,651 | 1,651 | 1,651 | 1,651 |

This table presents the effects of the pro and counter-attitudinal treatments on engagement with the potential pro- and counter-attitudinal potential outlets in the two weeks following the intervention, estimated using Poisson regressions. The dependent variables are exposure to posts in the Facebook feed, visits to news sites and posts shared. The sample includes participants with a liberal or conservative ideological leaning who installed the extension and provided permissions 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.8: Effect on Slant by Subsample

| | News Exposure | | | Browsing Behavior | | | Shared Posts | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Liberal Treatment | $-0.242^{***}$ | $-0.240^{***}$ | $-0.201^{***}$ | $-0.093^{**}$ | $-0.082^{**}$ | $-0.103^{**}$ | $-0.024^{*}$ | $-0.123^{**}$ | $-0.063$ |
| | (0.060) | (0.063) | (0.073) | (0.037) | (0.039) | (0.046) | (0.012) | (0.056) | (0.066) |
| Conservative Treatment | $0.352^{***}$ | $0.361^{***}$ | $0.455^{***}$ | $0.098^{**}$ | $0.101^{**}$ | $0.104^{**}$ | $0.044^{***}$ | 0.046 | 0.116 |
| | (0.067) | (0.069) | (0.082) | (0.040) | (0.041) | (0.050) | (0.013) | (0.060) | (0.073) |
| Cons. Treat. - Lib. Treat. | $0.59^{***}$ | $0.60^{***}$ | $0.66^{***}$ | $0.19^{***}$ | $0.18^{***}$ | $0.21^{***}$ | $0.07^{***}$ | $0.17^{***}$ | $0.18^{**}$ |
| | (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 | 1,557 | 1,434 | 1,011 | 1,787 | 1,654 | 1,167 | 18,322 | 982 | 687 |

This table presents the effect of the treatment on the slant of outlets participants engaged with by the subsample analyzed. 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., all participants who installed the extension for at least two weeks. *Posts Subsample* refers to the access posts subsample, i.e. all participants who provide permissions to access their posts for at least two weeks. *Ext + Posts Subsample* refers to all participants in both these subsamples. *Ext + Posts + Endline Subsample* refers to all 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.9: Primary Outcomes, Controlling for Covariates

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

| | (1) | (2) | (3) |
|---|---|---|---|
| Conservative Treatment | 0.011 | 0.001 | −0.001 |
| | (0.018) | (0.006) | (0.005) |
| Liberal Treatment | −0.007 | −0.005 | −0.004 |
| | (0.018) | (0.006) | (0.005) |
| Cons. Treatment - Lib. Treatment | 0.017 | 0.006 | 0.003 |
| | (0.019) | (0.006) | (0.005) |
| Common Controls | | X | X |
| Baseline Political Opinions Controls | | | X |
| Observations | 17,629 | 17,123 | 17,123 |

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

| | (1) | (2) | (3) |
|---|---|---|---|
| Pro-Att. Treatment | −0.021 | −0.003 | 0.005 |
| | (0.019) | (0.015) | (0.012) |
| Counter-Att. Treatment | $-0.055^{***}$ | $-0.039^{**}$ | $-0.027^{**}$ |
| | (0.019) | (0.015) | (0.012) |
| Pro-Att. Treat. - Counter-Att. Treatment | $0.033^{*}$ | $0.035^{**}$ | $0.032^{***}$ |
| | (0.019) | (0.015) | (0.012) |
| Common Controls | | X | X |
| Baseline Affective Polarization Controls | | | X |
| Observations | 16,889 | 16,889 | 16,889 |

These tables present the effects of the treatments on the political opinions index and the affective polarization index. 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. The specification and controls are described in more detail in Section 3.6. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.10: Effect of the Treatment on the Affective Polarization Index, Excluding Specific Measures

