Social Media and Polarization: Evidence from a Field Experiment

Ro'ee Levy*

August 26, 2019

Preliminary: Please do not circulate

There is growing concern that the rise of social media is increasing consumption of pro-attitudinal news and leading to greater polarization. I estimate the effect of social media news exposure by conducting a large field experiment randomly offering subscriptions to conservative and liberal outlets on Facebook. The intervention is designed to have high external validity: the news participants are exposed to is determined by Facebook's algorithm, the news supplied to participants is the actual news provided by leading media outlets and participants decided which news to consume. I collect novel data allowing me to analyze the entire chain of media effects: subscriptions to outlets, exposure to news on social media, visits to online news sites and news sharing.

I present three main findings. First, news consumption habits are substantially affected by the social media feed. A change of one standard deviation in the slant of the Facebook feed of compliers leads to a change of 0.31 standard deviations in the slant of news sites they visit, suggesting that individuals do not optimize their news consumption according to their preferred slant, but instead their decisions are often driven by search costs. Second, I find that exposure to cross-attitudinal news decreases affective polarization, defined as negative attitudes toward the opposing party. An increase of one standard deviation in the share of counter-attitudinal news in the social media feed decreases affective polarization by 0.13 standard deviations. Third, while attitudes toward parties change, political opinions are not affected by the slant of news individuals are exposed to. I decompose the mechanisms leading to greater consumption of pro-attitudinal news on social media and find evidence that Facebook is less likely to supply posts from counter-attitudinal outlets, conditional on subscription (a "filter bubble"). Together, these results imply that the algorithms governing social media may be amplifying the tendency to consume pro-attitudinal news and are thus increasing affective polarization.

*Yale University, roee.levy@yale.edu

1 Introduction

The share of Americans consuming news on social media has been steadily increasing. According to Pew surveys, 67% of Americans consume news on social media (Shearer and Matsa, 2018) 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 source of news, there is growing apprehension over its effects on public opinion. A primary concern is that due to the unique features of social media, individuals are exposed to more pro-attitudinal news and as a result polarization increases (Sunstein, 2017). Furthermore, several recent scandals, including the spread of fake news on social media, the Russian-based campaign to influence the 2016 US presidential elections, and Cambridge Analytica's attempt to affect political races, highlight a concern that individuals may be targeted and easily manipulated on social media. Yet, there is limited evidence on the effects of social media news consumption on political beliefs.

In this paper, I ask two questions: first, how does social media affect news consumption habits? Second, how does the consumption of pro-attitudinal and counter-attitudinal news through social media affect political opinions and affective polarization, defined as negative attitudes toward the opposing party? I study these questions by conducting a large online field experiment randomizing exposure to news outlets on social media, and collecting novel data on the news participants are exposed to, the sites they visit and their beliefs.

I first construct a new dataset by merging data on browsing behavior, voting and news outlets to show that Facebook, the most dominant social media platform for news consumption, is associated with consumption of pro-attitudinal news. Among individuals residing in the most liberal and most conservative deciles, 27% and 24% of visits to news sites through Facebook were to very liberal and very conservative outlets, respectively.[1] The corresponding figures for news sites not visited through Facebook are only 18% and 10%.

I distinguish between two main factors contributing to consumption of pro-attitudinal news on social media: news links shared by one's social network, and the wide variety of free outlets individuals can subscribe to when personalizing their feed. Using data from participants randomly assigned to the control group in my experiment, I find that links shared by outlets are driving the increased consumption of pro-attitudinal news. I conduct a field experiment to decompose the channels behind the association between subscriptions to outlets and consumption of pro-attitudinal news, to determine if social media is affecting overall news consumption or merely switching the medium where news is consumed, and to estimate the causal effects of the changes in news consumption on political beliefs.

The experiment has a straightforward design. Individuals were recruited to complete a baseline survey and at the end of the survey they were offered to subscribe to news outlets on Facebook according to their treatment assignment. Participants randomly assigned to the liberal and conservative treatments were offered subscriptions to four liberal or four conservative outlets, and participants in the control group were not offered any outlets. Approximately 52% of participants complied with the treatments and subscribed to at least one outlet. When individuals subscribe to an outlet by "liking" its Facebook

[1] Similarly to Bakshy et al. (2015), I define very liberal outlets as outlets in the most liberal quantile of the outlet slant distribution and very conservative outlets as outlets in the most conservative quantile.

page, posts shared by the outlet may start appearing in their social media feed, with the supply of posts determined by Facebook's algorithm. Participants exposed to the posts could view headlines directly on Facebook and they could click on links in the posts to consume the full news articles in the outlets' websites. Approximately eight weeks after the baseline survey, participants were invited to the endline survey to measure changes in their political beliefs.

Besides political opinions, my main outcome of interest is affective polarization since there is agreement that this measure has been increasing over time. When analyzing the effect on affective polarization, I redefine the treatments as a pro-attitudinal and counter-attitudinal treatment, where 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.

In addition to measuring changes in beliefs based on self-reported survey data, I collect novel data from two sources to analyze the entire chain of media effects. Facebook data on the subscriptions of participants along with data on the posts shared by most participants allow me to accurately measure compliance with the intervention and its effects on political behavior. A browser extension created for the experiment and offered to a subset of participants provides data on the news participants were exposed to on Facebook and the news sites they visited.

I present three main findings. First, participants' news consumption habits are substantially affected by their Facebook feed. As a result of the treatment, individuals consumed more news from the outlets they were randomly offered. The pro-attitudinal and counter-attitudinal treatments increased visits to the websites of the assigned outlets by approximately 25% and 86%, respectively, in the two weeks following the intervention. This implies that even though social media is typically associated with more segregated news consumption, individuals are willing to engage with counter-attitudinal news when it is accessible on social media.

In addition, individuals do not reoptimize their news consumption habits to keep the slant of their overall news consumption constant. The intervention made the slant of news sites visited in the liberal treatment 0.1 standard deviations more liberal, and the slant of news consumed in the conservative treatment 0.08 standard deviations more conservative. The difference between the treatment-on-the-treated (TOT) effects of the liberal and conservative treatments is similar to the difference between the mean slant of online news consumption in New York and Alabama. By instrumenting the change in the slant of participants' Facebook feed with the treatment, I find that as the slant of the feed became one standard deviation more conservative, the slant of news sites visited become 0.31 standard deviations more conservative. The news consumption induced by the treatments does not crowd in or crowd out additional content, as almost all of the effect on the slant of news sites visited is driven by the outlets offered to the participants, and not by spillovers to other outlets. The effect on news consumption gradually declines over time but does not disappear. In the eighth week following the intervention, the effect of the treatment arms on the slant of news sites visited was 47% of the effect in the first immediate week following the intervention.

The finding that social media has a strong effect on news consumption suggests that search costs play an important role when individuals decide which news to consume, as individuals shift their consumption

habits when articles from specific outlets become more easily accessible. Consequently, the algorithms determining which articles appear in one's social media feed can drastically alter one's 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, composed of five questions in the endline survey, and find that the counter-attitudinal treatment decreases the index by 0.03 standard deviations compared to the pro-attitudinal treatment.[2]

I estimate the magnitude of the effect of exposure to pro and counter-attitudinal news using a congruence scale measuring the degree of pro-attitudinal news. I find that an increase of one standard deviation in the congruence scale of news individuals are exposed to on Facebook decreases affective polarization by approximately 0.11 standard deviations. Similarly, an increase of one standard deviation in the share of counter-attitudinal news among all counter and pro-attitudinal news on Facebook decreases affective polarization by 0.13 standard deviations.

I conduct back-of-the-envelope calculations to measure the effect of changes in social media platforms on affective polarization. I focus on the feeling thermometer questions, measuring the difference between how participants feel toward their own party and the opposing party, to compare the result to existing benchmarks. I find that if individuals were exposed to an equal share of pro and counter-attitudinal news on Facebook, the difference between how they feel toward their own party and the opposing party would decrease by 3.91 degrees on a 0-100 scale. Alternatively, if Facebook would become only slightly more balanced, such that news consumed through Facebook would have the same score on the congruence scale as news consumed through other means, the difference in the feeling thermometer would have decreased by 1.09 degrees. For comparison, the difference in the feeling of Americans toward their own party and the opposing party increased by 3.83 degrees between 1996 and 2016.

My third finding is that while participants' attitudes toward parties are affected by the news they consume, exposure to conservative and liberal content does not have a strong effect on political opinions. The effect of the treatments on a political opinion index, focusing on issues discussed in the news during the study period, is economically small, precisely estimated and is not statistically significant.

The results support a model of sophisticated consumers who take into account the slant of the news they consume when updating beliefs shaping political opinions. Thus, consumers are not affected by the outlets they are randomly exposed to. However, since attitudes toward parties are affected by the treatment, the results suggest that these attitudes are not simply a function of the distance in political opinions. I suggest an alternative framework, where consumers have heterogeneous weights on the beliefs that combine to form political opinions. Attitudes toward a party depend on the distance between the party's political opinions and the opinions the party would have formed if it held the consumer's beliefs. 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, the participants may have learned the logic behind some of the arguments made by a party, even if they continued to disagree with the importance of these arguments.

[2] All estimates in the paper are intention-to-treat estimates (ITT) unless noted otherwise.

In the final section, I analyze the mechanisms leading to greater exposure to pro-attitudinal news on social media. The effect of the pro-attitudinal treatment on exposure to pro-attitudinal posts on Facebook is approximately twice as large as the effect of the counter-attitudinal treatment on exposure to counter-attitudinal posts. I decompose this gap in exposure into three main explanations: (1) Participants are more likely to subscribe to an offered outlet in the pro-attitudinal treatment; (2) Facebook is less likely to supply posts from counter-attitudinal outlets, conditional on subscription; (3) Participants use Facebook less frequently 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. The decomposition implies that the personalization of content, unique to online news consumption, and especially social media, can substantially affect the news people are exposed to.

Combining all the results paints a complicated picture: on the one hand, social media's algorithms limit exposure to counter attitudinal news which, in turn, affects the news sites individuals visit, and as a result, affective polarization increases. However, the study also shows that individuals are willing to engage with counter-attitudinal news and a subtle nudge can substantially increase consumption of cross-attitudinal news and decrease polarization.

This paper contributes to the literature on online news segregation by showing that the algorithms governing social media are increasing segregation. While there has been concern over the effect of social media's algorithms (Tufekci, 2015), research on the topic is limited. Gentzkow and Shapiro (2011) argue that news consumption online is not more segregated than offline news. Their study was published when social media was still in its infancy. Papers published since have both suggested that social media may be playing a role in increasing online segregation and argued that social media does not increase segregation and may even have a moderating influence (Flaxman and Rao, 2016; Guess, 2018; Peterson et al., 2018). Recent reviews of the literature have concluded that "*We lack convincing evidence of algorithmic filter bubble in politics*" (Guess et al., 2018)[3]. Most of the papers in this literature are based on browser data on the sites individuals visit, but they suffer from a lack of data on news individuals are exposed to on social media.[4] As a result, these papers cannot test for segregation *within* one's social media feed, where news is often consumed. Furthermore, these papers are not based on random variation and thus focus on correlations between social media and news consumption. Using unique data on social media feeds and experimental variation in subscriptions to outlets on social media, this paper decomposes the mechanisms which limit exposure to cross-attitudinal news and demonstrates that a filter bubble does exist, i.e., that conditional on subscription, individuals are more likely to be exposed to news matching their ideology on social media.

This paper contributes to the literature on the internet, pro-attitudinal news consumption and polarization by testing the effect of varying the main mechanism through which social media is suspected to increase polarization: decreasing 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 of social media on

[3]See also (Zuiderveen Borgesius et al., 2016)

[4]One notable exception is a paper by Bakshy et al. (2015) which studies the mechanism behind online segregation. It analyzes Facebook's internal data and argues that exposure to cross-attitudinal news is mostly limited by individual choices and not by algorithmic ranking. Bakshy et al. (2015) focus on individuals' social networks, while I focus on outlets to which individuals subscribe.

polarization and have come to mixed conclusions (Allcott et al., 2019; Boxell et al., 2017; Lelkes et al., 2015). Since these papers focus on social media generally, they do not identify the causal effect of pro-attitudinal news on polarization. Indeed, a recent review of the literature argued that "*it is far from clear, however, that partisan news actually causes affective polarization*" (Iyengar et al., 2019).

This study also contributes to a well-established literature on media persuasion. This is the first study to randomize exposure to news outlets on social media. Lab experiments have found that individuals are persuaded by the news they consume (e.g. Coppock et al. 2018) and several papers with quasi-experimental designs show that viewers are persuaded by Fox News (DellaVigna and Kaplan, 2007; Martin and Yurukoglu, 2017). However, in other settings, the persuasive effects of the media on political opinions are not as clear. Gentzkow (2006) measures the effect of the entry and exit of newspapers and does not find an effect on party vote shares, and Allcott and Gentzkow (2017) find that fake news probably had a negligible effect on the 2016 election. In a survey of the literature, Strömberg (2015) argues that it is still not clear if, and to what extent, consumers' political behavior is affected by the news they consume. In many contexts, the "gold standard" for measuring causal effects is field experiments, since they combine the strong identification offered by lab experiments with higher external validity. However, with the notable exception of Gerber et al. (2009), there have been almost no field experiments exposing participants to different media outlets.[5]

Methodologically, this paper contributes to a growing literature conducting online media-related experiments (Allcott et al., 2019; Bail et al., 2018; Chen and Yang, 2018; Jo, 2018), by demonstrating how an experiment can exploit social media's existing infrastructure to gradually distribute news to participants in a natural setting. The main contribution in the design of this experiment is its high external validity, as besides the initial offer to subscribe to outlets, which is common on social media, it does not intervene in any behavior. Facebook's algorithm determined which posts participants were exposed to. The news supplied to participants was the actual news provided by leading media outlets during the study period. Finally, participants decided whether to read, skip, comment on or share specific articles that appeared in their feed. In other words, the treatment participants experienced due to the intervention is almost identical to the experience of tens of millions of American who subscribe to news outlets on Facebook.[6]

The remainder of the paper is organized as follows. In the next section, I provide background on Facebook and present exploratory analyses showing that social media is associated with consumption of pro-attitudinal news. Section 3 describes the experimental design, the datasets used and the empirical strategy. The fourth and fifth sections present results: Section 4 analyzes the effect of the experiment on news exposure, consumption and news sharing behavior, and Section 5 analyzes the effect on political opinions and affective polarization. Section 6 places the results in a theoretical framework on the effect of news consumption on political beliefs. Section 7 decomposes the mechanisms increasing consumption of and exposure to pro-attitudinal news on social media. The final section concludes.

[5] 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 media organizations but also from elected officials, opinion leaders and non-profit groups.

[6] The only differences are that the participants in the experiment also completed a baseline survey and that they received the suggestion to subscribe to an outlet within the survey and not within Facebook.

2 Social Media News Consumption

2.1 Background: Facebook

This study focuses on Facebook since it is the most dominant social network, used by 68% of Americans. Despite its prominence, Facebook has been understudied, especially compared to Twitter, since 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 the company's algorithm. Posts can be shared by either friends or pages of organizations, such as media outlets that users subscribed to, and may include text, video, pictures, and links.[7]

Facebook is a very popular source for news consumption. Approximately 43% of Americans get news on Facebook, more than twice the share getting news through other social networks (Shearer and Matsa, 2018). While this study focuses on US news consumers, understanding the effect of Facebook has global implications due to Facebook's popularity worldwide. 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).

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

2.2 Exploratory Analysis: Social Media is Associated with Pro-Attitudinal News

There is a gap between the public concern over the effect of social media on news consumption, expressed often in press reports discussing echo chambers and filter bubbles, and rigorous studies on the topic which have not found conclusive evidence that social media is segregating news consumption. To motivate the experiment, this section explores the link between Facebook and news consumption. I present two stylized facts: I show that Facebook is associated with greater consumption of pro-attitudinal news and that Facebook is associated with consumption of more extreme news.

To estimate the association between Facebook usage and news consumption, I rely on three datasets. First, the 2017 Comscore Web Behavior Database Panel provides a sample of the browsing behavior of approximately 93,000 US internet users, where each observation is a domain visited by a specific computer, along with the referring domain. It is a subset of Comscore's opt-in Media Matrix Panel (Comscore, 2017). Second, to determine the slant of each website, I rely on a dataset of news domains constructed by Bakshy et al. (2015). The dataset defines the slant of 500 leading news sites according to the self-reported ideology of Facebook users sharing articles from these websites. The dataset correlates

[7] To subscribe to content from a page an individual "likes" the page on Facebook. To simplify terminology, throughout the paper I will describe this action as subscribing to an outlet, instead of liking a page.

well with other datasets (e.g. Gentzkow and Shapiro, 2010) measuring the slant of outlets and is used here since it provides the slant to a large number of outlets, at the website level. Throughout the paper, I refer to outlets on this list as *leading news outlets*. For more details on processing these outlets see Appendix A.1. Third, as a proxy for individuals' ideology, I use 2008 zip code level voting data (Mummolo and Nall, 2016). 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. The results in this section are robust to using 2016 county-level election data and 2017-2018 donation data.

The data confirms that Facebook is an important source of news consumption and only Google refers more visits to news sites. 7.4% of visits to leading news sites in the sample are referred to by Facebook, and the share increases to 15.6% among individuals who visited at least one site through Facebook.[8]

Figure 1 presents a binned scatter plot of news consumption by ideology and shows a clear correlation between the consumer's 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 for news consumed through other means (the dashed black line), indicating that news consumed through Facebook tends to better match one's ideology. To construct this figure, I calculate the mean slant of news sites visited for each individual in the sample when the sites 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.

Table 1 tests the robustness of this result 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 0.05 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 the effect is robust to adding individual and month fixed effects, indicating that the positive relationship between social media and pro-attitudinal news consumption does not stem only from different individuals selecting to consume news through social media. A potential issue with these regressions is that they do not take into account spillovers. Individuals could merely be switching the medium in which they consume specific domains. For example, if a conservative clicks on links to Fox News articles through Facebook, she may, as a result, consume less news by directly entering the Fox News home page. Columns (3)-(4) deal with 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 consuming a greater share of news through Facebook tend to consume news better matching their ideology. Column (4) shows that the result is robust to controlling for individual and month fixed effects, and column (5) finds a similar effect when using general Facebook usage as the independent variable.

While the mean slant of news consumed through Facebook is not extreme even at the most liberal and most conservative zip codes,[9] the share of pro-attitudinal news increases substantially when news is

[8]Reports using different datasets confirm that Facebook is an important source of news consumption. For example, according to Parse.Ly, which collects data from a large network on online publishers, Facebook is the second most important referral source for publishers and accounts for 26% of external traffic to news sites. Parse.Ly - 2018 Traffic Sources by Content Categories and Topics. For updated data on external referrals in the Parse.Ly's network see: https://www.parse.ly/resources/data-studies/referrer-dashboard/

[9]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.

consumed through Facebook. On average, in zip codes in the most liberal decide, 27% percent of news consumed through Facebook is from very liberal outlets, compared to 18% of all other news consumed. In an average household in the most conservative zip codes, 24% of news consumed through Facebook is from the most conservative outlets, compared to 10% of other news consumed. Similarly to Bakshy et al. (2015), throughout the paper I define very liberal outlets as outlets in the bottom quantile of the slant distribution, liberal outlets are defined as outlets in the second quantile, conservative outlets as outlets in the fourth quantile and very conservative outlets as outlets in the top quintile of the news slant distribution.

One implication of individuals consuming pro-attitudinal news through Facebook is that less moderate news is consumed. Indeed, Figure 2 presents the density of the mean slant of news consumption at the individual level and shows that the mean slant tends to be more extreme when news sites are visited through Facebook. For example, when accessing news sites through Facebook, more individuals have a news diet with a mean slant similar to Fox News and The Huffington Post, and fewer individuals have a news diet resembling CNN and USA Today.

Table 2 tests this result in a regression framework, similar to the regression presented in Table 1. The dependent variable is defined as the absolute value of the slant, and the independent variable is whether, or how often, news was consumed through Facebook. My preferred specification is column (4) which includes data at the individual*month level and controls for individual and month fixed effect. The column shows that a month when an individual consumed all news through Facebook is associated with an increase of 0.5 standard deviations in the slant's absolute value, compared to a month when news was not consumed through Facebook. The rest of the specifications are consistent with this result.

The evidence in this section showing that social media is associated with pro-attitudinal news helps explain the public concern over social media news consumption. While Comscore data provides a large and diverse panel which is useful in understanding the news consumption habits of American consumers, these regressions do not provide clean identification, nor can they shed light on the mechanisms leading to these results or their implications. Therefore, an experiment generating random variation in exposure to news on social media is required.

3 Design and Data

3.1 Experimental Design

The goal of this study is to measure the effect of exposure to news on social media on online news consumption and political beliefs.

I recruited participants to the study in February-March 2018 using Facebook ads. Individuals who clicked the ads were directed to the survey landing page, where they could begin the baseline survey by logging in using their Facebook account. Toward the end of the survey, participants were randomly assigned to a liberal treatment, conservative treatment or control group, with the randomization blocked

by participants' self-reported baseline ideology.[10] Participants in the conservative treatment were offered to subscribe to four potential conservative outlets, participants in the liberal treatment were offered to subscribe to four potential liberal outlets and participants in the control group were not offered any outlets. The four potential liberal outlets and four potential conservative outlets were defined for each participant before the treatment. The potential outlets did not include outlets which participants already subscribe to on Facebook to ensure participants would only be offered to subscribe to *new* outlets. To clarify, the intervention did not provide exclusive access to these outlets, and any individual can subscribe to these outlets on Facebook at no cost or effort, regardless of the intervention. Therefore, the intervention should be thought of as a nudge encouraging exposure to new sources of news. In April-May 2018, approximately eight weeks after the baseline survey, participants were invited to the endline survey.

The experiment is intentionally designed to be as natural as possible. The only intervention is the initial nudge asking participants to subscribe to an outlet. Similar interventions occur naturally when individuals encounter suggestions for pages they can subscribe to, either placed by ads or by Facebook. When participants subscribed to an outlet, posts from the outlet appeared in their Facebook feeds according to Facebook's algorithm, just as they would if anyone else would have subscribed to the outlet. Since leading news outlets were chosen, the posts observed by the participants were also observed by millions of other Americans during the study period. Finally, at no point were participants asked to engage with any posts or read any news content. Participants 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 decision they make regarding other posts appearing in their feed.

3.2 The Setting: Media Outlets and the News Environment

Figure 3 presents the primary outlets offered in the experiment. These outlets were offered to participants in the liberal and conservative treatments, who did not already subscribe to them. The primary liberal outlets are MSNBC, Slate, Huffington Post and The New York Times, and the primary conservative outlets are The Wall Street Journal, The Washington Times, Fox News and The National Review.

The news outlets were chosen according to several criteria. First of all, they have a relatively clear ideological slant. Secondly, preference was given to popular outlets (Fox News and the New York Times are the second and third most popular news pages on Facebook) so the effects found would have high external validity. Finally, outlets of varying quality and extremity are included to allow participants several different options when choosing whether to subscribe to an outlet and thus increase the likelihood that participants engage with at least one of the outlets offered.

[10] Randomization was blocked to increase power and 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 for "I haven't thought about it much." Participants were assigned to a treatment 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 options in the ideological scale. The first participant in each stratum 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.

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. The alternative outlets are presented in Appendix Figure A.2, and Appendix Table A.1 displays the full list of outlets, along with the number of times they were offered and number of new subscriptions among participants who completed the endline survey.[11]

Figure 4 displays the most prominent men and women mentioned in posts shared by the primary outlets during the study period and shows that the outlets mostly discussed political figures. Unsurprisingly, President Trump is the dominant figure mentioned and is mentioned more than the next 17 individuals combined. 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 Vladimir Putin, Michael Cohen, Scott Pruitt, and Stormy Daniels much more often than conservative outlets.

3.3 Data Collection

The analysis of the experiment relies on three datasets: self-reported survey data, Facebook data on posts individuals shared and pages they subscribed to, and browser data on the Facebook feed and news related browsing behavior. To the best of my knowledge, this is the first study combining experimental variation with social media and browsing data.

Survey data is used to measure baseline and endline self-reported political beliefs and news consumption habits. 37,492 participants validly completed the baseline survey and 19,693 participants validly completed the endline survey.[12]

After reviewing the consent form, participants logged in to the survey using their Facebook account, through a Facebook app created for the project. They were asked to provide permissions to outlets they subscribed to and posts they shared. Providing permissions was completely voluntarily and permissions could be revoked at any time and were revoked automatically approximately two months of the logged in to the baseline or endline survey. Since data on baseline subscriptions was required to define new potential outlets for each participant, only participants who provided permissions to access their subscriptions are included in the baseline sample.[13] Approximately 92% of baseline participants provided

[11] To reduce the number of final outlets, alternative outlets which were defined as potential outlets for fewer than 20 participants were excluded from the experiment, along with the participants for which these outlets were defined. This removed less than 0.1% of participants from my sample.

[12] Approximately 3.64% of participants who completed the baseline survey validly were excluded from the study since they already subscribed to too many outlets and there were no available four new liberal outlets or four new conservative outlets which could potentially be offered to them. Participants were also excluded if the initial pages they subscribed to could not be observed, if they responded carelessly, if they took the survey a second time or if a technical error prevented some of their data from being collected. See additional details in Appendix A.2.

[13] 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.

access to the posts they shared for at least two weeks.[14]

To collect browser data, participants who completed the baseline survey using Google Chrome on a desktop computer were offered to install a browser extension collecting data on the Facebook feed and news-related browsing behavior, in exchange for a small reward. 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 was designed to only collect the URLs of news sites visited. Approximately 2,447 of the 8,082 participants who were offered the extension, installed it. Most of the analysis in this paper will focus on data from 1,839 participants who kept the extension installed for at least two weeks.[15]

Table 3 summarizes the main outcomes and datasets used. Most of the analysis will focus only on the relevant sub-sample participants, e.g. when analyzing media outcomes I focus on the sub-sample of participants who installed the browser extension. For more details on the surveys, Facebook data and extension data see Appendix Sections A.2, A.3, and A.4, respectively.

3.4 Primary Outcomes

3.4.1 Media

I measure media outcomes along four main dimensions: subscriptions to outlets, posts observed on Facebook, news sites visited and posts shared.

Subscriptions are collected using Facebook data. Any participant who subscribed to at least one of the outlets offered is considered a complier. Since the intervention only offers new outlets to participants, defiers do not exist in this experiment. While always-takers are theoretically possible, they are extremely rare. To simplify the analysis, I will define compliance as subscribing to an offered outlet when it was offered.[16]

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 collected using Facebook data, suggesting data was collected properly and participants answered questions truthfully in most cases.

Facebook news exposure is measured based on the browser extension. I collect data on posts appearing on the participants' Facebook feed, and attribute a post to a news outlet if it is shared by the outlet or contains links to the outlet's domains.[17] While variation generation by the experiment is in subscriptions

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

[15] 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 for several weeks. 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 an early preview of the study results.

[16] Defying the experiment would mean unsubscribing from an offered outlet, but in the baseline sample 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 0.6% of participants subscribed to their potential liberal outlets, and only 0.4% subscribe to their potential conservative outlets in the two weeks following the intervention.

[17] In order to match URLs with news outlets, I first convert over 10 millions URLs to their final endpoint, allowing redirects along the way. For more details see Appendix A.4

to the outlets' Facebook pages, I include in this measure news articles shared by Facebook friends as well, to accurately measure total exposure to news outlets on Facebook.

Browsing behavior is also measured based on the browser extension and tests whether offering subscriptions to new outlets affects media consumption. The extension can greatly reduce measurement error as other studies have suggested that individuals self-reported media habits are more polarized than their actual media habits (Guess et al., 2017).

Posts shared by the participants are used to analyze the effect of the treatments on political behavior. 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. I focus on posts shared in the two weeks following the intervention and exclude posts sharing photos, albums, music, and events.[18]

For each dimension, the following media outcomes are constructed and analyzed. First, to test the direct effect of the experiment, I measure the number of times participants engaged with the *potential outlets*. Second, I measure the mean slant of all *leading news outlets* participants engaged with, where the slant of each outlet is based on Bakshy et al. (2015). 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 consumption, 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.

3.4.2 Opinions and Attitudes

In this paper, I analyze the effect of news exposure on two primary outcomes: affective polarization and political opinions. The construction of both primary outcomes is defined in the study's pre-analysis plan along with the control covariates used in the primary regressions.[19] The pre-analysis plan is discussed in more detail in Appendix C.

Affective polarization is measured as an index composed of five outcomes. First, I use standard feeling thermometer question and calculate the difference in feeling towards the participant's party and the other party (*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", and I calculate the difference in the answers (*difficult perspective).* Third, participants are asked

[18]The remaining posts typically include a link or an embedded video. I focus on these posts since they are more likely to contain political content relevant to the experiment. However, the outlets offered to participants may also publish posts that contain only a photo and text (for example Fox News published posts with quotes related to the news without an accompanying link or video). 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

[19]AEA RCT Registry - Trial 0002713.

a similar question on the following statement: "I think it is important to consider the perspective of Democrats/Republicans" (*consider perspective*). The *difficult perspective* and *consider perspective* questions 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, some, a few or almost no good ideas (*party ideas*). I take the difference between the answer's relative ranking among the possible options (i.e., no good ideas is coded as 3 and a lot of good ideas is coded as 0). Fifth, to measure social-distance participants are asked how they would feel if they had a son or daughter who married a Democrat/Republican (*marry opposing party*). This question is asked among participants who identify with a party, and asks how they would feel if their son or daughter married someone who identifies with the opposing party. The answer is coded as 2 for very upset, 1 for somewhat upset and 0 for not upset at all. Each outcome variable is defined such that a higher value is associated with more polarization and then standardized by subtracting the control group mean and dividing by the control group's standard deviation.

I focus on affective polarization since scholars agree that this measure has been increasing over time (Gentzkow, 2016; Lelkes, 2016) and there is growing concern over its implications on political accountability, governance, social selection, professional behavior, and even product and labor markets (Iyengar et al., 2019; Iyengar and Krupenkin, 2018; Abramowitz and Webster, 2016). The increase in affective polarization has not escaped the public and in a recent survey 85% of Americans stated that the tone and nature of political debate in this country have become more negative over the past several years, compared to only 3% who said that the tone has become more positive (PEW, 2019).

Political opinions are measured based on an index composed of twenty endline survey questions. The questions focus on domestic political issues and figures discussed 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.[20] Each outcome variable is defined such that a higher value is associated with a more conservative opinion, and then standardized.

For both the affective polarization and the political opinion outcomes, the final index is composed by taking an average of all the index components and the index is then standardized with respect to the control group so all effects are measured in standard deviations.

3.5 Balance and Attrition

Table 4 presents descriptive statistics for participants who completed the endline survey, broken by treatment group. Similarly to other opt-in panels (Yeager et al., 2011), the sample is not nationally representative. Participants tend to be more liberal than the general population, and as expected, more participants say that they get most of their news on social media (20%), compared to the national population (13%).[21]

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

[21]There are several likely explanations for why the sample is different than 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 than samples of Mechanical Turk users (Berinsky et al., 2012), for example. One advantage of this sample is that Facebook users are not experienced, semi-professional survey takers, in contrast to many Mechanical Turk workers. Participants were asked in the followup survey how many additional surveys they completed in the past month, the median answer is 1 and the mean answer

In contrast to the ideological composition, the gender composition of participants and their average age is very similar to the US population. Since some of the analysis compares individuals by whether they were exposed to a pro-attitudinal or counter-attitudinal treatment, Table 5 presents a balance table according to whether the treatment matched the participant's ideology, and shows that the sample is balanced also along the redefined treatment arms.

Tables 4 and 5 also test for differential attrition among the three endline samples: participants who completed the endline survey, participants who provided access to posts they shared for at least two weeks and participants who provided access to browsing data for at least two weeks. While there is no concern of differential attrition in the latter two samples, there is slightly greater retention of participants who completed the endline survey in the control group (48%), compared to the liberal (45%) and conservative groups (45%).[22] The differential attrition mostly stems from the fact that some participants in the conservative and liberal treatments did not complete the final screen of the baseline survey after they encountered the intervention, either due to a technical issue that affected a small share of participants or since they preferred not to complete the survey at that stage.[23]

Appendix Table A.2 includes only participants who completed the endline survey and shows that despite the differential attrition, there are no substantial differences in observables between the treatment arms. Similarly, Appendix Table A.3 shows there are no substantial differences in observables according to whether participants were assigned to the pro-attitudinal treatment or counter-attitudinal treatment. In both cases, based on an F-test, I reject the hypothesis that the treatment arms are different from each other based on observables.

Most importantly, there is no differential attrition in completing the endline survey between the conservative treatment and liberal treatment and no differential attrition between the pro-attitudinal treatment and counter-attitudinal treatment. The primary endline survey outcomes are analyzed by comparing the two treatment arms to each other and thus do not suffer from differential attrition.

3.6 Empirical Strategy

To estimate the effect of exposure to news on political beliefs, two primary models are used throughout the analysis: a model comparing the conservative and liberal treatments, and a model redefining the treatments as counter-attitudinal and pro-attitudinal. Below, I describe the main intention-to-treat (ITT) estimates. To estimate the effect of subscribing to outlets, I also analyze TOT estimators, where the independent variable is compliance with the treatments and the instrument is the random treatment assignment.

is 7. For comparison, a 2014 study found that the median Mechanical Turk reported participating in 20 academic studies in the *week* before the question was asked (Rand et al., 2014).

[22] Table 4 also shows that there is some difference between the conservative treatment and the other groups in 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*.

[23] Participants who did not complete the survey did not provide their email address in most cases and therefore it was more challenging to recruit them to the endline survey.

3.6.1 Liberal and Conservative Treatments

The following intention-to-treat regression is used to measure the average effect of the intervention on political opinions:

$$Y_i = \beta_1 T_i^L + \beta_2 T_i^C + \alpha X_i + \varepsilon_i \quad (1)$$

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. The dependent variables in this specification are engagement of participant i with potential outlets, the slant of all leading outlets participant i engaged with and i's political opinions, where a higher Y_i is associated with more conservative outcomes.

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). Comparing these treatments, 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 compliers to more conservative content, compared to the control group). In addition, there is no evidence for differential attrition between the liberal and conservative treatments, so comparing these treatments mitigates any concern that differential attrition may be driving the results. Finally, if individuals are persuaded by both treatments, comparing them directly takes advantage of the experiment's design and provides more power than comparing each treatment to the control group.

To increase power, when estimating the effect on political opinions and affective polarization, I control for a set of pre-registered covariates, X. I control for self-reported ideology, party affiliation, approval of President Trump, ideological leaning, age, age squared, gender and baseline questions similar to the questions used in the endline survey. For the full list of control variable and their definitions, see Appendix B.1. When estimating the effect on media outcomes I only control for baseline outcomes, when they exist.

3.6.2 Pro-Attitudinal and Counter-Attitudinal Treatments

When measuring the effect on polarization, the independent variable expected to have an impact is not whether news outlets were conservative or liberal, but whether they matched the participant's ideology. Therefore, the following model is analyzed:

$$Y_i = \beta_1 T_i^P + \beta_2 T_i^A + \alpha X_i + \varepsilon_i \quad (2)$$

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

with a conservative ideology or a conservative treatment assigned to a participant with a liberal ideology.[24] The dependent variables in this specification are the engagement of participant i with the potential pro-attitudinal or counter-attitudinal outlets, the congruence scale and share of counter-attitudinal news among all outlets the participant engaged with and the participant's affective polarization. $\beta_2^A < \beta_2^P$ tests whether individuals become more polarized when assigned to pro-attitudinal news, compared to counter-attitudinal news.

4 Findings: Demand for News on Social Media

In this section, I show that participants' news consumption habits are substantially affected by their Facebook feed. I first show that when given an option, participants were willing to engage with counter-attitudinal news. Participants in the counter-attitudinal treatment subscribed to the counter-attitudinal outlets, visited the outlets' websites and even shared their posts. In the second sub-section, I focus on news consumed from all leading outlets and show that the slant of news sites visited is strongly affected by the treatment.

4.1 Individuals Willing to Engage With Counter-Attitudinal News

Overall, 56% of participants who were offered pro-attitudinal outlets complied with the treatment and subscribed to at least one outlet, compared to 46% of participants who were offered counter-attitudinal outlets. Appendix Table A.4 presents descriptive statistic on the compliers by treatment. Liberals, women, and participants who subscribe to more outlets on Facebook were generally more likely to comply with the treatments.

The difference between the share of participants subscribing to counter-attitudinal and pro-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% comply with the counter-attitudinal treatment, compared to 59% who comply with the pro-attitudinal treatment. While unsubscribing from the outlets offered was more common in the counter-attitudinal treatment, in only 12% of cases compliers in the counter-attitudinal treatment unsubscribed from the offered outlets within two weeks. Panel 1 of Figure 5 shows that two weeks following the intervention, participants assigned to the counter-attitudinal

[24]The ideological leaning of participants is determined according to the party they identify with, or party they lean toward. For participants who do not identify with the Republican or Democratic party, the ideological leaning is determined according to where participants place themselves on the ideological scale. For moderate participants or participants who do not know where to place themselves ideologically, the ideological leaning is determined according to the candidate they supported in the 2016 elections. Approximately 3% of participants did not self-identify as liberal or conservative, did not identify with the Republican or Democratic party and did not vote for Trump or Clinton, and are excluded from the analysis when analyzing the effect of the pro and counter-attitudinal treatments.