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Counter-Att. Treatment | $-0.027^{**}$ | $-0.042^{***}$ | $-0.028^{*}$ | $-0.041^{***}$ | $-0.046^{***}$ | $-0.032^{**}$ |
| | (0.012) | (0.016) | (0.016) | (0.015) | (0.016) | (0.015) |
| Pro-Att. Treatment | 0.005 | $-0.007$ | 0.0001 | $-0.004$ | $-0.006$ | 0.001 |
| | (0.012) | (0.016) | (0.015) | (0.015) | (0.016) | (0.015) |
| Counter - Pro | $-0.032^{***}$ | $-0.035^{**}$ | $-0.029^{*}$ | $-0.037^{**}$ | $-0.040^{**}$ | $-0.033^{**}$ |
| | (0.012) | (0.016) | (0.016) | (0.015) | (0.016) | (0.015) |
| Excluded Measure | | Feeling Thermometer | Difficult Perspective | Consider Perspective | Party Ideas | Marry Opposing Party |
| Observations | 16,889 | 16,889 | 16,889 | 16,889 | 16,888 | 16,889 |

This table presents the effect of the treatment 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 3.6. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.11: Primary Outcomes, According to Outlets Offered

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

| | (1) | (2) | (3) |
|---|---|---|---|
| Liberal Treatment | −0.004 | −0.007 | −0.005 |
| | (0.005) | (0.007) | (0.005) |
| Conservative Treatment | −0.001 | −0.007 | −0.002 |
| | (0.005) | (0.007) | (0.005) |
| Cons. Treat - Lib. Treat | 0.003 | 0.001 | 0.003 |
| | (0.005) | (0.007) | (0.005) |
| Controls | X | X | X |
| Include Participants Who Already Subscribed To A Primary Outlet In Baseline | X | | X |
| Potential Outlets FE | | | X |
| Observations | 17,123 | 9,257 | 17,123 |

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

| | (1) | (2) | (3) |
|---|---|---|---|
| Pro-Att. Treatment | 0.005 | −0.002 | 0.004 |
| | (0.012) | (0.016) | (0.013) |
| Counter-Att. Treatment | −0.027** | −0.031* | −0.032** |
| | (0.012) | (0.016) | (0.013) |
| Pro-Att. Treat. - Counter-Att. Treat | 0.032*** | 0.029* | 0.036*** |
| | (0.012) | (0.017) | (0.013) |
| Controls | X | X | X |
| Include Participants Who Already Subscribed To A Primary Outlet In Baseline | X | | X |
| Potential Outlets FE | | | X |
| Observations | 16,889 | 9,127 | 16,889 |

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 3.6. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.12: Primary Outcomes, by Subsample

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

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Liberal Treatment | −0.004 | −0.005 | −0.010 | −0.019 |
| | (0.005) | (0.005) | (0.018) | (0.019) |
| Conservative Treatment | −0.001 | −0.003 | 0.004 | 0.002 |
| | (0.005) | (0.005) | (0.018) | (0.018) |
| Cons. Treat - Lib. Treat | 0.003 | 0.003 | 0.015 | 0.020 |
| | (0.005) | (0.005) | (0.018) | (0.018) |
| Controls | X | X | X | X |
| Sample | Endline | Endline+Posts | Endline+Ext | Endline+Posts+Ext |
| Observations | 17,123 | 15,858 | 1,252 | 1,162 |

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

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Pro-Att. Treatment | 0.005 | 0.007 | 0.015 | 0.027 |
| | (0.012) | (0.013) | (0.044) | (0.046) |
| Counter-Att. Treatment | −0.027** | −0.027** | −0.069 | −0.054 |
| | (0.012) | (0.013) | (0.043) | (0.045) |
| Pro-Att. Treat. - Counter-Att. Treat | 0.032*** | 0.034*** | 0.085* | 0.081* |
| | (0.012) | (0.013) | (0.043) | (0.044) |
| Controls | X | X | X | X |
| Sample | Endline | Endline+Posts | Endline+Ext | Endline+Posts+Ext |
| Observations | 16,889 | 15,635 | 1,242 | 1,152 |