All the results are robust to defining ideological leaning first by self-reported ideology, then by party and then by candidate, which how I ideological leaning is defined in the pre-analysis plan. I use the first definition to make the study comparable to other paper which tend to focus on party affiliation (Druckman and Levendusky, 2019) and since the affective polarization question asked about Republicans and Democrats and not liberals and conservatives.

treatment still subscribed to 1.4 new counter-attitudinal outlets on average. Each row in the figure is estimated by regressing engagement with the four potential pro-attitudinal outlets or four potential counter-attitudinal outlets on the treatment. Panel 2 of Figure 5 also shows that following the intervention, participants in the pro-attitudinal and counter-attitudinal treatments were exposed to more posts from the pro-attitudinal and counter-attitudinal outlets, respectively.

Participants could consume news from the outlets they subscribed to by simply reading the posts in their Facebook feed or by clicking links included in the posts to visit the outlets' websites directly.[25] Panel 3 of Figure 5 shows that the counter-attitudinal treatment increased visits to the counter-attitudinal outlets by approximately 82%, an ITT effect of 1.3 additional visits over a baseline of 1.6. Less surprisingly, the figure also shows that the pro-attitudinal treatment increased the number of visits to pro-attitudinal outlets by 23%, an ITT effect of 3 additional visits over a baseline of 13.[26]

Figure 6 shows that participants not only consumed news from counter-attitudinal outlets, they also shared them with their social network. The fact that participants chose to share these posts suggests that participants noticed the posts and considered them important. Sharing the posts also implies that participants indirectly amplified the treatment effect, by expanding some of the treatment to their social network. One possibility is that participants shared posts while commenting negatively on their content. Panel 2 of Figure 6 focuses on posts which were shared with no commentary by the participants and shows that even among these posts, the counter-attitudinal treatment has a positive significant effect on the number of posts shared.

To conclude, merely offering individuals an option to subscribe to counter-attitudinal outlets increases engagement with these outlets. Even though individuals in the experiment were not required to subscribe to any outlet, and almost of half of the individuals who were offered outlets did not subscribe to any of them, some individuals may have subscribed due to Hawthorne effects. Fortunately, the offer to subscribe to an outlet is the only stage in which Hawthorne effects can plausibly play a role. Since participants are not nudged to consume or share any content and since these activities take place separately from the survey, it is likely that their browsing and sharing behavior is not affected by the experimenter and represents an actual interest in these outlets.

4.2 The Social Media Feed has a Strong Effect on News Consumption

The previous section demonstrated that individuals engage with the potential outlets when they appear in their feed, implying that search costs play a role in deciding what to consume. 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 offered the New York Times may start consuming more

[25] Approximately 81% of posts from outlets participants subscribed to contained links.

[26] The extension data was only collected when participants used Facebook or browsed news sites on a desktop computer, when 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 times individuals were exposed to outlets in their feed or visited news sites are lower bounds for the actual intention to treat effect. In the baseline survey, participants were asked how many links to articles about government and politics they clicked on Facebook in the past 24 hours using a computer and on a mobile phone. Among participants who installed the extension, 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.

articles from the outlet's website, but as a response choose to consume less news from the Washington Post's website, which offers a similar perspective.

Row 1 in Figure 7 shows that when participants were randomly offered liberal or conservative outlets, their feed became more liberal or conservative, respectively. The effect on the news feed is dramatic. A combination of a conservative treatment assigned to a liberal complier and a liberal treatment assigned to a conservative complier closes approximately half 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, the change in feed provides an opportunity to test whether participant's change the slant of news sites visited as a result or whether they optimize the websites they visited according to their preferred slant.

I find that individuals do *not* reoptimize the slant of their news consumption. Row 2 of Figure 7 shows that the treatment has a strong and significant effect on the slant of news sites visited by the participants. The difference between the TOT estimator of the liberal and conservative treatments is about 23% 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 consumption of an individual in Pennsylvania, a swing state, to the diet similar to an individual in New York, a blue state, while the TOT effect of the conservative treatment would have shifted her online news consumption to a news diet similar to an individual in Alabama, a red state.[27]

Table 6 shows that when the compliers' news feed become one standard deviation more conservative, the slant of the sites they visit become 0.31 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 sites visited through Facebook (instead of all news sites visited in the two weeks following the intervention) and find that when the feed becomes one standard deviation more conservative, the slant of sites visited through Facebook becomes 0.72 standard deviations more conservative.

To minimize attrition, the analysis so far has focused on the two weeks following the intervention. Figure 8 shows that the change in media behavior was persistent over eight weeks. Panel 2 displays raw data for the mean slant of all news consumed among participants in the liberal and conservative treatments, along with the control group. All the numbers are negative since liberals are over-represented in my sample. The panel shows that before the intervention there was almost no difference between the news consumption in the two treatment arms and that the intervention induces an immediate gap in the news consumption of participants assigned to the conservative and liberal treatments. The effect of the treatment gradually declines but does not disappear. Eight weeks following the intervention, the effect of the treatments on the slant of news consumed online declined by 53% compared to the immediate effect in the first week.

[27] The slant of news consumption at the state level is calculated using Comscore data. For each individual, 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 intentionally focuses on states where there is a larger sample of at least 750 Comscore panelists who visited news sites more than once.

I estimate that subscription to the new media outlets did not have a strong crowd-in or crowd-out effect on consumption of other news. To test for spillovers across news outlets, I recalculate the effect of the treatments on the slant of news consumption and exposure excluding the eight potential outlets defined for each individual. Appendix Figure A.4 shows that the mean slant of news consumption is not substantially affected by the treatments when the potential outlets are excluded.

So far I have focused on the effect of the Facebook feed on news sites visited. Rows 3 and 4 of Figure 7 show that even the mean slant of posts shared by the participants was affected by the change in participants' social media feed.

This section raises concerns regarding the power of social media companies in shaping news consumption habits. The extremely low costs of subscribing to an outlet along with the fact that individuals do not re-optimize their 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 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 purchasing ads to promote their organizations on social media or even foreign agents promoting specific pages on Facebook in attempt to influence the American electorate.[28]

5 Findings: Opinions and Attitudes

So far it has been established that the intervention had a large effect on participants' media behavior. This section tests whether the new media environment affected their political beliefs. I focus on the effect on political opinions and affective polarization. The effect on political knowledge is discussed in Appendix D.2.

5.1 No Evidence for an Effect of Social Media News Exposure on Political Opinions

Panel 1 of Figure 9 shows the treatment did not affect the political opinions index. The conservative treatment increased the political opinions index by 0.004 standard deviations. While the point estimate has the expected sign, the effect of is very small economically and is not statistically significant. 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.

One possibility is that no effect is found because the first stage was weak, i.e. the intervention did not change the news participants were exposed to. However, this explanation is unlikely since the intervention was strong enough to affect attitudes (as shown in the next section). Furthermore, instead of estimating the intervention directly, I can use the treatment as an instrument for the slant of participants'

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

Facebook feed and then measure how the feed's slant affects opinions. By extrapolating this result, 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 3% in the direction of the opinions of the conservative (see Appendix Table A.6). The effect is not statistically significant and I can reject a shift in opinion of more than 5% at a significance level of 0.05. Finally, perhaps I do not find a strong effect on opinions since participants consume a very small share of their news on Facebook. However, this explanation is unlikely as participants who report getting most of their news on social media were affected similarly to other participants (see Appendix Section D.1).

A second possibility is that the null effect masks important heterogeneity. Perhaps some consumers were persuaded by the outlets they consumed, while in other cases there was a backlash effect and consumers' opinions moved in the opposite direction of the treatment. For example, some conservatives may have become more conservative when exposed to liberal outlets while liberals became more liberal and as a result, the average treatment effect is close to 0. I test this hypothesis by estimating the effect of the interaction of ideology and treatment on the political opinions index and find no evidence for a backlash effect. Appendix Figure A.5 shows that liberals did not become significantly more liberal as a result of the conservative treatment and conservatives did not become more conservative as a result of the liberal treatment.

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. Two major differences between the studies can explain the different results. First, Bail et al. expose individuals to different *views* on Twitter and not only to news outlets. Individuals plausibly became more upset when they are exposed to the twitter feed of opposing elected officials and opinion leaders, compared to counter-attitudinal news outlets. Second, Bail et al. provided participants with financial incentives to continuously follow the new information they were exposed to. While financial incentives provide a stronger treatment effect, they may also encourage individuals to continue consuming news which increases their partisan hostility, and that they would not have consumed without the incentive.

To conclude, the results suggest that individuals' political opinions are not significantly affected by the news outlets they are exposed to.

5.2 Exposure to Counter-Attitudinal News Increases Affective Polarization

In contrast to the null effect on political opinions, Panel 2 of Figure 9 shows that the counter-attitudinal treatment decreased affective polarization by 0.03 standard deviations, compared to the pro-attitudinal treatment. The treatment on treated effect is approximately 0.06 standard deviations. Appendix Table A.5 shows that the result is robust to not controlling for covariates.

Figure 10 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. While I am underpowered to detect a statistically significant effect for all coefficients, in all cases the pro-attitudinal treatment made participants more polarized compared to the counter-attitudinal treatment and the coefficients are similar in

magnitude to the point estimate of the index measure.

Appendix Figure A.6 shows that the effects are relatively homogeneous. Appendix Figure A.7 shows that the treatment arms affected the polarization index in a consistent fashion: the counter-attitudinal treatment decreased polarization and the pro-attitudinal treatment increased it.[29]

In the rest of this section I analyze the magnitudes of the effect using three methods. First, I compare the effect of the intervention to benchmarks in the control group and outside the experiment. Second, I use the extension data to estimate the effect of a change in exposure to pro and counter-attitudinal news on affective polarization. Third, I calculate back-of-the-envelope calculations to estimate how affective polarization would have changed if Facebook had more balanced news exposure.

To compare the results to existing benchmarks, I focus on the feeling thermometer question, which has been 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 is 0.95. For comparison, in the past 20 years, the feeling thermometer measure (the difference between how individuals view their own party and the opposing party) increased by 3.83 degrees. 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.

To estimate the effect of a change in exposure to pro or counter-attitudinal news, I focus on participants who installed the extension and completed the endline survey. I use two summary statistics for exposure to pro and counter-attitudinal news: counter-attitudinal news share and the congruence scale (both statistics are defined in Section 3.4.1). Each statistic is instrumented with the treatment to measure its effect on affective polarization.

I find that an increase of one standard deviation in the share of exposure to counter-attitudinal news decreases affective polarization by 0.13 standard deviations. Similarly, an increase of one standard deviation in the congruence scale decreases affective polarization by 0.11 standard deviations. One challenge in studying affective polarization 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. I divide the effects of news exposure on affective polarization by coefficients obtained using a cross-sectional regression among the control group, and find that approximately 31%-36% of the correlation between pro-attitudinal news and affective polarization is due to a causal effect (see Table 7).

The IV regressions determining the effect of exposure to news rely on the exclusion restriction that the treatment could only affect affective polarization through variation in the Facebook feed. Although this assumption cannot be tested directly, it is unlikely that the treatment affected beliefs through any

[29] I only have enough power to reject a null effect for the effect of the counter-attitudinal treatment. One concern when comparing each treatment arm separately to the control group is that differential attrition may affect the results. Therefore, in the primary specification I compare the treatments to each other.

other channel. The intervention only encouraged participants to subscribe to outlets on Facebook, and subscriptions only change the Facebook feed. However, it is less clear what specific changes in the feed affect beliefs. I focus on two reasonable summary statistics which were strongly affected by the treatment. While it is encouraging that the measures lead to similar results, it is possible that other changes in the feed, not captured in these measures, affected beliefs.

Finally, I use two back-of-the-envelope calculations to estimate how affective polarization would have changed if Facebook had more balanced news exposure. In Appendix Table A.7, I find that if Facebook had an equal share of pro-attitudinal and counter-attitudinal news, affective polarization would have decreased by 0.19 standard deviations and the feeling thermometer measure would have decreased by 3.91 degrees. 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 then 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.

However, perhaps having a balanced news feed is not a realistic counterfactual since 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 news consumed through Facebook had a similar congruence scale to news consumed through other means. In Appendix Table A.8, I find that affective polarization would have decreased by 0.06 standard deviations and the feeling thermometer outcome would have decreased by 1.09 degrees.[30] These back-of-the-envelope calculations should be interpreted carefully since they ignore general equilibrium effects (e.g., it is likely that if Facebook drastically changed its feed, individuals would use other social networks instead). Nevertheless, they suggest that the feed can play an important role in amplifying or mitigating polarization.

6 Interpretation

How should we interpret the fact that individuals attitudes toward parties change, while their political opinions remain stable? The results are consistent with a secular trend of a sharp increase in affective polarization over the last couple of decades, while issue-based polarization has not increased in a similar fashion (Gentzkow, 2016; Lelkes, 2016; Mason, 2015). In this section, I place the results in a broader theoretical framework. First, I discuss existing models of persuasion and argue that the results fit a model of rational informed consumers. Second, I discuss the effect on affective polarization and suggest a framework explaining why attitudes toward parties are affected without a similar effect on political opinions.

[30]The result is based on the following calculation: First, I find the difference in the control group between the congruence scale of news sites visited through Facebook and the congruence scale of all other news sites. Then I calculate the effect of a change in the congruence scale of the Facebook feed on the congruence scale of news sites visited. Using these two numbers I estimate by how much the congruence scale of participant's Facebook feed would have to decrease in order for the participant to consume news through Facebook with the same congruence scale as other news consumed. Finally, I estimate the effect of such a change on affective polarization.

6.1 Persuasion and Political Opinions

Consider the following model of media persuasion, based on DellaVigna and Kaplan (2007). There is some state of the world θ, where a higher θ represents a more conservative belief. Consumer i has a prior on the state of the world $\theta_i \sim (\theta_i^0, \frac{1}{h_i})$, where θ_i^0 is the consumer's initial belief and h_i is the precision of the belief.

Outlet j receives a signal on the state of the world $s \sim N(\theta^*, \frac{1}{h_S})$, where s is the signal, θ^* is the true state of the world and h_S is the precision of the signal received. While all outlets receive the same signal, each outlet reports the signal with a bias according to the outlet's ideological slant: $r_j = s + b_j$ where a larger b_j is a more conservative bias. The ideological slant of outlets can be explained by owner incentives (Anderson and Mclaren, 2012), or by an attempt to maximize market share (Gentzkow and Shapiro, 2006).

Since consumers know that the outlets are biased, they do not take the reports at face value, but instead interpret them as $f(r_j, b_j)$. The consumer's posterior is the weighted average of her prior and the adjusted report: $\theta_i^1 \sim N(\frac{h_i * \theta_i^0 + h_S * f(r_j, b_j)}{h_i + h_S}, \frac{1}{h_i + h_S})$. I consider three theories for how beliefs are updated according to different interpretation functions f:

- **Rational informed consumers**: $f(r_j, b_j) = r_j - b_j = s$

  Rational consumers, who know the slant of the outlet, should adjust the report and take into account only the original signal observed by the outlet. Thus, they completely ignore the outlet's bias.

- **Naive consumers**: $f(r_j, b_j) = r_j = s + b_j$

  Naive consumers do not take into account the slant of the outlet, for example, due to failing to account for correlated signals and repetition of information ("persuasion bias") (DeMarzo et al., 2003).

- **Motivated reasoning**: $f(r_j, b_j) = \begin{cases} s & \tau < |r_j - \theta_i| \text{ 人 } |r_j - b_j - \theta_i| \leq \tau \\ s + b_j & \text{otherwise} \end{cases}$

  Consumers suffering from motivated reasoning prefer to keep their original opinions. Therefore, they take into account the bias of the outlet only when the report they receive is sufficiently different than their current opinion and taking into account the bias will allow them to adjust their opinions less. Note that this could lead to a backlash effect. Suppose that a liberal outlet receives a very conservative signal, but reports the news as if it supports the liberal position. A conservative taking into account the slant of the outlet may become more conservative when consuming news from the outlet, while a liberal who does not take into account the slant of the outlet may become more liberal.

I assume that a consumer's political opinion is a weighted average of K beliefs $\gamma_i = \sum_{k \in \{1..K\}} w_{ik} \theta_{ik}$, where $0 \leq w_{ik} \leq 1$ 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 the specific belief. For example,

a consumer's support for a bill to price carbon will probably depend on whether she believes climate change is happening, whether the bill will mitigate emissions and whether it will increase consumer prices. A liberal might put more weight on the effect of the bill on emissions while a conservative may care more about prices.[31]

Applying the theories to the experiment leads to clean predictions. If a consumer is rational, it should not matter if she is assigned to the liberal or conservative treatment when updating her priors. Since the bias of the reports of all outlets is taken into account, they should lead the consumer to update her beliefs in the same way. Furthermore, assuming individuals in the control group are still exposed to news, they should also update their beliefs similarly. In contrast, if the consumer is naive, she does take into account the slant of the outlet and should become more conservative when exposed to conservative outlets, compared to the effect of exposure to liberal outlets.

The results of the experiment support the first theory. Despite large effect of the intervention on news consumption, and even though the political opinion index focused on topics discussed in the news for which consumers would be expected to have weaker priors, the treatment did not significantly affect political opinions. Moreover, as shown in Appendix Figure A.5 there is no evidence for liberals being persuaded only by the liberal treatment or conservatives being persuaded only by the conservative treatment, which could have been expected according to the motivated reasoning theory.[32]

6.2 Persuasion and Affective Polarization

In contrast to political opinions, most of the literature on affective polarization is new and there is no overarching framework for the topic. In this section, I suggest a framework which explains why affective polarization can increase without a change in political opinions.

A straightforward way to model the attitude of individual i toward party p is to define it as a function of the distance between the political opinion of the party (γ_p) and a benchmark for the "correct" opinion according to individual i, $\hat{\gamma}_{ip} = \phi(\theta_{i1},...,\theta_{ik},w_{i1},...,w_{ik},\theta_{p1},...,\theta_{pk},w_{p1},...,w_{pk})$, where $\hat{\gamma}_{ip}$ is the opinion individual i thinks party p should hold. Assume affective polarization is a linear function of the distance between the opinion of the party and the benchmark: $g(\gamma_p - \hat{\gamma}_{ip})$. I consider two functions ϕ determining the benchmark opinion.

- **Affective polarization due to political distance**: $\hat{\gamma}_{ip} = \gamma_i = \sum_k w_{ik}\theta_{ik}$

[31] The example was intentionally chosen to be general. The questions forming the political opinions index are on more specific topics, but the same logic holds. For example, participants' favorability of the March for Our Lives Movement could depend on their beliefs on whether gun violence is a serious problem, whether banning certain weapons will help deal with the problem,on whether gun owners may not be able to purchase their preferred guns if the movement accomplishes its goals, and whether the leaders of the movement attack gun-right supporters. Consumers will probably place different weights on different beliefs. For example, a conservative might place a higher weight on the effect of the movement on gun right supporters, while a liberal may place lower weights on those effects.

[32] The precise predictions of the motivated reasoning theory require additional assumptions. If the report the consumer receives from the counter-attitudinal outlets is distant enough from her beliefs, and the report received from the pro-attitudinal outlets is not too distant, the report from the pro-attitudinal outlets would be taken at face value, while the consumer would respond to the actual signal when consuming the report from the counter-attitudinal outlets. If consumers in the control group are already exposed to the signal, due to other news they consume, we would expect only the pro-attitudinal outlet to affect opinions.

Consumers who only care about political opinions will use their own opinions as the benchmark for the correct opinions and determine their attitudes toward a party based on the difference between their political opinions and the party's political opinions. When political opinions change from γ_i^0 to γ_i^1, the following change in expected in affective polarization (ΔA_i)

$$\Delta A_i = g(\gamma_i^1 - \gamma_p) - g(\gamma_i^0 - \gamma_p) = g(\sum_k w_{ik}(\theta_{ik}^1 - \theta_{ik}^0)) \tag{3}$$

This theory predicts that an update in consumer beliefs should only affect attitudes towards a party through its effect on the consumer's political opinions. The effect on affective polarization should have the same sign as the predictions for political opinions (e.g., if a consumer forms a more conservative opinion after updating her beliefs, she should also form a more positive attitude towards the Republican party). 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. If a consumer is randomly exposed to liberal outlets and as a result her support for a climate bill increases, her attitude toward the Democratic party should become more positive and her attitudes towards the Republican party should become more negative.

- **Affective polarization due to unreasonable opinions:** $\hat{\gamma} = \sum_k w_{pk}\theta_{ik}$

  Alternatively, consumers may judge whether the political opinion of a party is reasonable according to the party's own weights. Hence, their attitude toward a party is based on the distance between the party's political opinion and the political opinion the party is expected to hold according to the consumer's beliefs (but keeping the party's own weights). According to this theory, individuals understand and accept the fact that different parties place different priorities, but develop negative attitudes toward a party when they believe that even according to the party's weights, the party should change its political opinions. The change in affective polarization following updated beliefs is now:

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

  If the consumer and the party place the same weight on beliefs, there is no difference between the two theories. However, if not all beliefs are similarly updated, 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 not increase consumer prices, will support the bill. The consumer will hold negative attitudes toward a party opposing the bill, since even if the party does not place a high weight on decreasing emissions, it should support the bill as long as it has no negative consequences. If the liberal is exposed to conservative outlets and learns that the bill may increase prices, she may still support the bill since she places higher weights on mitigating emissions, but will develop less negative attitudes toward a party that places a high weight on consumer prices and thus opposes the bill. In other words, the liberal consumer will better understand the rationale behind the argument objecting to the bill even if she does not agree with the importance of the argument.

As shown in Figure 9, consumers attitudes toward parties are affected by news exposure. The result is not consistent with the first theory, defining affective polarization as a function of distance in political opinions, since affective polarization was affected by the treatment, in contrast to political opinions. The results do not contradict the second theory. If consumers naively updated only beliefs on which they place relatively low weights, their political opinions may stay the same, but their attitudes could change.[33]

To further test the theories, I analyze the effect of the experiment on attitudes towards parties. Affective polarization is usually measured as the difference between individuals' attitudes toward their own party and their attitudes toward the opposing party, but we can focus on each party separately. If media outlets are delegates of their consumers (Gentkow et al., 2015), we would expect pro-attitudinal outlets to cover more issues for which $w_{OWN} > w_{OPPOSING}$ and counter-attitudinal outlets to cover more issues for which $w_{OPPOSING} > w_{OWN}$, where w_{OWN} are the weights placed by the individual's party and $w_{OPPOSING}$ are the weights placed by the opposing party.

The two theories differ in their predictions regarding the effect of the treatments on attitudes toward each party. If affective polarization is merely a function of political distance, attitudes will mostly be affected when consumer i updates beliefs on which she places higher weights (see equation 3). Therefore, attitudes toward both parties should be strongly affected by exposure to pro-attitudinal outlets, that are more likely to cover beliefs on which i places high weights. On the contrary, if affective polarization is a function of unreasonable opinions, attitudes toward party p will be affected more by beliefs for which p places higher weights (see equation 4). Therefore, pro-attitudinal outlets are expected to have a greater effect on one's attitudes toward their own party, while counter-attitudinal outlets will affect attitudes toward the opposing party. Table 8 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 (not formally presented) which argues that affective polarization is increasing because of negative coverage of pro-attitudinal outlets focusing on the opposing party (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 channel 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 mostly explained by increased negative coverage in the news media. 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. The theory predicts that affective polarization will decrease when individuals are more likely to believe the arguments supporting the other side's point of view, even if that will not change their ultimate political opinions due to differences in priorities (weights). While I provide evidence supporting the theory,

[33] Alternatively, it is possible that consumers are generally naive when updating all beliefs, but have much stronger priors regarding beliefs on which they place higher weights when determining their political opinions. As a result, beliefs on which the consumers place higher weights are not strongly affected by the experiment and the consumers' political opinions do not change. If consumers have weaker priors on beliefs for which they place low weights and the opposing party places higher weights, their attitudes toward the party could be affected by the intervention.

there could be other explanations for the change in affective polarization,[34] and more research is needed to pinpoint the precise mechanisms explaining how affective polarization evolves.

7 Mechanism: Why Does Social Media Increase Exposure to Pro-Attitudinal News?

The previous sections showed that changes in the social media feed affect browsing behavior and that modified news consumption habits affect partisan hostility. Therefore, it is important to understand what influences the news individuals are exposed to. I first approach this question using descriptive statistics on the sources of news segregation on Facebook. I then exploit variation induced by the treatment assignments to decompose the difference in exposure to posts from pro-attitudinal outlets and posts from counter-attitudinal outlets offered in the experiment.

7.1 Subscription to Outlets vs. Social Network

Figure 11a uses extension data from the experiment's control group to show that news consumed through Facebook tends to better match consumers' ideologies. This provides additional confirmation, using a different sample, to the results presented in Section 2.2. 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 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 the news they consume? The first theory is tested in the first row of Figure 11b, which presents the distribution of the slant of news sites visited through a link shared by friends on Facebook. The second row in the figure tests the second theory by presenting the slant for news sites visited when a post was shared by a Facebook page.[35]

I find that links clicked through Facebook pages (which include media outlets individuals subscribe to) are driving the increased consumption of pro-attitudinal news. For example, when websites are not visited specifically through Facebook, approximately 17% of news sites visited by conservatives consumers are very conservative. When conservative consumers visit news sites through posts shared by their Facebook friends the share is similar, while when they visit news sites through posts shared by Facebook pages they subscribe to the share of very conservative websites increases to 28%.

[34] For example, Mason (2015) explains that partisan bias may increase without changes in position extremity as a result of stronger partisan identity. In other words, the intervention may have amplified or mitigated tribalism, which can increase affective polarization, without having a strong effect on political opinions. In Appendix Figure A.9 I find that the conservative treatment did not increase identification with the Republican Party. The liberal treatment is associated with a slight increase of identification with the democratic party, however the point estimate is small, it is precisely estimated and it is not statistically significant.

[35] Based on the control group, I estimate that 59.9% of visits to news sites are through Facebook pages. Both posts supplied by friend and posts supplied by pages are affected by Facebook's algorithm

7.2 Social Media's Algorithm Decreases Exposure to Counter-Attitudinal News

Since posts from news outlets lead to greater consumption of pro-attitudinal news, it is important to understand what determines exposure to these posts in the Facebook feed. This section decomposes the segregation in social media news exposure while focusing only on posts shared by the news outlets included in the experiment. As shown in Panel 2 of Figure 5 the pro-attitudinal treatment increased exposure to pro-attitudinal posts by approximately twice as much as the effect of the counter-attitudinal treatment on counter-attitudinal posts.

Figure 12 decomposes the increased exposure to pro-attitudinal posts to three main forces: Individuals preferring to subscribe to pro-attitudinal news outlets ("selective exposure"); Facebook's algorithm supplying more posts from pro-attitudinal outlets, conditional on subscription (the "filter bubble" effect); and individuals using Facebook less often when offered counter-attitudinal outlets.

This accounting exercise is based on the following framework: $E_{ij} = S_{ij}F_{ij}U_i$

where E_{ij} is the number of posts from outlet j individual i was exposed to. Exposure is a product of whether individual i subscribed to outlet j ($S_{ij} \in \{0,1\}$), the share of posts supplied from the outlet among all posts the participant observed (F_{ij}) and the total number of posts individual i observed in Facebook (U_i).

I decompose exposure using the following formula

$$\Delta E = \underbrace{S_\Delta * F_A * U_A}_{\text{Selective Exposure}} + \underbrace{S_C * F_\Delta * U_C}_{\text{Filter Bubble}} + \underbrace{S_C * F_C * U_\Delta}_{\text{Time on Platform}} + \underbrace{S_\Delta * F_\Delta * U_C + S_C * F_\Delta * U_\Delta + S_\Delta * F_\Delta * U_\Delta + S_\Delta * F_\Delta * U_C}_{\text{Combinations}} \tag{5}$$

where for each variable the $_C$ subscript is used to notate the value for the counter-attitudinal treatment and $_\Delta$ subscript is used to notate the difference between the pro-attitudinal and counter attitudinal treatments.

- ΔE is the difference between exposure to posts from the potential pro-attitudinal outlets in the pro-attitudinal treatment and exposure to posts from potential counter-attitudinal outlets in the counter-attitudinal treatment.
- S is subscriptions to outlets and is estimated by pooling all potential outlets and participants and regressing subscription to an outlet on the full interaction of whether an outlet was offered to the participant and whether the outlet is pro-attitudinal (Appendix Table A.9, column 1).
- F is the share of posts supplied from outlets, conditional on subscription to the outlets, among all posts the participant was exposed to. F is estimated by pooling all potential outlets and participants and regressing the share of posts supplied from an outlet, among all posts a participant was exposed to, on the full interaction of whether the participants subscribed to the outlet and whether the outlet is pro-attitudinal, with subscription to an outlet instrumented with whether the outlet was randomly offered (Appendix Table A.9, column 2).[36]

[36] F 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

- U is the total number of posts participants were exposed to on Facebook. U is estimated by regressing the total number of posts participants were exposed to on whether they were assigned to the pro-attitudinal or counter-attitudinal treatment (Appendix Table A.9, column 3).

Selective exposure can be interpreted as the additional posts from counter-attitudinal outlets participants in the counter-attitudinal treatment would have been exposed to if they would have subscribed to the same number of outlets as participants in pro-attitudinal treatment. The filter bubble measures the additional posts subscribers to counter-attitudinal outlets would have been exposed to, if Facebook would have supplied them with the same share of posts from these outlets, as the share supplied from the pro-attitudinal outlets. The time on platforms measures the additional posts which would have been viewed by participants assigned to the counter-attitudinal treatment, if they would have used Facebook as much as participants in assigned to the pro-attitudinal treatment.

I find evidence for all three effects: participants prefer to subscribe to pro-attitudinal news outlets, participants are supplied with a smaller share of posts from counter-attitudinal outlets, conditional on subscription, and participants decrease their Facebook usage after they are offered to subscribe to counter-attitudinal outlets.

The strongest force driving exposure to pro-attitudinal news in the experiment is the supply of posts by the platform. This provides clear 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. The result also suggests why personalization is leading to segregation online - when consumers are exposed to more counter-attitudinal news, they decrease their Facebook usage and therefore platforms have a clear incentive to provide individuals with news that matches their opinion, in order to maximize engagement. While I focus on Facebook, the logic probably 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.[37]

Throughout this section, I focus on outlets to measure the slant of the news participants engaged with, as is typical in media studies (Gentzkow and Shapiro, 2011; Guess, 2018). An additional margin of personalization which can increase segregation is variation in posts within an outlet. A platform could supply conservatives with the more conservative articles posted by an outlet and supply liberals with the more liberal articles posted by the same outlet.

I find no evidence for personalization within an outlet. I focus on the subset of articles that were shared by at least one Member of Congress in 2018 and define the slant of an article according to the mean

treatments might be due to treatment heterogeneity and selection into compliance, and not due to different treatment effects. In a future robustness section, I plan on re-weighting the IV estimators according to the baseline covariates of compliers (Angrist and Fernandez-Val, 2013; Aronow and Carnegie, 2013) in order to estimate the ATE for never-takers and for the entire sample, and testing whether this affects the decomposition results.

[37] Factor taken into account when determining the relevancy of tweets include the tweet's author and the user's past relationship with the author, therefore it is likely that tweets from pro-attitutidnal account will receive a higher ranking. For more details see: Using Deep Learning at Scale in Twitter's Timelines. https://blog.twitter.com/engineering/en_us/topics/insights/2017/using-deep-learning-at-scale-in-twitters-timelines.html

DW-Nominate score of Congress Members who shared the article.[38] Using this measure, I find that generally conservatives 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 from the eight potential outlets defined for each consumer, I do not find any correlation between the slant of the articles and consumers' ideologies (see Appendix Table A.10). This suggests that Facebook's algorithm does not lead to conservatives being supplied with more conservative articles, *within* an outlet. It also implies that conservatives and liberals were exposed to similar content from the outlets they subscribed to in the intervention, conditional on a post from the outlet appearing in their feed.

To conclude, this section finds that selection, time spent on a platform and especially platform algorithms decrease exposure to counter-attitudinal news. It is worth noting that this section does not suggest that Facebook's algorithm intentionally targets news outlets according to whether they share the consumer's beliefs. It is more likely that the platform learns what the consumer may be interested in based on the consumer's behavior, her social network and other pages she is subscribed to. Still, the effect of personalization on news exposure is an important departure from how news was supplied and consumed in the past. This result apply not only to social media platforms, but generally to most online news, as major news outlets are also personalizing their websites and the articles they suggest to their customers, and as a result, may be increasing the consumption of pro-attitudinal news.[39]

8 Conclusions

Consumption of news through social media is increasing, but the effect of social media on public opinion remains controversial. This paper sheds a light on how social media affects news consumption, political opinions, and affective polarization.

The study shows that individuals are willing to be exposed to new opinions through social media. Participants in the experiment do not only subscribe to cross-attitudinal news outlets, but they also consume and share news from those outlets. However, the algorithms determining news exposure mitigate potential exposure to counter-attitudinal news. This "filter bubble" effect did not exist until recently and may have stronger impacts in the future, with the development of more sophisticated machine learning algorithms customizing news options. These effects are important since I find that the social media feed has a large influence on online news consumption habits.

The study suggests that a more nuanced view is needed regarding the effect of media on political beliefs. On the one hand, exposure to pro-attitudinal news increases affective polarization, compared to exposure to counter-attitudinal news. Therefore, there is room for concern that exposure to news on

[38]The list of the Facebook pages of Members of Congress is based on the Congress Members project (https://github.com/unitedstates/congress-legislators). The list was used to collect 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).

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

social media is leading to more negative attitudes across parties. On the other hand, it seems that individuals are not so easily persuaded by what they consume online, and exposure to more liberal or conservative news does not affect the political opinions of participants. Perhaps concerns that elections were determined by manipulation of opinions on social media are overstated. The fact that the results are in line with long term trends in polarization and political opinions suggests that a segregated news environment can explain the increase in affective polarization over the past several decades.

Similarly to other randomized control trials, one should be careful when extrapolating the results of the study. The experiment took place in the first half of 2017. In this period, Facebook was often criticized for the effect of news-related content users are exposed to on social media, and seemed to have responded by decreasing the exposure to news outlets on the platform. It is possible that in a different period, the effects would have been larger. Similarly, Trump's presidency is exceptional in the stability of the president's approval ratings.[40] If other opinions were relatively stable throughout the period as well, the minimal effect found on political opinion could partly be explained by the period when the survey took place. Finally, the experiment compared outlets with a clear ideological slant. It would be interesting in a future study to also randomize access to moderate outlets, such as USA Today, and measure their effects on beliefs. While acknowledging these limitations, the experiment has high external validity when it comes to analyzing actual behavior on Facebook around 2017, as supply and consumption of news occurred just as they do when individuals subscribe to any other outlet on Facebook.

Finally, this study has clear policy implications. Scholars have suggested that new tools are required to expose individuals to more cross-cutting news (Sunstein, 2017). The experiment described in this paper effectively measures the effect of such an intervention and shows that since individuals are willing to be exposed to other viewpoints, nudges which lower the search costs for more diverse information may be effective. For example, if social media companies were to show users leading articles from outlets that usually do not appear in their feed, individuals might find it easier to understand other opinions and develop less negative attitudes toward supports of opposing parties.[41] However, the study also shows why companies may not have an incentive to implement these policies, since increased exposure to counter-attitudinal news may decrease engagement with the social media platform. To conclude, while social media may be increasing affective polarization through the increased consumption of pro-attitudinal news, it also has the potential to mitigate these effects.

[40] For example: Dann, Carrie and Murray, Mark - NBC/WSJ poll: Trump Approval 'Remarkably Stable' After a Stormy Week of Bad News. NBC News. August 26, 2018; Yokley, Eli - New Poll Suggests Voters Have Made Up Their Minds on What Trump Is Like. Morning Consult. July 26, 2018.

[41] For example, in 2017 Facebook implemented a feature which showed users articles from additional outlets related to a post in their feed. Facebook Newsroom - New Test With Related Articles. April 25, 2017. Online: https://newsroom.fb.com/news/2017/04/news-feed-fyi-new-test-with-related-articles/. Facebook recently confirmed plans for a new News Tab which will focus exclusively on news and will be curated by both humans and algorithms. Such a tab may decrease polarization if it exposes users to diverse viewpoints, but the algorithms will work similarly to the rest of the platform and expose individual to more pro-attitudinal news.

In August 2018 Twitter announced that it will allow users to follow topics instead of specific accounts, which may also expose users to more diverse news sources. Newton, Casey - Twitter tests letting users follow topics in the same way they follow accounts. The Verge. August 13, 2019. Online: https://www.theverge.com/2019/8/13/20804476/twitter-interests-follow-topics-feature-accounts-timeline

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.

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.

Anderson, S. P. and J. Mclaren (2012). Media Mergers and Media Bias with Rational Consumers. *Journal of the European Economic Association 10*(4), 831–859.

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.

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*.

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

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 (2017). Is the Internet Causing Political Polarization? Evidence from Demographics. *NBER Working Paper*.

Chen, Y. and D. Y. Yang (2018). The Impact of Media Censorship: Evidence from a Field Experiment in China. *Job Market Paper*.

Comscore (2017). Comscore Web Behavior Database.

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.

Day, M. V., S. T. Fiske, E. L. Downing, and T. E. Trail (2014). Shifting Liberal and Conservative Attitudes Using Moral Foundations Theory. *Personality and Social Psychology Bulletin 40*(12), 1559–1573.

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

DeMarzo, P. M., D. Vayanos, and J. Zwiebel (2003). Persuasion Bias, Social Influence, and Unidimensional Opinions. *The Quarterly Journal of Economics 118*(3), 909–968.

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

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

Gentkow, 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.

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.

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.

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.

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

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

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

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.

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. *Job Market Paper*.

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.

Martin, G. J. and A. Yurukoglu (2017). Bias in Cable News: Persuasion and Polarization. *American Economic Review*.

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.

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

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

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

PEW (2019). Public Highly Critical of State of Political Discourse in the U.S. Technical report.

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 (2019). Digital News Report 2019.