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 only participants who also provided participants to access their posts for at least two weeks (participants in the endline survey subsample and the access posts subsample). Column (3) includes only participants who installed the extension for at least two weeks (participants in the endline survey subsample and the extension subsample. Column (4) includes only participants who both provided access to posts and installed the extension (participants in all subsamples). The specification and controls are described in more detail in Section 3.6. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.13: Effect of Exposure to Pro- and Counter-Attitudinal News on Affective Polarization

(a) Cross-Sectional Correlation in Control Group

| | OLS | OLS |
|---|---|---|
| | Affective Polarization | |
| | (1) | (2) |
| FB Counter-Att. Share, Std. Dev. | $-0.384^{***}$ | |
| | (0.052) | |
| FB Congruence Scale, Std. Dev. | | $0.406^{***}$ |
| | | (0.054) |
| Data | Control Group | Control Group |
| Observations | 353 | 353 |

(b) Causal Effect Based on Experimental Variation

| | IV | |
|---|---|---|
| | Affective Polarization | |
| | (1) | (2) |
| FB Counter-Att. Share, Std. Dev. | $-0.121^{*}$ | |
| | (0.066) | |
| FB Congruence Scale, Std. Dev. | | $0.099^{*}$ |
| | | (0.058) |
| Controls | X | X |
| First Stage F | 66.22 | 64.55 |
| Share of Correlation in Control Group | 0.31 | 0.24 |
| Observations | 1,071 | 1,071 |

These tables measure the association between exposure to pro- and counter-attitudinal news and affective polarization. The tables use two summary statistics. *Counter-Att. Share* is the share of news form counter-attitudinal outlets the participant was exposed to on Facebook between the baseline and endline surveys, among all news from pro- and counter-attitudinal outlets. The *Congruence Scale* is the mean slant of all news exposed to on Facebook, multiplied by (-1) for liberal participants. Sub-table (a) 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). Sub-table (b) shows the results of IV regressions, where the independent variables are instrumented with the treatment. The regressions control for the covariates specified in Section 3.6 and include all participants who installed the Chrome extension and completed the endline survey. The row titled *Share of Correlation in Control Group* divides the causal effect found in sub-table (b) by the correlation found in sub-table (a). Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.14: Counterfactuals

| | IV | | | |
|---|---|---|---|---|
| | Affective Pol., Std. Dev. | Feeling Thermo., Degrees | Affective Pol., Std. Dev. | Feeling Thermo., Degrees |
| | (1) | (2) | (3) | (4) |
| FB Counter-Att. Share | $-0.522^{*}$ | | $-11.413$ | |
| | (0.285) | | (7.677) | |
| FB Congruence Scale | | $0.446^{*}$ | | 9.911 |
| | | (0.260) | | (7.048) |
| Controls | X | X | X | X |
| Control Group: Counter Share | 0.17 | | | |
| Effect of Balanced Facebook Feed | -0.17 | | -3.76 | |
| Control Group: Congruence | | 0.33 | | |
| Effect of Balanced Facebook Feed | | -0.15 | | -3.27 |
| Control Group: Diff in Counter Share | 0.04 | | | |
| Effect of Equating Counter Share | -0.02 | | -0.46 | |
| Control Group: Diff in Congruence | | -0.06 | | |
| Effect of Equating Congruence | | -0.03 | | -0.58 |
| Observations | 1,071 | 1,071 | 1,030 | 1,030 |