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

Shearer, E. and K. E. Matsa (2018). News Use Across Social Media Platform 2018. *PEW Research Center*.

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

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

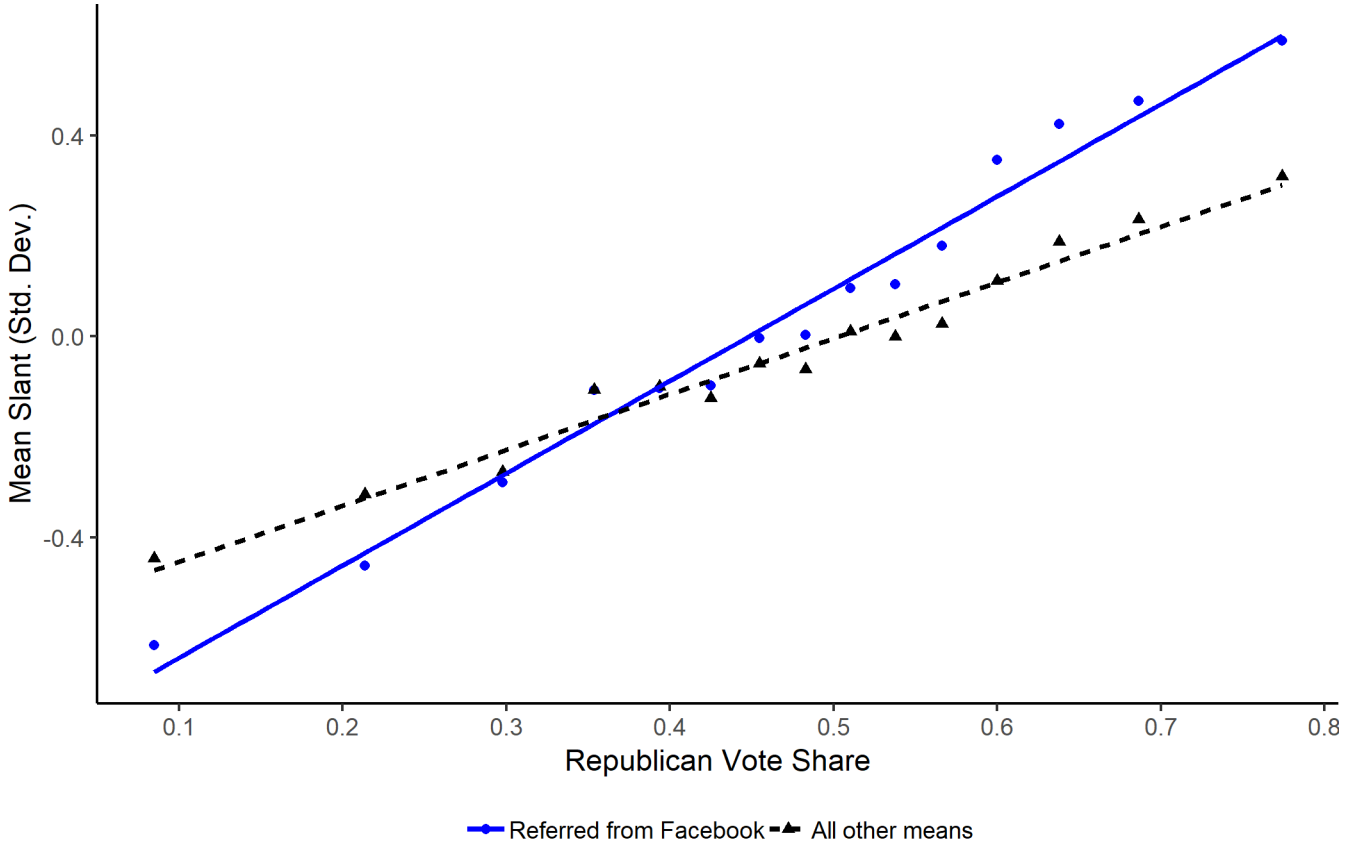
Tucker, J. A., A. Guess, P. Barbera, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan (2018). Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature.

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

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.

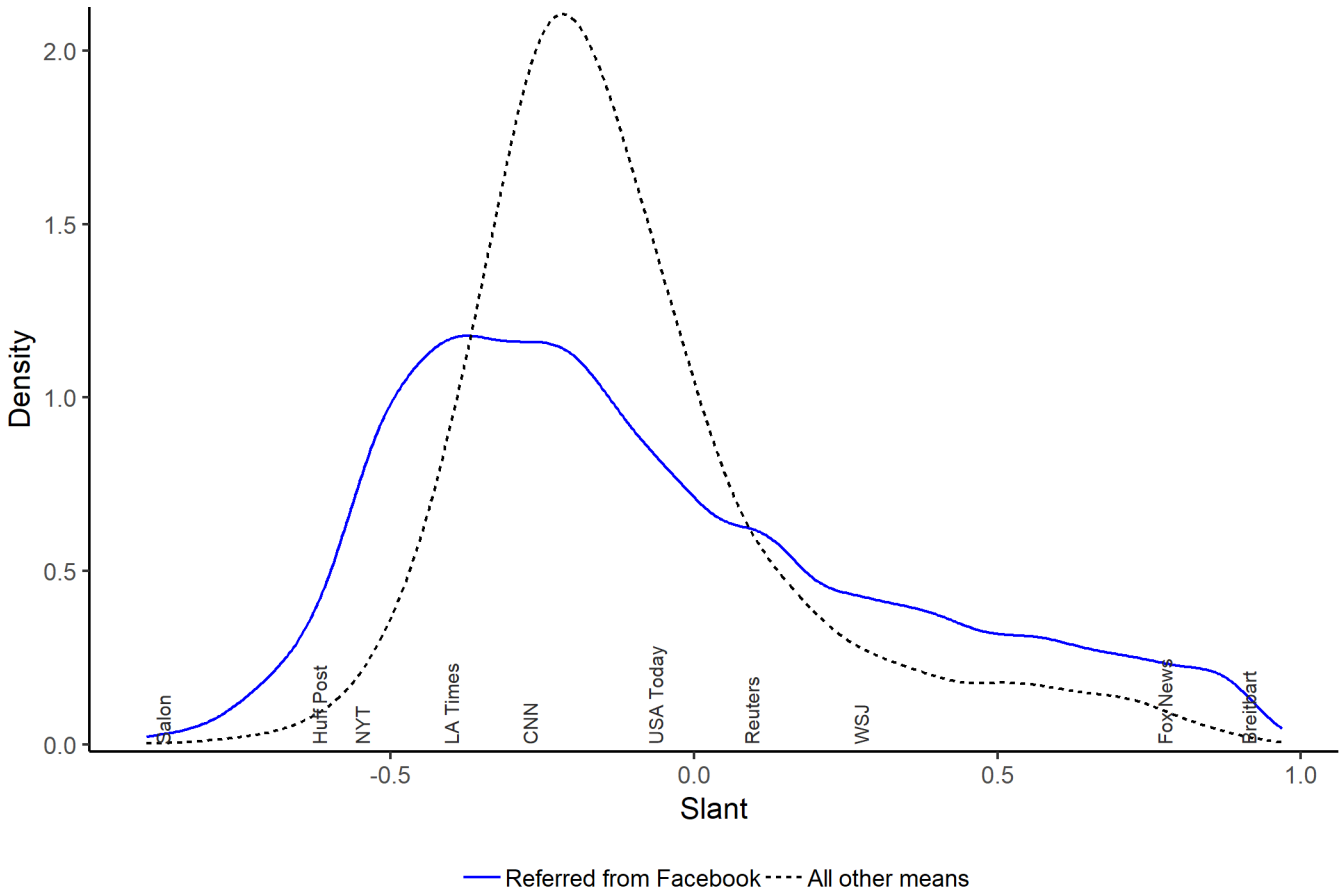
Zuiderveen Borgesius, F., D. Trilling, J. Möller, B. Bodó, C. De Vreese, and N. Helberger (2016). Should We Worry about Filter Bubbles? *Internet Policy Review 5*(1), 1–16.

Figure 1: Ideology and Slant of News Consumption



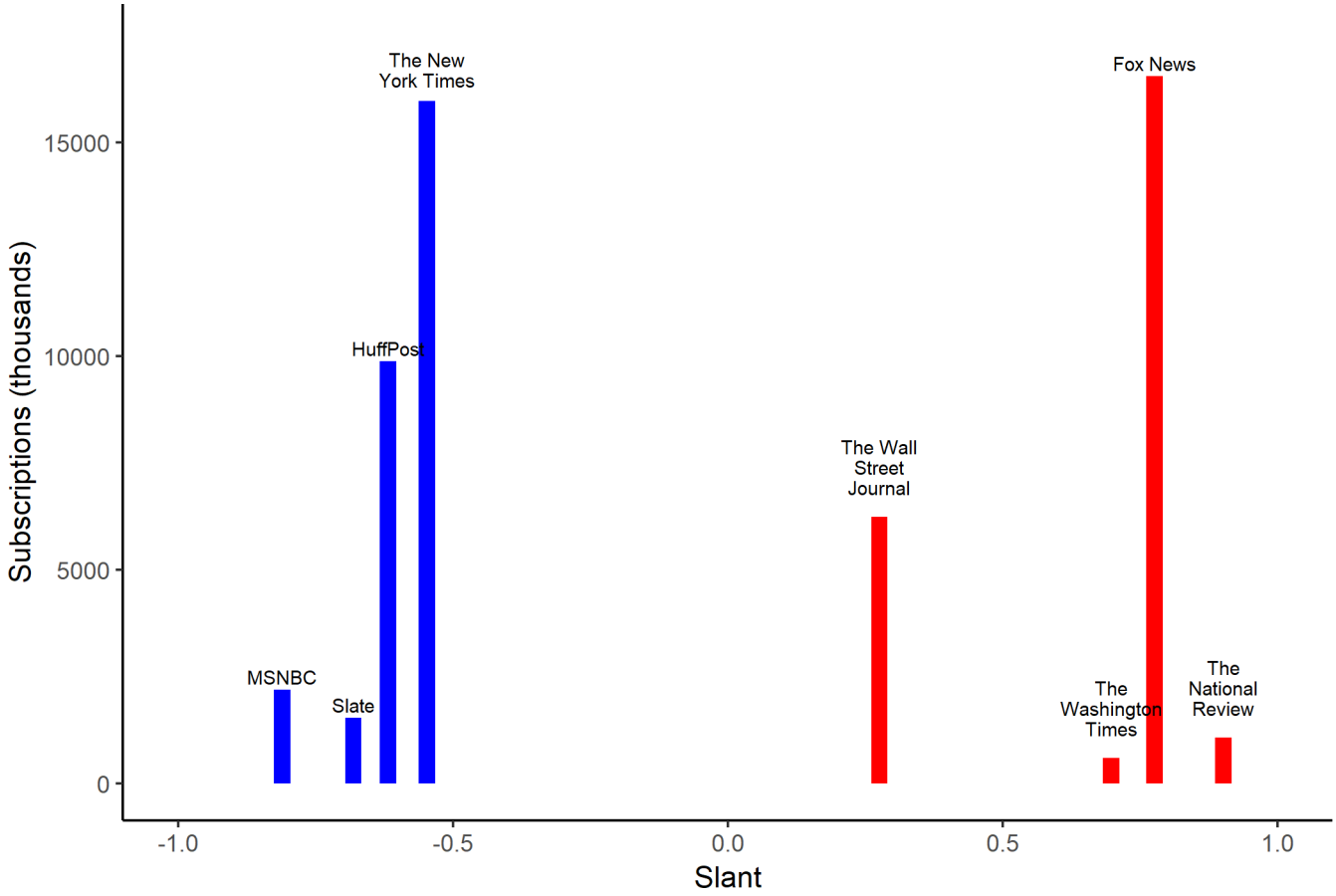
This figure shows the correlation between political ideology and online news consumption. It presents a binned scatter plot based on 2017 Comscore data. The Republican vote share in the x axis is based on 2008 zip code level voting data (Mummolo and Nall, 2016). The mean slant in the y axis is calculated as the mean slant of all news sites visited, where 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". Only users who visited at least two news sites through Facebook and two news sites not through Facebook are included in the sample.

Figure 2: Distribution of Mean News Slant



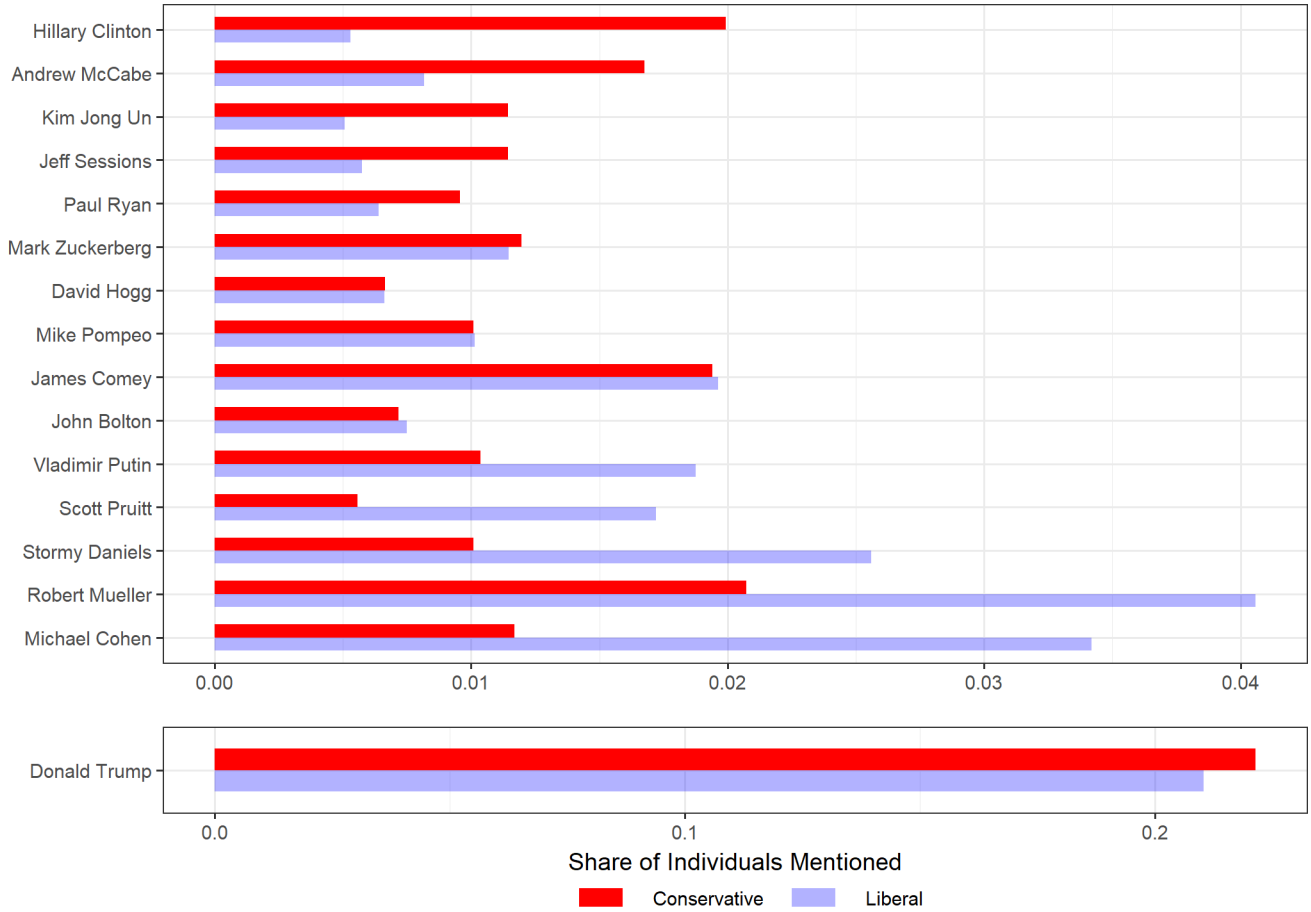
This figure shows the distribution of the mean slant of news consumed by individuals based on 2017 Comscore data. The slant of each domain is based on Bakshy et al. (2015). A visit to a news site is referred from Facebook if the referring domain is "facebook.com". Only users who visited at least two news sites through Facebook and two news sites not through Facebook are included in the sample.

Figure 3: Primary Assigned Outlets



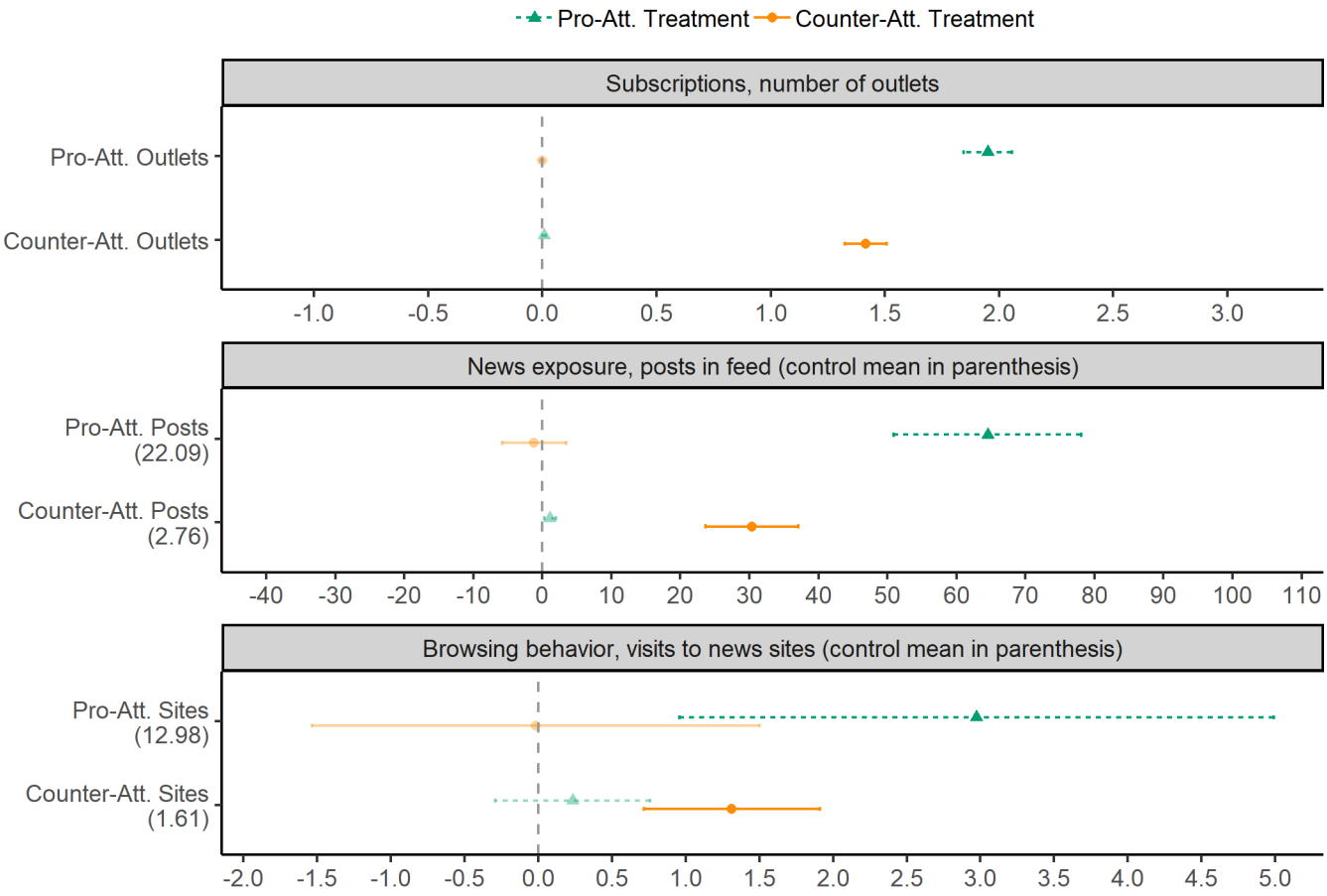
This figure displays the primary liberal and conservative outlets offered in the experiment. The x-axis is the slant of the outlets, as determined by Bakshy et al. (2015), and the y-axis is the total number of individuals who have subscribed to each outlet in April 2018.

Figure 4: Figures Discussed in the News During the Study Period



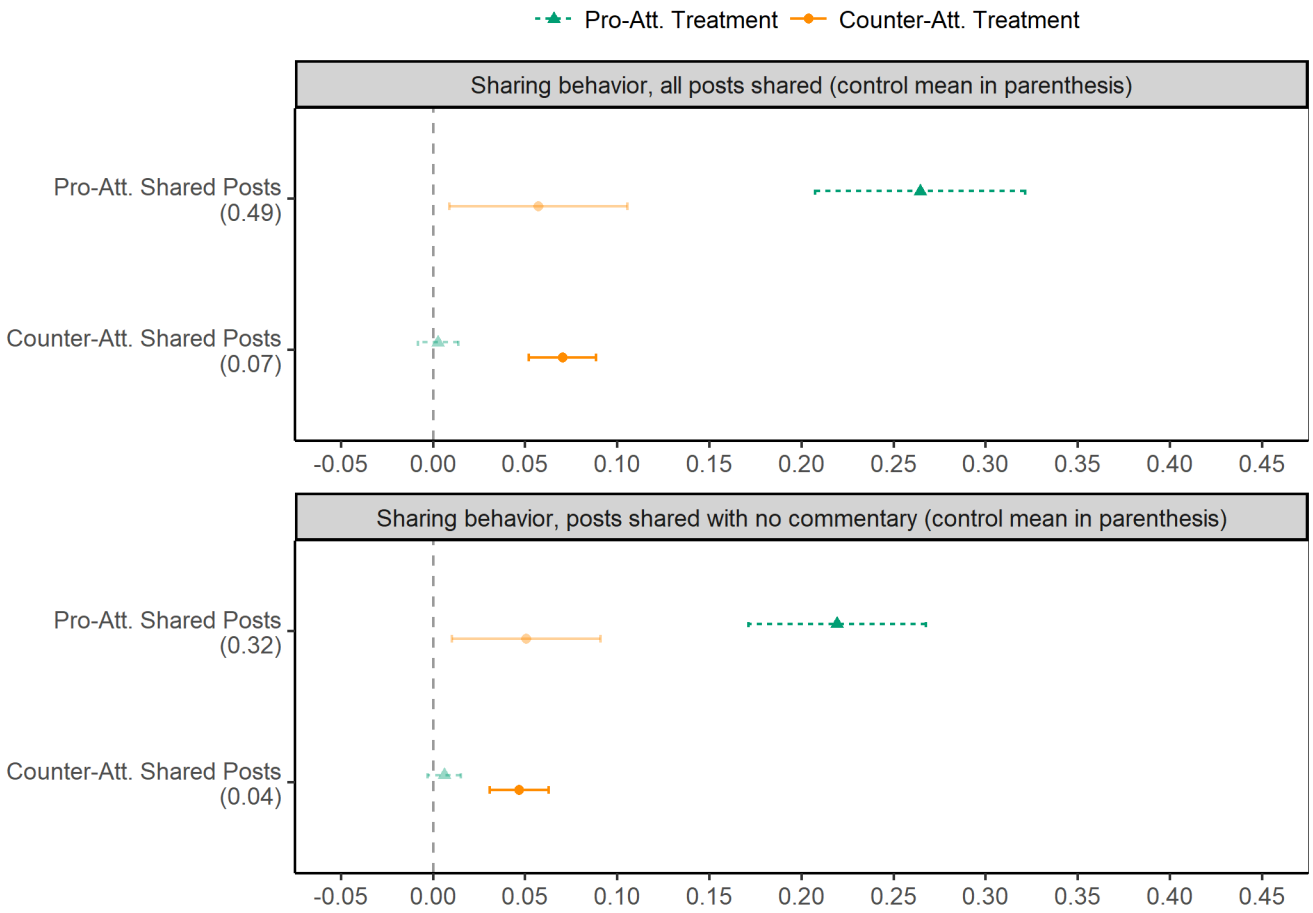
The figure shows the prominent men and women mentioned in posts shared by the primary outlets between February 28 and April 25, the median dates the baseline survey and endline survey were taken. The x axis is the share of times an individual was mentioned in a post by one of the four primary conservative outlets (top bars, in red) and by one of the four primary liberal outlets (bottom bars, in blue), 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 figure mentioned. The figures were identified using the Spacy Natural Language Processing algorithm. To simplify the graph, the names 'Trump' and 'Donald Trump' were determined to be the same individual, even though 'Trump' could refer to other members in President Trump's family.

Figure 5: Effect of the Treatment on Subscriptions, News Exposure and Browsing Behavior, Two Weeks Following the Intervention



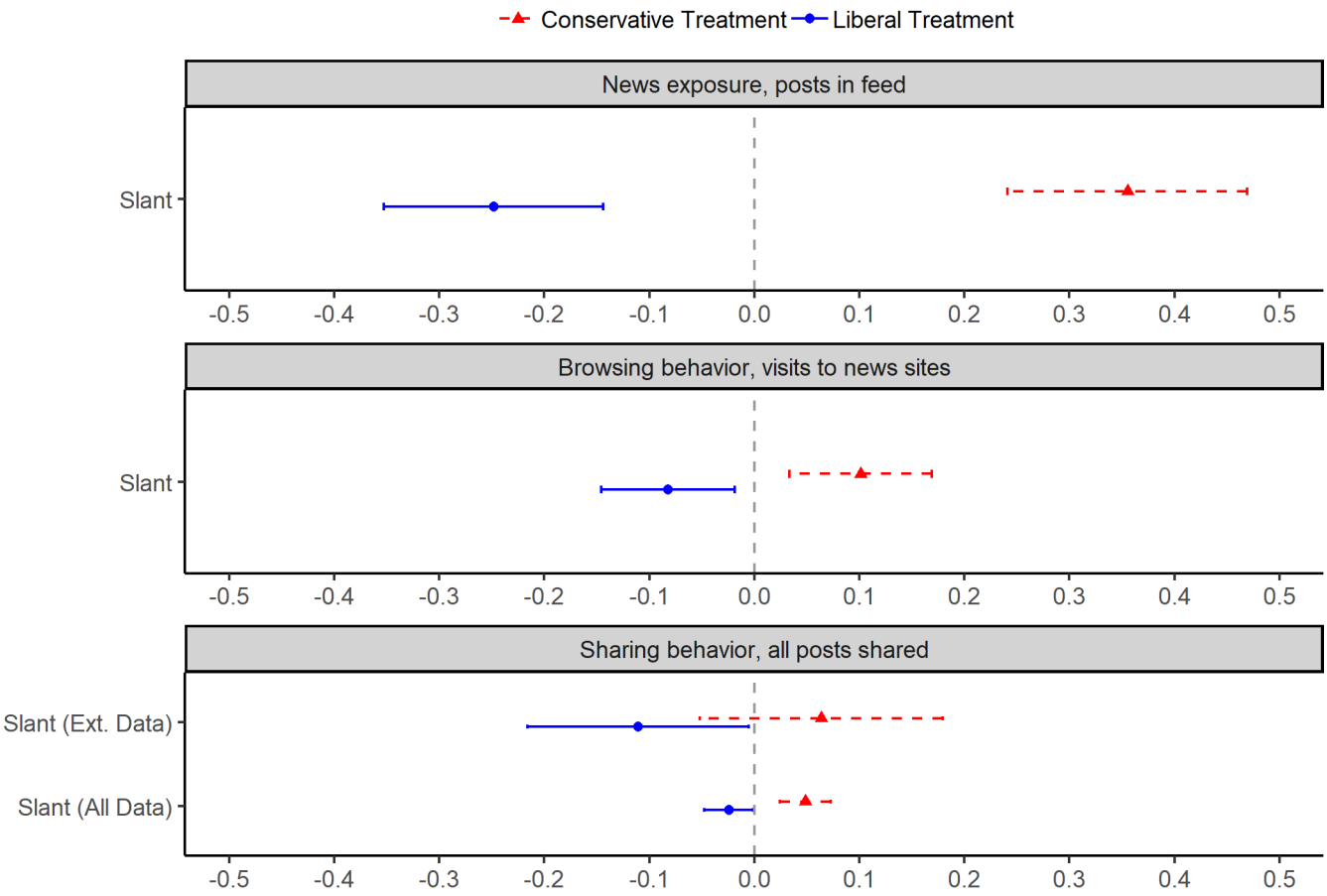
This figure shows the effect of the pro-attitudinal and counter-attitudinal treatments on engagement with each individual's potential outlets. Each row in the figure is estimated by regressing engagement with the four potential pro-attitudinal outlets or four potential counter-attitudinal outlets on the treatment. The outcomes are the number of outlets individuals subscribed to, the number of posts from the outlets that appeared in their feed and the number of times they visited the outlets' websites. The figure includes data from 1,839 participants for which full data from the extension is available for the two weeks following the intervention. For clarity, the effect of the pro-attitudinal treatment on pro-attitudinal outcomes and the effect of the counter-attitudinal treatment on counter-attitudinal outcomes are highlighted with darker color. The regressions control for the outcome measure in baseline if it exists. Error bars reflect 90 percent confidence intervals.

Figure 6: Effect of the Treatment on Posts Shared



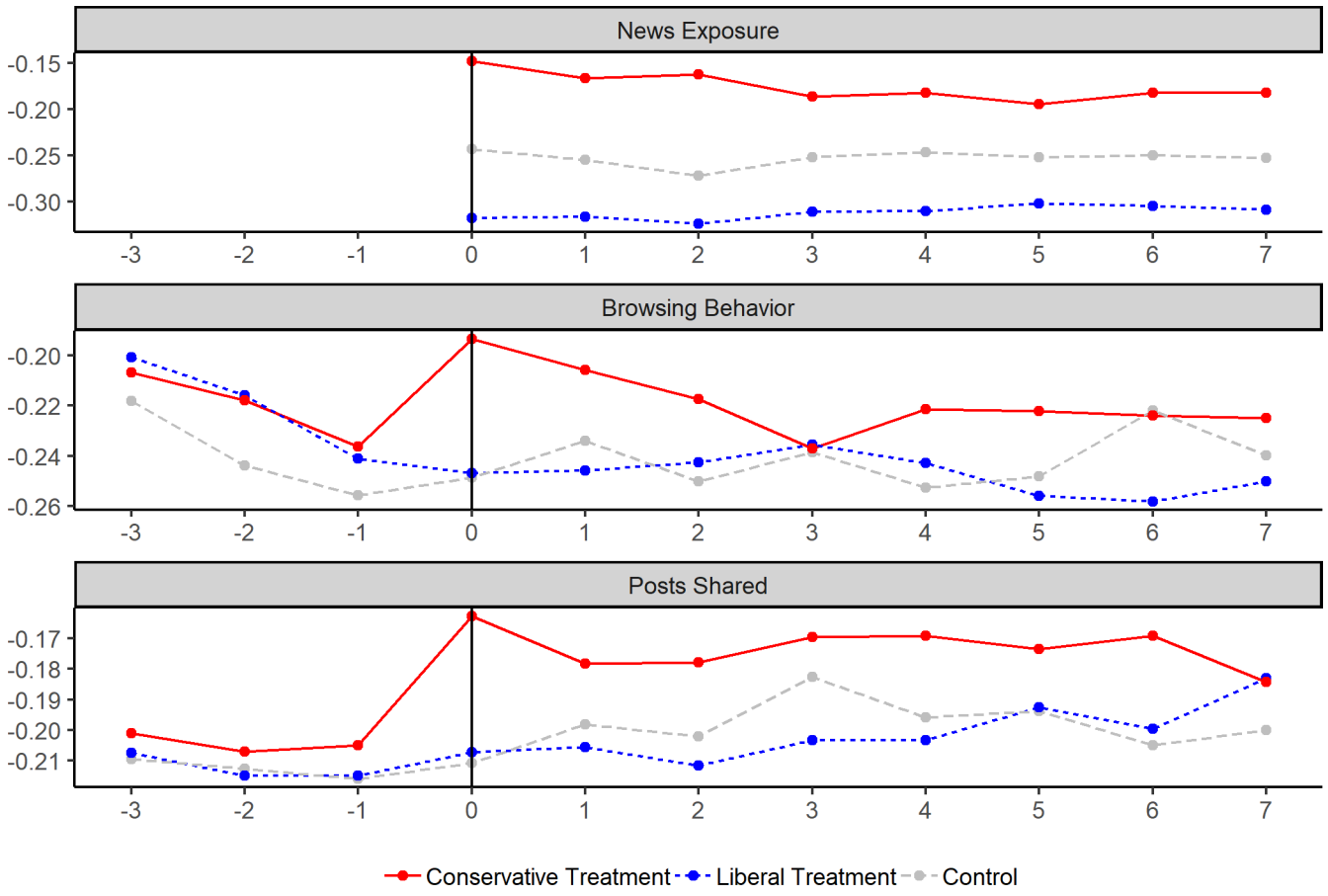
This figure shows the effect of the pro-attitudinal and counter-attitudinal treatments on engagement with each individual's four potential pro-attitudinal outlets and four potential counter-attitudinal outlets. Each row in the figure is estimated by regressing engagement with the four potential pro-attitudinal outlets or four potential counter-attitudinal outlets on the treatment. The outcomes in both panels are the number of posts from the outlets individuals shared. The first panel includes all posts and the second panel includes only posts that were shared without any commentary by the participant. The figure includes data from 34,589 participants for which data on shared posts is available for at least two weeks following the intervention. The regressions control each outcome measure in baseline. For clarity, the effect of the pro-attitudinal treatment on pro-attitudinal outcomes and the effect of the counter-attitudinal treatment on counter-attitudinal outcomes are highlighted with darker color. Error bars reflect 90 percent confidence intervals.

Figure 7: 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. Each row in the figure is estimated by regressing engagement with the four potential liberal outlets or four potential conservative outlets on the treatment. Regression control for the outcome in baseline if it exists. The figure display the slant for three outcomes: Exposure to posts on Facebook (panel 1), news sites visited (panel 2) and posts shared (panel 3). The first three outcomes include participants who installed the browser extension for at least two weeks and provided permission to access their posts for two weeks. The last outcome includes a much larger sample of all participants who provided permissions to access their posts for at least two weeks. Error bars reflect 90 percent confidence intervals.

Figure 8: Mean Slant of News Exposure, Browsing Behavior and Posts Shared in the Conservative, Liberal and Control Groups

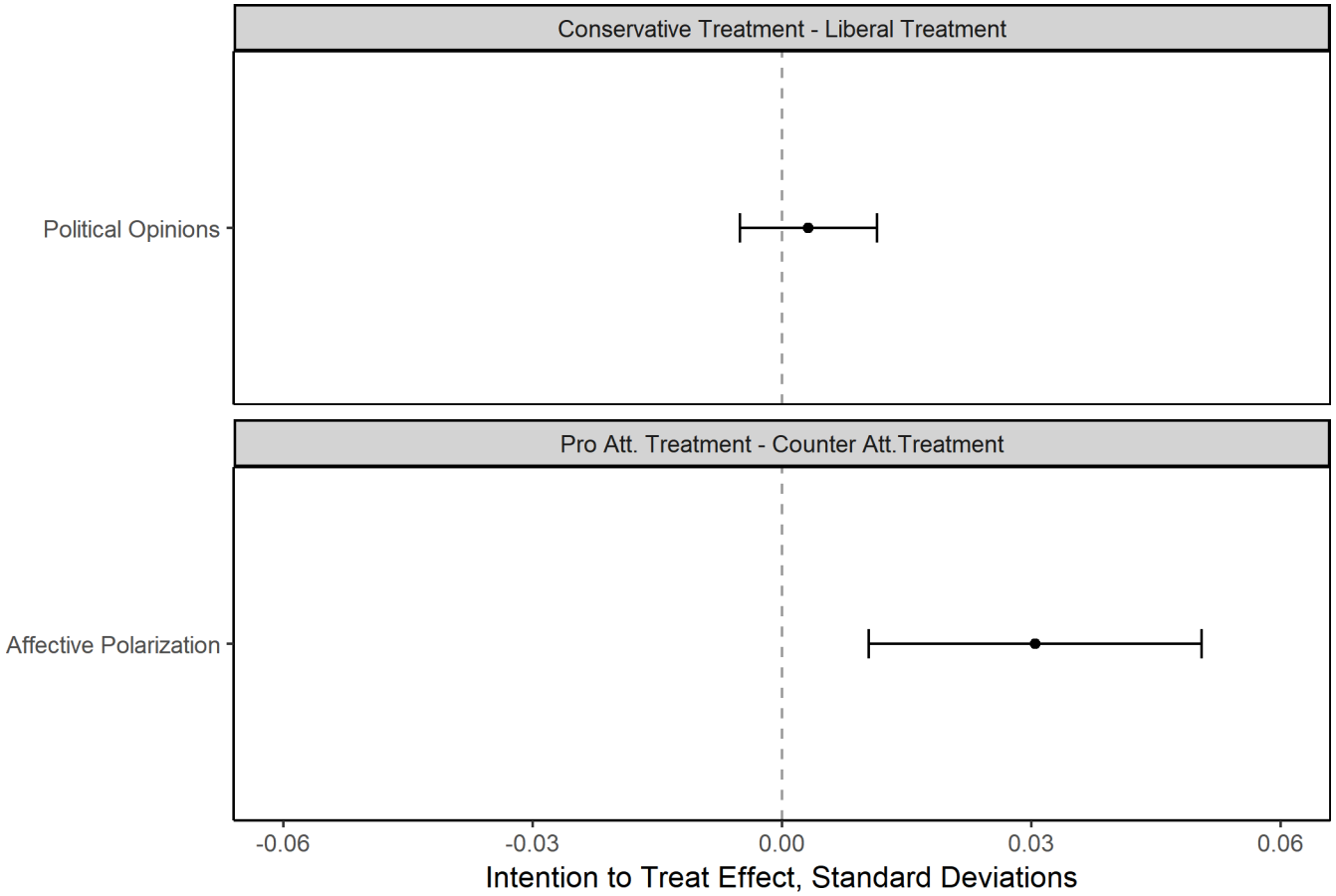


This figure shows the affect of the treatments over time. Each panel shows the mean slant for a specific week.