The table shows how affective polarization would have changed under two counterfactuals. *FB Counter-Att. Share* shows the effect of the share of counter-attitudinal news on Facebook on the affective polarization index and the feeling thermometer measures. *FB Congruence Scale* shows the effect of the congruence scale of Facebook news on the outcomes. The regressions are IV regression with the treatment as the instrument and controlling for the covariates specified in Section 3.6. *Control Group: Counter Share* and *Control Group: Congruence* show the share of counter-attitudinal posts and the congruence scale in the control group Facebook feed, respectively. *Effect of Balanced Facebook Feed* presents the results of the first counterfactual and shows how the outcomes would have decreased if the share of counter-attitudinal posts increased to 50% or the congruence scale decreased to 0. *Control Group: Diff in Counter Share* is the control group differences between the share of counter-attitudinal news in the Facebook feed and in news sites not visited through Facebook. *Control Group: Diff in Congruence* is the difference in the congruence scale. *Effect of Equating Counter Share* and *Effect of Equating Congruence* present the results of the second counterfactual and show how the outcomes would have decreased if news exposure on Facebook was similar to online news sites not visited through Facebook. The data includes all posts participants in the extension subsample were exposed to between the baseline and endline surveys. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.15: Primary Outcomes Using Different Index Methods

(a) Political Opinions

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Liberal Treatment | −0.004 | −0.005 | −0.031 | −0.004 | −0.003 |
| | (0.005) | (0.017) | (0.079) | (0.007) | (0.005) |
| Conservative Treatment | −0.001 | 0.022 | −0.051 | 0.006 | 0.004 |
| | (0.005) | (0.017) | (0.080) | (0.007) | (0.005) |
| Cons. - Lib. Treatment | 0.003 | 0.027 | 0.010 | -0.020 | 0.006 |
| | (0.005) | (0.017) | (0.007) | (0.053) | (0.005) |
| Controls | X | X | X | X | X |
| Index Method | Standard | Inv-Cov | Inv-Cov | Inv-Cov | Inv-Cov |
| Include Responders With Missing Outcomes | - | No | Yes | No | Yes |
| Replace Negative Weights With 0 | - | Yes | Yes | No | No |
| Observations | 17,123 | 9,247 | 17,123 | 9,247 | 17,123 |

(b) Affective Polarization

| | (1) | (2) | (3) |
|---|---|---|---|
| Pro-Att. Treatment | 0.005 | 0.005 | 0.001 |
| | (0.012) | (0.017) | (0.010) |
| Counter-Att. Treatment | −0.027** | −0.030* | −0.026*** |
| | (0.012) | (0.017) | (0.010) |
| Pro-Att. Treat. - Counter-Att. Treatment | 0.032*** | 0.035** | 0.026*** |
| | (0.012) | (0.017) | (0.010) |
| Controls | X | X | X |
| Index Method | Standard | Inv-Cov | Inv-Cov |
| Include Responders With Missing Outcomes | - | No | Yes |
| Observations | 16,889 | 10,053 | 16,889 |

These tables estimate the effects of the treatments on the primary outcomes using different summary indexes. Column (1) uses equal weights for all outcomes included in the index. Column (2) uses inverse covariates weights and includes participants that have no missing values for any of the index components. In column (3), inverse-covariance weights are used for all participants with non-missing outcomes. For participants with missing outcomes, the weights are renormalized to sum to 1, 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 1. The specification and controls are described in Section 3.6. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.16: Effect of the Treatment 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** | −0.027** | −0.010 |
| | (0.014) | (0.012) | (0.018) |
| Counter-Att. Treatment - Pro-Att. Treat. | 0.034** | 0.032*** | 0.008 |
| | (0.014) | (0.012) | (0.019) |
| Controls | X | X | X |
| Observations | 17,154 | 16,889 | 16,635 |

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 and column (3) include the three behavioral components. The specification and controls are described in Section 3.6. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.17: The Association Between the Slant of News Consumed and the Interaction of Facebook Usage and Consumer's Ideology

| | Site Slant, Std. Dev. | Site Slant, Std. Dev. | Mean Slant, Std. Dev. | Mean Slant, Std. Dev. | Mean Slant, Std. Dev. |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Rep. Vote | 0.821*** | | 0.213*** | | |
| | (0.097) | | (0.027) | | |
| FB News Ref. * Rep. Vote | 0.932*** | 0.415*** | | | |
| | (0.182) | (0.075) | | | |
| FB News Ref. Share * Rep. Vote | | | 2.603*** | 0.874*** | |
| | | | (0.211) | (0.106) | |
| FB Visits Share * Rep. Vote | | | | | 1.213*** |
| | | | | | (0.451) |
| Unit of Observation | Site | Site | Ind. | Ind.*Month | Ind.*Month |
| Individual FE | | X | | X | X |
| Month FE | | X | | X | X |
| Demographics | | | X | | |
| Observations | 2,181,674 | 2,181,674 | 57,839 | 263,285 | 263,285 |

This table shows the association between the slant of news consumed and the interaction of ideology and Facebook usage. *Rep. Vote* is the 2008 Republican vote share at each individual's zip code. The slant of outlets is based on Bakshy et al. (2015) and is standardized. In columns (1)-(2) each observation is a website visited, *FB News Ref.* is a visit to a news site where the referring domain is "facebook.com" and the dependent variable is the mean slant of sites visited calculated based on Bakshy et al. (2015), where a higher slant is more conservative. In column (3), each observation is an individual, and in columns (4)-(5), each observation is an individual by month. In columns (3)-(5), the dependent variable is the mean slant of all sites visited by the individual or by the individual at a specific month. *FB News Ref. Share* is the share of news sites an individual visited through Facebook and *FB Visits Share* is the share of visits to Facebook among all websites visited. The sample is based on the 2017 Comscore Web Behavior Database Panel and includes all individuals who visited at least two news sites through Facebook and two news sites through other means. Standard errors are clustered at the individual level. *p<0.1 **p<0.05 ***p<0.01

Table A.18: The Association Between the Absolute Value of the Slant of News Consumed and Facebook Usage

| | Site Slant Abs. Value, Std. Dev. | Site Slant Abs. Value, Std. Dev. | Mean Slant Abs. Value, Std. Dev. | Mean Slant Abs. Value, Std. Dev. | Mean Slant Abs. Value, Std. Dev. |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| FB News Ref. | 0.474*** | 0.335*** | | | |
| | (0.025) | (0.013) | | | |
| FB News Ref. Share | | | 0.885*** | 0.496*** | |
| | | | (0.033) | (0.018) | |
| FB Visits Share | | | | | 0.463*** |
| | | | | | (0.085) |
| Unit of Observation | Site | Site | Ind. | Ind.*Month | Ind.*Month |
| Individual FE | | X | | X | X |
| Month FE | | X | | X | X |
| Demographics | | | X | | |
| Observations | 2,260,860 | 2,260,860 | 59,707 | 272,236 | 272,236 |

This table shows the association between the absolute slant of news consumed and Facebook usage. In columns (1)-(2) each observation is a website visited, *FB News Ref.* is a visit to a news site where the referring domain is "facebook.com" and the dependent variable is the absolute value of the slant of all sites visited, calculated based on Bakshy et al. (2015). In column (3), each observation is an individual and in columns (4)-(5), each observation is an individual by month. In columns (3)-(5), the dependent variable is the mean absolute value of the slant of all sites visited by the individual or by the individual at a specific month. *FB News Ref. Share* is the share of news sites an individual visited through Facebook and *FB Visits Share* is the share of visits to Facebook among all websites visited. The sample is based on the 2017 Comscore Web Behavior Database Panel and includes all individuals who visited at least two news sites through Facebook and two news sites through other means. Standard errors are clustered at the individual level. *p<0.1 **p<0.05 ***p<0.01

Table A.19: Primary Outcomes When Samples Is Reweighted to Match the US Population

(a) Political Opinions

| | (1) | (2) |
|---|---|---|
| Liberal Treatment | −0.004 | −0.004 |
| | (0.005) | (0.007) |
| Conservative Treatment | −0.001 | 0.00001 |
| | (0.005) | (0.008) |
| Cons. Treat - Lib. Treat | 0.003 | 0.004 |
| | (0.005) | (0.008) |
| Controls | X | X |
| Reweighted | | X |
| Observations | 17,123 | 17,123 |