The first panel shows the slant of posts observed in the participants' Facebook feed by treatment. The panel does not include data from the pre-period since I do not observe the Facebook feed before the intervention. The second panel shows the slant of the websites of outlets visited, by treatment, and the third panel shows the slant of posts shared by treatment.

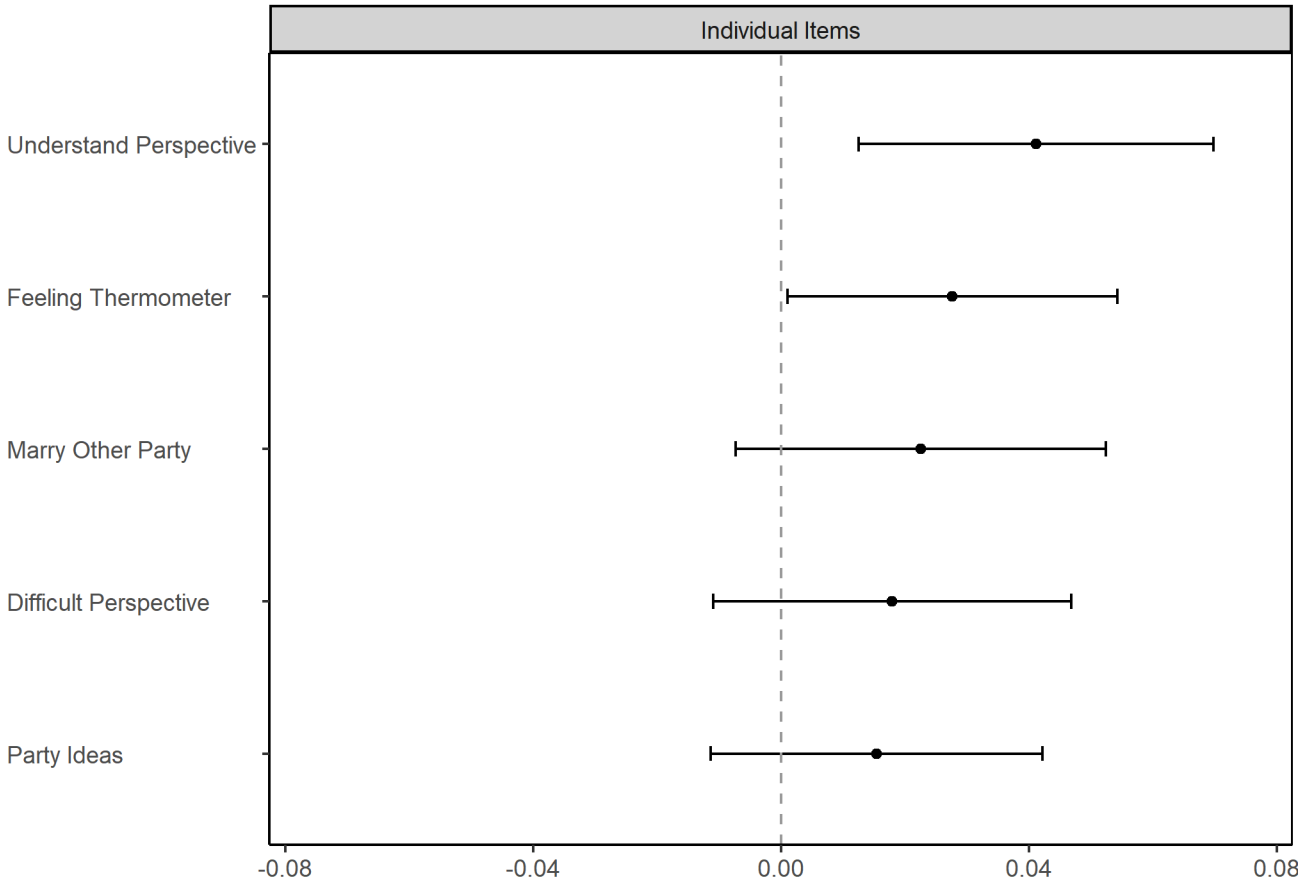
The first two panels are based on participants who installed the extension for at least eight weeks and third panel is based on participants who provided access to their shared posts for at least eight weeks.

Figure 9: Effect of the Treatment on Political Opinions and Polarization



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 pro-attitudinal treatment on the affective polarization index, compared to the counter-attitudinal treatment. A higher value is associated with a more polarized outcome. The indices are described in section 3.4.2. The regressions specifications are detailed in section 3.6. Error bars reflect 90 percent confidence intervals.

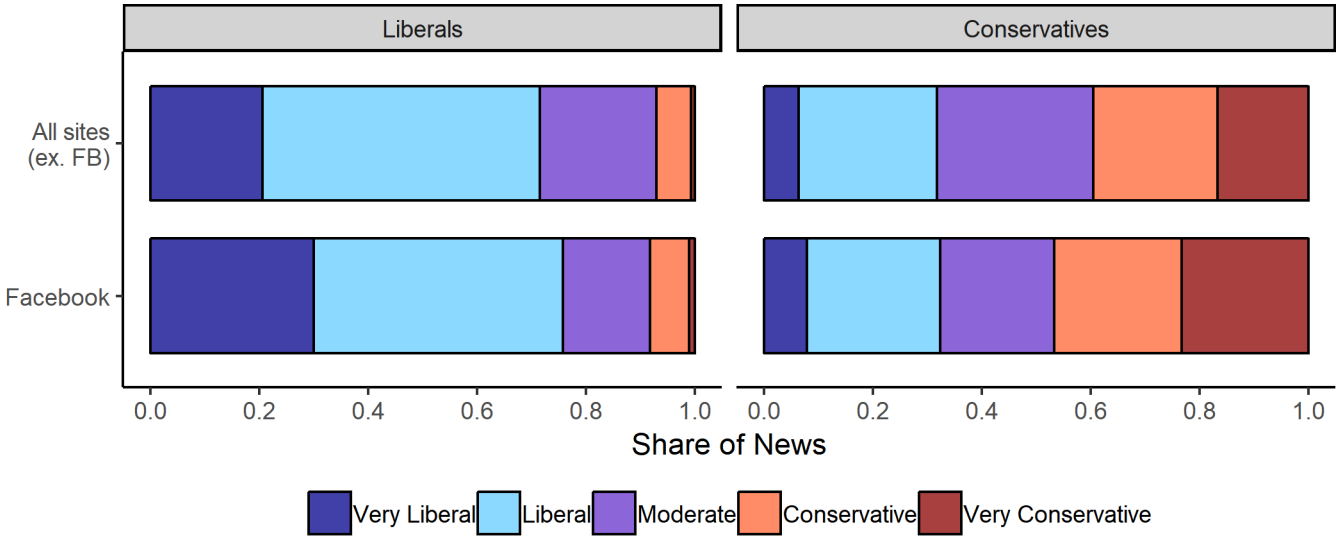
Figure 10: Effect of the Treatment on Affective Polarization - Individual Measures



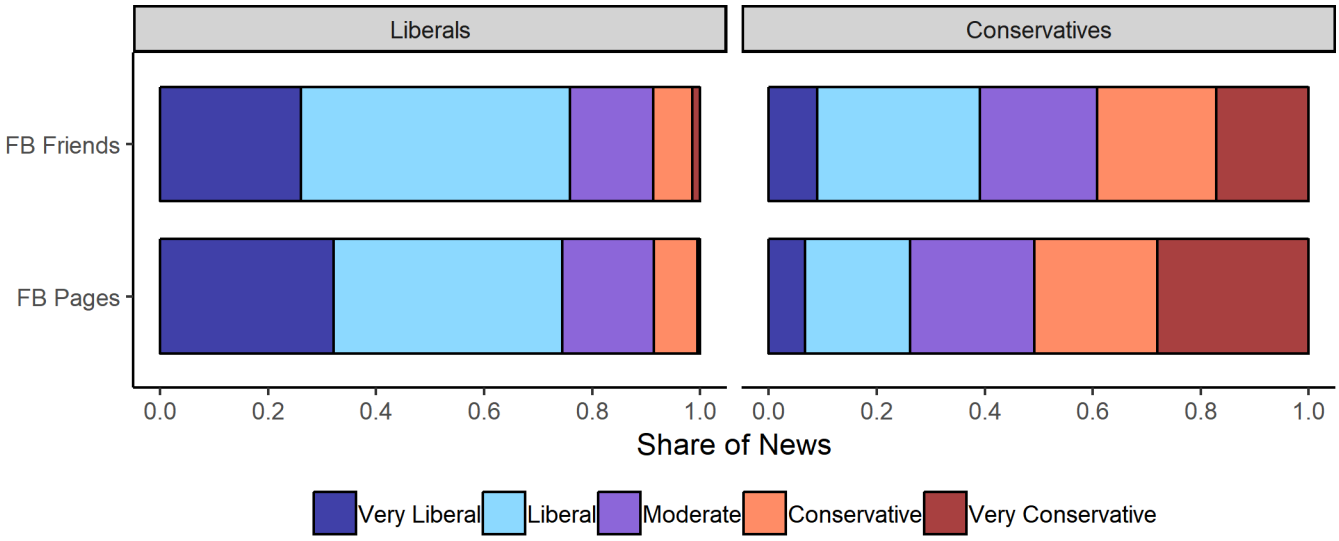
Each row represents the result of a regression estimating the effect of the treatment on the dependent variables composing the affective polarization index.
Difficult Perspective and *Consider Perspective* measure political empathy (Reit et al., 2017). The former is the difference in how difficult it 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. *Feeling Thermometer* is the difference in a feeling thermometer question asking participants how warm they feel toward each party. *Marry Opposing Party* is how participants would feel if their son/daughter married someone from the opposing 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 11: Referral to News Sites, Control Group Extension Data

(a) News Referred to through Facebook and not through Facebook



(b) News Referrals Within Facebook

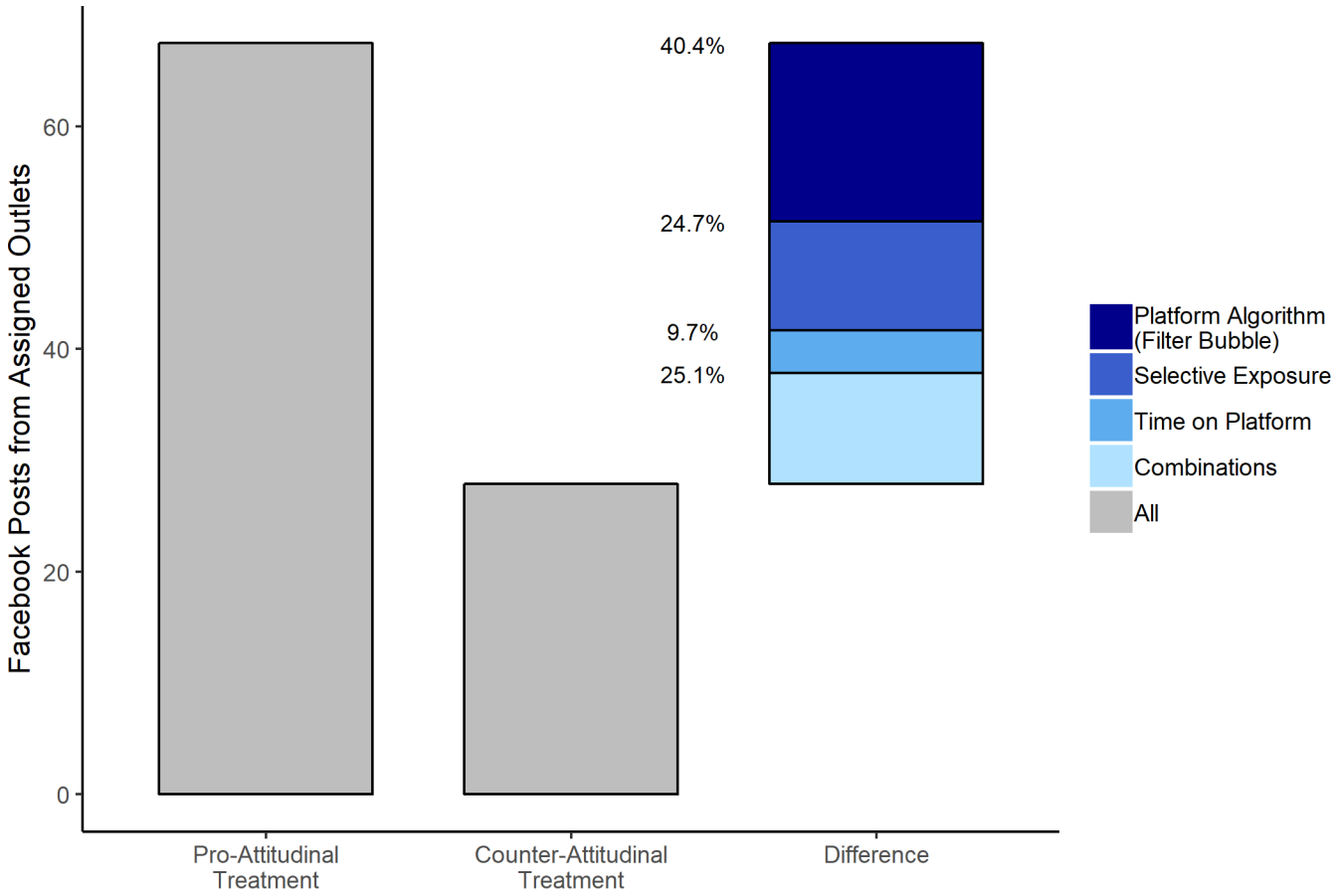


These figures show the mean distribution of news consumed online from leading outlets according to the consumer's ideology. The share of news consumed for each category is first calculated at the individual level and then the mean is calculated for each group. For more details on processing leading outlets see Appendix A.2.

The first sub-figure shows all news consumed when a user clicked any link in 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.

The figure is based on the browser extension data for participants who kept the extension installed for at least two weeks. Only participants 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 not through Facebook are included. To increase power the figure is based on one month of extension data.

Figure 12: The Gap Between the Effect of the Counter-attitudinal and Pro-attitudinal Treatments on Exposure



This figure explains the gap between the number of posts from potential pro-attitudinal outlets participants in the pro-attitudinal treatment were exposed to and the number of posts from potential counter-attitudinal outlets participants in the counter-attitudinal treatment were exposed to. The y-axis is the number of posts seen per day and the x-axis is the treatment. The selective exposure effect measures the effect of individuals subscribing to more pro-attitudinal news. The terms are described in Section 7.2 and the calculations appear in Appendix Table A.9.

Table 1: Consuming News through Facebook and the Match between Individual and Outlet Slant

	Slant		Mean Slant		
	(1)	(2)	(3)	(4)	(5)
Rep. Vote Share	0.405***		0.057***		
	(0.048)		(0.007)		
FB References * Rep. Vote Share	0.459***	0.205***			
	(0.090)	(0.037)			
FB Share of News Ref. * Rep. Vote Share			0.696***	0.292***	
			(0.056)	(0.035)	
FB Share of Visits * Rep. Vote Share					0.405***
					(0.151)
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 effect of ideology and social media on the slant of news consumed. The 2008 Republican vote share is determined by each individual's zip code. In columns (1)-(2) each observation is a website visited, *FB Reference* refers to a visit 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*month. In columns (3)-(5) the dependent variable is the mean slant of all sites visited. *FB Share of News Ref* refers to the share of news sites an individual visited through Facebook and *FB Share of Visits* refers to the share of visits to Facebook among all websites visited.

Table 2: Consuming News through Facebook and the Absolute Slant Consumed

	Absolute Slant		Mean Absolute Slant		
	(1)	(2)	(3)	(4)	(5)
FB References	0.474***	0.335***			
	(0.025)	(0.013)			
FB Share of News Ref.			0.885***	0.496***	
			(0.033)	(0.018)	
FB Share of Visits					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 effect of social media on the absolute slant of news consumed. In columns (1)-(2) each observation is a website visited, *FB Reference* refers to a visit 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*month. In columns (3)-(5) the dependent variable is the mean absolute value of the slant of all sites visited. *FB Share of News Ref* refers to the share of news sites an individual visited through Facebook and *FB Share of Visits* refers to the share of visits to Facebook among all websites visited.

Table 3: Outcomes and Data Collection

Outcome	**Data Source**	**Participants**	**Observations from the four potential conservative outlets and four potential liberal outlets in the two weeks following the intervention**	**Observation from leading outlets in the two weeks following the intervention**
Immediate Compliance - Subscriptions	Facebook	All (37,492 completed baseline survey)	-	-
Exposure - Posts Seen on Facebook Feed	Extension	1,839	106,813	469,304
Browsing Behavior - Websites Visited	Extension	1,839	27,658	149,131
Sharing Behavior	Facebook	34,589	23,452	173,809
Political Opinions and Affective Polarization	Endline Survey	17,634	-	-

This table describes the sources for the main outcome variables analyzed. Extension refers to the Chrome extension a subset of participants installed and Facebook refers to data from Facebook's API for participants who provided access to the data. The third column presents the sample size for each dataset, and the fourth and fifth columns describe the number of observations related to the *potential outlets* and *leading outlets*. The outlets are defined in section 3.2.