(b) Affective Polarization

| | (1) | (2) |
|---|---|---|
| Pro-Att. Treatment | 0.005 | 0.019 |
| | (0.012) | (0.020) |
| Counter-Att. Treatment | $-0.027^{**}$ | −0.014 |
| | (0.012) | (0.022) |
| Pro-Att. Treat. - Counter-Att. Treat | $0.032^{***}$ | 0.033 |
| | (0.012) | (0.020) |
| Controls | X | X |
| Reweighted | | X |
| Observations | 16,889 | 16,889 |

These tables estimate the effect of the treatment 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 feeling toward their party and the opposing party, age and the share of females. This analysis is discussed in Appendix D.3. The specification and controls are described in Section 3.6. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.20: Effect of the Treatment 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 | 0.005 | −0.004 | 0.008 | 0.001 | 0.012 | −0.002 | 0.002 |
| | (0.006) | (0.007) | (0.006) | (0.008) | (0.005) | (0.009) | (0.010) | (0.006) |
| Conservative Treatment | −0.002 | 0.001 | −0.002 | 0.021*** | 0.008 | −0.001 | −0.009 | 0.001 |
| | (0.006) | (0.007) | (0.006) | (0.008) | (0.005) | (0.009) | (0.010) | (0.006) |
| Cons. Treat - Lib. Treat | 0.00 | -0.00 | 0.00 | 0.01* | 0.01 | -0.01 | -0.01 | -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 | 17,123 | 16,924 | 17,123 | 16,957 | 15,735 | 13,513 | 11,833 | 15,245 |

This table estimates the effect of the treatment 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.21: Effect of the Treatment on Exposure to Words on the Facebook Feed

| | Michael Cohen | Clark Shooting | Louis Farrakhan | Clinton Speech |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Liberal Treatment | 1.830*** | 1.170*** | 0.165 | 0.039 |
| | (0.589) | (0.357) | (0.120) | (0.043) |
| Conservative Treatment | 0.670 | 0.140 | 0.378*** | 0.081** |
| | (0.432) | (0.265) | (0.103) | (0.033) |
| Cons. Treat - Lib. Treat | -1.16* | -1.03*** | 0.21* | 0.04 |
| | (0.60) | (0.33) | (0.13) | (0.05) |
| Controls | X | X | X | X |
| Expected Effect | Lib. Treat | Lib. Treat | Cons. Treat | Cons. Treat |
| Observations | 1,670 | 1,670 | 1,670 | 1,670 |

This table estimates the effect of the treatment on topics appearing in participants' Facebook feeds. *Michael Cohen* is the number of times the expression "Michael Cohen" appeared. *Clark Shooting* is the number of times the expression "Stephon Clark" appeared. *Louis Farrakhan* is the number of times the expression "Louis Farrakhan" appeared. *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 participants who kept the extension installed for at least two weeks. I include all posts until April 15, 2018, and not only the first two weeks after the extension was installed because different participants installed the extension at different dates, and the purpose of the table is to test whether the participants were exposed to specific events when they were covered in the media. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.22: Estimations Decomposing the Segregation in News Exposure

| | Subscribed OLS | Usage OLS | Exposed IV |
|---|---|---|---|
| | (1) | (2) | (3) |
| Pro-Att. Treatment | 0.501*** | 18.000* | |
| | (0.087) | (10.800) | |
| Subscriptions | | | 0.010*** |
| | | | (0.001) |
| Subscriptions * Pro-Att. | | | 0.005*** |
| | | | (0.002) |
| Unit | Participant | Participant | Participant*Outlet Group |
| Baseline Controls | | X | |
| Mean in Counter-Att. Treatment | 1.537 | 147.181 | 0.008 |
| Observations | 1,058 | 1,058 | 2,116 |

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 the first two columns, the independent variable is whether the participant was assigned to the pro-attitudinal or counter-attitudinal treatment. 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 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. In the first two columns robust standard errors are used. In the third column, standard errors are clustered at the individual 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