Table 4: Balance Table by Assignment to the Liberal and Conservative Treatments

	Mean					Difference		
Variable	All	US	Control	Liberal Treat.	Cons. Treat.	Control - Lib.	Control - Cons.	Cons. - Lib.
Baseline Survey								
Ideology (-3, 3)	-0.61	0.15	-0.61	-0.62	-0.62	0.013	0.012	0.001
Republican	0.17	0.25	0.17	0.17	0.17	-0.006	0.001	-0.007
Independent	0.37	0.29	0.36	0.37	0.37	-0.004	-0.003	-0.000
Democrat	0.38	0.3	0.39	0.38	0.38	0.009	0.003	0.006
Vote Support Clinton	0.53		0.52	0.53	0.53	-0.004	-0.003	-0.001
Vote Support Trump	0.26		0.26	0.26	0.27	0.004	-0.001	0.005
Feeling Therm, Rep.	29.1	43.1	29.2	29.1	28.9	0.126	0.253	-0.127
Feeling Therm, Dem.	47.0	48.7	47.3	46.9	46.8	0.375	0.449	-0.075
Difficult Pers., Rep. (1, 5)	3.13		3.14	3.12	3.14	0.018	0.003	0.015
Difficult Pers., Dem. (1, 5)	2.39		2.39	2.39	2.39	-0.000	0.006	-0.006
Most News Radio	0.09	0.08	0.09	0.09	0.09	0.002	0.001	0.001
Most News Web	0.48	0.21	0.48	0.48	0.47	0.006	0.010	-0.004
Most News TV	0.20	0.53	0.19	0.19	0.20	0.000	-0.008	0.008*
Most News Social	0.18	0.13	0.18	0.18	0.17	-0.002	0.002	-0.004
Device								
Mobile	0.67		0.67	0.68	0.67	-0.011*	-0.000	-0.011*
Facebook								
Female	0.52	0.51	0.52	0.52	0.52	-0.005	-0.002	-0.004
Age	48.5	47.5	48.5	48.3	48.7	0.229	-0.216	0.446*
Total Subscriptions	474		479	474	470	5.419	9.207	-3.788
News Outlets Subscriptions	8.47		8.42	8.50	8.49	-0.081	-0.075	-0.006
News Outlets Slant (-1, 1)	-0.20		-0.20	-0.20	-0.20	0.002	0.001	0.001
Access Posts, Pre-Treatment	0.98		0.98	0.98	0.97	0.001	0.005***	-0.004**
Attrition								
Took Followup Survey	0.47		0.49	0.46	0.46	0.028***	0.028***	-0.001
Access Posts, 2 Weeks	0.92		0.93	0.92	0.92	0.001	0.008**	-0.007**
Ext Install, 2 Weeks	0.05		0.05	0.05	0.05	0.002	-0.002	0.004
N	37,492		12,504	12,495	12,493			
F-Test						1.337	0.985	1.036
P-value						(0.103)	(0.489)	(0.412)

This table presents descriptive statistics by whether participants were assigned to the liberal treatment, conservative treatment or control group. *Ideology* is mean self-reported ideology on a 7 seven point scale. *Republican, Independent and Democrat* are party affiliation, including leaners. *Vote Support* is the share of participants who voted for the candidate or did not vote and supported the candidate. Other options include not supporting any candidate or voting for a third candidate. *Feeling Therm.* is the feeling thermometer score on a 0-100 degree scale. *Difficult Pers.* is whether participants find it difficult to see things from Democrats/Republicans point of view. *Most News* is where participants report getting most of their news. Other options include print and do not know. *Total Subscriptions* is the number of Facebook pages participants subscribed to in baseline. *News Outlets Subscriptions* is the number of subscriptions among leading news outlets. *News Outlets Slant* is the slant of news outlets subscriptions. *Access Posts* is whether participants provided access to the posts they shared. F-test calculated by regressing the treatment on all pre-treatment variables in the table, with missing values replaced with a constant and an indicator for a missing value. The sources for US data are detailed in Table 5. *p<0.1 **p<0.05 ***p<0.01

Table 5: Balance Table by Assignment to the Pro-attitudinal and Counter-attitudinal Treatments

	Mean					Difference		
Variable	All	US	Control	Pro-Att.	Counter-Att.	Control - Pro.	Control - Counter.	Pro. - Counter.
Baseline Survey								
Ideology, Abs. Value (0, 3)	1.75	1.22	1.80	1.80	1.80	0.002	-0.001	-0.003
Republican	0.17	0.25	0.17	0.17	0.18	0.001	-0.006	-0.007
Independent	0.37	0.29	0.35	0.36	0.35	-0.010	0.002	0.012*
Democrat	0.38	0.3	0.40	0.39	0.39	0.009	0.004	-0.005
Vote Support Clinton	0.53		0.54	0.54	0.54	-0.004	-0.003	0.001
Vote Support Trump	0.26		0.27	0.27	0.27	0.001	0.004	0.003
Feeling Therm, Difference	49.9	39.2	50.1	49.8	49.8	0.247	0.235	-0.012
Difficult Pers., Difference	1.91		1.92	1.89	1.91	0.031	0.012	-0.020
Most News Radio	0.09	0.08	0.09	0.09	0.09	0.001	0.004	0.003
Most News Web	0.48	0.21	0.48	0.47	0.48	0.010	0.005	-0.005
Most News TV	0.20	0.53	0.20	0.20	0.20	-0.007	-0.001	0.006
Most News Social	0.18	0.13	0.17	0.17	0.18	0.004	-0.004	-0.008
Device								
Mobile	0.67		0.67	0.68	0.67	-0.010	-0.002	0.007
Facebook								
Female	0.52	0.51	0.52	0.52	0.52	-0.004	-0.006	-0.001
Age	48.5	47.5	48.7	48.8	48.6	-0.061	0.096	0.157
Total Subscriptions	474		476	468	474	7.854	2.743	-5.111
News Outlets Subscriptions	8.47		8.52	8.59	8.59	-0.065	-0.072	-0.007
News Outlets Slant (-1, 1)	-0.20		-0.20	-0.20	-0.21	-0.003	0.006	0.009
Access Posts, Pre-Treatment	0.98		0.98	0.98	0.98	0.002	0.002	-0.000
Attrition								
Took Followup Survey	0.47		0.49	0.46	0.46	0.028***	0.029***	0.001
Access Posts, 2 Weeks	0.92		0.93	0.92	0.92	0.006*	0.002	-0.004
Ext Install, 2 Weeks	0.05		0.05	0.05	0.05	0.002	0.000	-0.002
N	37,492		12,504	12,494	12,494			
F-Test						1.244	0.621	0.986
P-value						(0.182)	(0.933)	(0.484)

This table presents descriptive statistics by whether participants were assigned to the pro-attitudinal treatment, counter-attitudinal treatment or control group. *Abs Ideology* is the absolute value of self-reported ideology. *Feeling Therm, Difference* is the difference between the feeling toward the participants 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. For all other variables see Table 4.
US data for *Female* and *Age* based on Census 2017 Population Estimates, data for *Party, Feeling Thermometer* and *Ideology* based on 2016 American National Election Survey, data for *Most News* based on Pew Research Center American Trends Wave 23.

Table 6: Effect of the Treatment on the Social Media Feed on News Sites Visited

	Slant of News Sites	Slant of News Sites Visited through FB
	(1)	(2)
Slant of FB Feed	0.314***	0.717***
	(0.069)	(0.087)
Controls	X	X
F Stat	61.21	71.94
Observations	1,525	1,204

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

Table 7: 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.	-0.356^{***}	
	(0.051)	
FB Congruence Scale, Std. Dev.		0.370^{***}
		(0.052)
Data	Control Group	Control Group
Observations	353	361

(b) Causal Effect Based on Experimental Variation

	IV	
	Affective Polarization	
	(1)	(2)
FB Counter-Att. Share, Std. Dev.	-0.128^{*}	
	(0.066)	
FB Congruence Scale, Std. Dev.		0.115^{*}
		(0.059)
Controls	X	X
First Stage F	62.36	58.94
Share of Correlation in Control Group	0.36	0.31
Observations	1,071	1,088

These tables measure the effect of exposure to pro and counter-attitudinal news on affective polarization. The tables use two summary-statistics. *Counter-Att. Share* is the share of counter-attitudinal news the participant was exposed to in Facebook, calculated as the share of news from counter-attitudinal outlets among all pro-attitudinal and counter-attitudinal outlets. The *Congruence Scale* is the mean slant of all news exposed to in 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 dependent variable is the affective polarization index, the independent variables are the two summary statistics and the instrument is the treatment. The regressions control for the covariates specified in section 3.6. The row titled *Share of Correlation in Control Group* divides the causal effect found with the correlation in the control group.
*p<0.1 **p<0.05 ***p<0.01

Table 8: Effect of the Treatment on Opinions toward Own vs Opposing Party

	Opinion Own Party	Opinion Opposing Party
	(1)	(2)
Pro-Att. Treatment	0.010	−0.003
	(0.013)	(0.014)
Counter-Att. Treatment	0.002	0.031**
	(0.014)	(0.013)
Pro - Counter	0.008	-0.034**
	(0.014)	(0.014)
Observations	16,894	16,894

This table presents the effect of the pro-attitudinal treatment and counter-attitudinal treatment 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 supported) 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 consumers were only asked how they would feel if their son/daughter married someone from the opposing party. The controls and specified in section 3.6. *p<0.1 **p<0.05 ***p<0.01

Appendix

A Data Collection and Processing

A.1 Leading News Outlets

Throughout the paper, I analyze participants' engagement with leading outlets. The list of outlets and their slant are based on a dataset constructed by Bakshy et al. (2015). The authors use Facebook's internal data and classify links to hard and soft news based on the words in the link. The alignment of each website is determined according to the self-reported ideology of Facebook users who share hard-content links from the website.

Before using the dataset I removed the following popular domains which are not related to news outlets: Amazon, The White House, Twitter, Vimeo, Wikipedia, and YouTube. I then remove the webreference ("www.") from all outlets' websites, so all outlets only contain the domain used. I then merge websites which have similar URLs 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. After processing the data, the list of leading outlets contains 488 websites. Similarly to Bakshy et al. (2015), I define very liberal outlets as outlets in the bottom quantile of the distribution of news slant outlets, liberal outlets as outlets in the second quantile, moderate outlets as outlets in the third quantile, conservative outlets as outlets in the fourth quantile and very conservative outlets are defined as outlets in the top quintile of the news slant distribution.

I match Facebook pages with the list of leading outlets by searching for all Facebook pages with names similar to the outlet's domain and manually checking the three leading pages. Overall, Facebook pages were found for 374 out of 488 websites included from the Bakshy et al. (2015) dataset.

A.2 Surveys

A.2.1 Baseline Survey

The baseline survey took place from early February to mid-March 2018. Participants were recruited to the survey using Facebook ads. The 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 for people who follow politics (see Figure A.1). 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 Chrome 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.[42]

40,514 participants took the survey and reached the screen where the intervention occurs. Of those participants, 37,492 are included in the final sample. Participants are excluded from the final sample for the following reasons: missing information on outlets the participant subscribes to either because the participant did not provide permissions to access that data or since the data was not collected properly in real time (2.38%); participant already subscribed to too many of the outlets such that it was not possible to offer the participant four new liberal and four new conservative outlets (3.64%); technical issues with the Qualtrics survey which prevented some data from being collected (0.28%); or taking the survey a second time (0.01%). Finally, to ensure quality responses, participants who were likely to respond carelessly were also excluded. These include participants who completed all survey section until the intervention exceptionally quickly (in under three minutes where the median time was eleven minutes) and participants who did not answer at least half of the closed-ended, non-required questions (0.03%). The criteria determining whether to exclude a participant are all based on survey data submitted before the intervention occurs.

A.2.2 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.[43] To match endline survey responses with baseline survey responses, participants who began the survey were asked to log in to the endline survey through Facebook or supply an email address. Endline responses are matched with baseline 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 and last name if the combination is unique, and combination of first and last name if the combination is unique. 98.73% of the individuals who took the followup survey were matched.

19,693 participants began the endline survey and were matched with valid baseline survey responses. Responsdents are included in the sample 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 responses that matched to valid endline 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).

[42] To test whether many participants clicked the ad since someone shared it with them, I provided participants with a slightly modified link to the survey after the survey was completed, 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 the ad could also share the ad or the link that appears in the ad with other individuals, and in such cases it is difficult to distinguish participants who reached the survey through an ad and those who reached it after the ad was shared with them. However, approximately 95% of exposures to the recruitment ad 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 users entered the survey since a sponsored ad appeared in their feed.

[43] A small share of participants were recruited through an invitation in the Chrome extension or through a Facebook notification.

A.3 Facebook Data on Subscriptions and Posts Shared

I collect data on outlets participants subscribed to and posts they shared. Only posts shared by participants with their social networks of the following types are included in the analysis: status update, video, link or note (photos, albums, events, and music were excluded). I exclude posts not matched with a news outlet in the analysis. Posts are matched with outlets based on either the Facebook page sharing the post or the domain of a link contained in the post. I match Facebook posts to leading outlets based on the Facebook page which shared the posts and based on links contained in the post.

Matching links is based on the following steps:

1. For each outlet included in the experiment, a list of relevant domains is created. The list is created 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 "huffingtonpost.ca", "huffpost.com" and other similar domains. For outlets not included in the experiment, I define the main domain in the Bakshy et al. (2015) dataset as the only relevant domain.

2. Links from posts cannot be immediately matched with the list of domains since they may refer to a short alias, created by URL-shortening services such as tinyurl.com, which refers to the full long URL. Therefore, each URL is first converted to the final redirected URL before being matched to the list of domains.

A.4 Extension Data on Facebook Feed and Browsing Behavior

Data on news-related posts appearing in participant's Facebook feed was matched to outlets using the same method explained in the previous section. News sites visited were matched to outlets based on the list of domains defined in the previous section. 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 for a second time within a 20-minute window, the second visit is excluded.

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.

B Additional Details on Empirical Strategy

B.1 Controls

To increase power when estimating the effect on political opinion and affective polarization, I control for a set of pre-registered covariates. Appendix Table A.5 shows that the results regarding both political opinions and affective polarization are robust to excluding all covariates. I control for self-reported

ideology, party affiliation, approval of President Trump, ideological leaning, age, age squared, gender and baseline question on political beliefs similar to the questions used in the endline survey.[44]

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 defined in Section 3.6.2.

The following baseline questions are used as controls when estimating the effect on political opinions: feeling toward President Trump on the thermometer scale (0-100 continuous scale), does the participant personally worry about illegal immigration (nominal variable with four options ranging from "not at all" to "great deal"), is Mueller conducting a fair investigation (nominal variable with three options: yes, no and do not know), has Trump attempted to obstruct the investigation into Russian interference in the election (nominal variable with three options: yes, no and do not know). The following baseline questions are used when estimating the effect on affective polarization: baseline value of the *feeling thermometer* measure and baseline value of the *difficult perspective* measure.

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 list of potential outlets is controlled for to prevent a possible omitted variable bias since individuals who were assigned the alternative outlet may have different characteristics than individuals who were assigned the primary outlets.

C Pre-Analysis Plan

The main outcome and hypothesis tested in this study were pre-registered in the AEA RCT Registry.[45] The analysis deviates from the pre-analysis plan in two important ways. First, in the plan I state that the weights for the index variables will be determined by the inverse of the covariance between variables at endline (Anderson, 2008). However, the method generates negative weights. When using negative weights the interpretation of the index is no longer 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. Therefore, the indexes are created using equal weights instead.

Appendix Table A.11 repeats the analysis with the inverse-covariance weighting for affective polarization and political opinions. Column (2) of Appendix Table A.11b shows that the effect of the counter-

[44] Ideological leaning was not explicitly mentioned as a control variable in the pre-analysis plan. This covariate was added 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.

[45] AEA RCT Registry Trial 0002713

attitudinal treatment on affective polarization, compared to the pro-attitudinal treatment, remains almost exactly the same when using inverse-covariance weights. One challenge with inverse-covariance weights is that this method does not cleanly generate weights for individuals with missing outcomes. In column (3), each outcome variable is first scaled, weights are then created using the inverse-covariance method based on participants with no missing outcomes, and the weights are 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.11a shows the results of a similar analysis with the political opinion index. Since the inverse-covariance method generates negative weights, for some of the outcomes, columns (4) and (5) repeat the analysis after replacing all negative weights with zero and renomarlizing the weights 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 affective polarization index originally included five attitudinal measures and three behavioral measures, while only the attitudinal measures are analyzed in this study. 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 their posts. Furthermore, the behavioral measure turned out not to measure polarization well. While a measure of affective polarization should typically be correlated with partisanship, there was almost no correlation between being partisan and the behavioral outcomes.[46]

Column (1) of Appendix Table A.12 shows that the primary estimate is still significant when using all eight variables in the affective polarization index. Column (3) measures the effect only on the three behavioral outcomes. The effect of the treatments is almost exactly zero. 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 individuals' willingness to share a post regarding the importance of seeking both sides of a story.

D Additional Analysis

D.1 Heterogeneous Effects

I test for heterogeneous effects based on the following covariates:

[46]The correlation for the behavioral polarization measure was between 0.03 and 0.06, while the correlation for the affective polarization measures was between 0.17 and 0.36

- Baseline News Subscriptions - baseline subscription to news outlets on Facebook is above or equals median
- Seen Outlet - self-reported exposure to posts from the eight potential outlets is above or equals median in baseline
- Know Outlet Slant - the distance between perceived slant of potential outlets and average perceived slant by participants with the same self-reported ideology is below median
- Politics Ad - recruitment ad mentioned that the survey discusses politics
- High CRT - answer to at least one the following questions is correct: "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". This covariate is based on the Cognitive reflection test (Shane, 2005)
- Older - age above or equals median
- Most news social media - answer to "Thinking specifically about government and politics, do you get most of your news about this topic..." is "Through social networking sites (such as Facebook or Twitter)"
- Follow news - answer to "How often do you pay attention to what's going on in government and politics?" is above median
- Ideological - absolute value of ideology on the 7 point scale (from -3 for very liberal to +3 for very conservative) is above or equals median
- Conservative - participant's ideological leaning is conservative
- Certain - answer to "Generally speaking, how certain are you of your political opinions?" is above or equals median
- Echo chamber - 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
- Personality: Openness - answer to index of the following two questions is above or equals median: "I see myself as open to new experiences, complex" and reverse values of "I see myself as conventional, uncreative" (based on Gosling et al. 2003).

Panel 1 of Appendix Figure A.6 shows that while the effect on political opinions is generally homogeneous (i.e. most people were not persuaded by the treatment), there are several exceptions. Participants who self-report following the news more frequently were more likely to be affected by the treatment. In addition, conservatives were slightly more likely to be persuaded by the treatment (while not statistically

significant, the effect can also be seen in Appendix Figure A.5). The stronger effect found for conservative complements other studies finding that conservatives are more likely to be affected when exposed to new information (Bail et al., 2018; Day et al., 2014), and is in line with anecdotal evidence that fake news creators target conservatives since they are more likely to consume the articles.[47] Finally, somewhat surprisingly, participants who scale higher on the openness index were less likely to be persuaded by the interventions.

Panel 2 of Appendix Figure A.6 shows that there is no evidence for substantial heterogeneous effects of the treatment on affective polarization.

D.2 Knowledge

While the 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, 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 (e.g., hearing about the Stephon Clark shooting) and for other measures, the conservative treatment was expected to have an effect (e.g. hearing about a controversial statement by Hillary Clinton).

Table A.13 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.

The extension data was used to test whether participants were exposed to differential coverage on topics included in the self-reported familiarity index. Table A.14 shows that the intervention affected news exposure. The regression measures the effect of the treatment on the number of posts mentioning each topic which appeared in the participants' social media feeds.[48]

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 two of the four topics, the effect is also significant when the treatments are compared to each other. This suggests that while the slant of one's social media news feed can determine what news events an individual is exposed to, the feed does not necessarily affect their political awareness of topics, probably since individuals consume news through other means as well.

[47] Sydell, L. - We Tracked Down A Fake-News Creator In The Suburbs. Here's What We Learned. All Things Considered, NPR. November 23, 2016

[48] The number of posts discussing Michael Cohen, Louis Farrakhan or the shooting of Stephon Clark are measured by counting the number of posts mentioning the expression "Michael Cohen", "Louis Farrakhan" and "Stephon Clark", respectively. The fourth outcome is Hillary Clinton's speech in India suggesting that many white women voted for Trump since they took their voting cues from their husbands. To measure mentions of the speech I count the number of posts mentioning the words Clinton, vote and either India or husband.

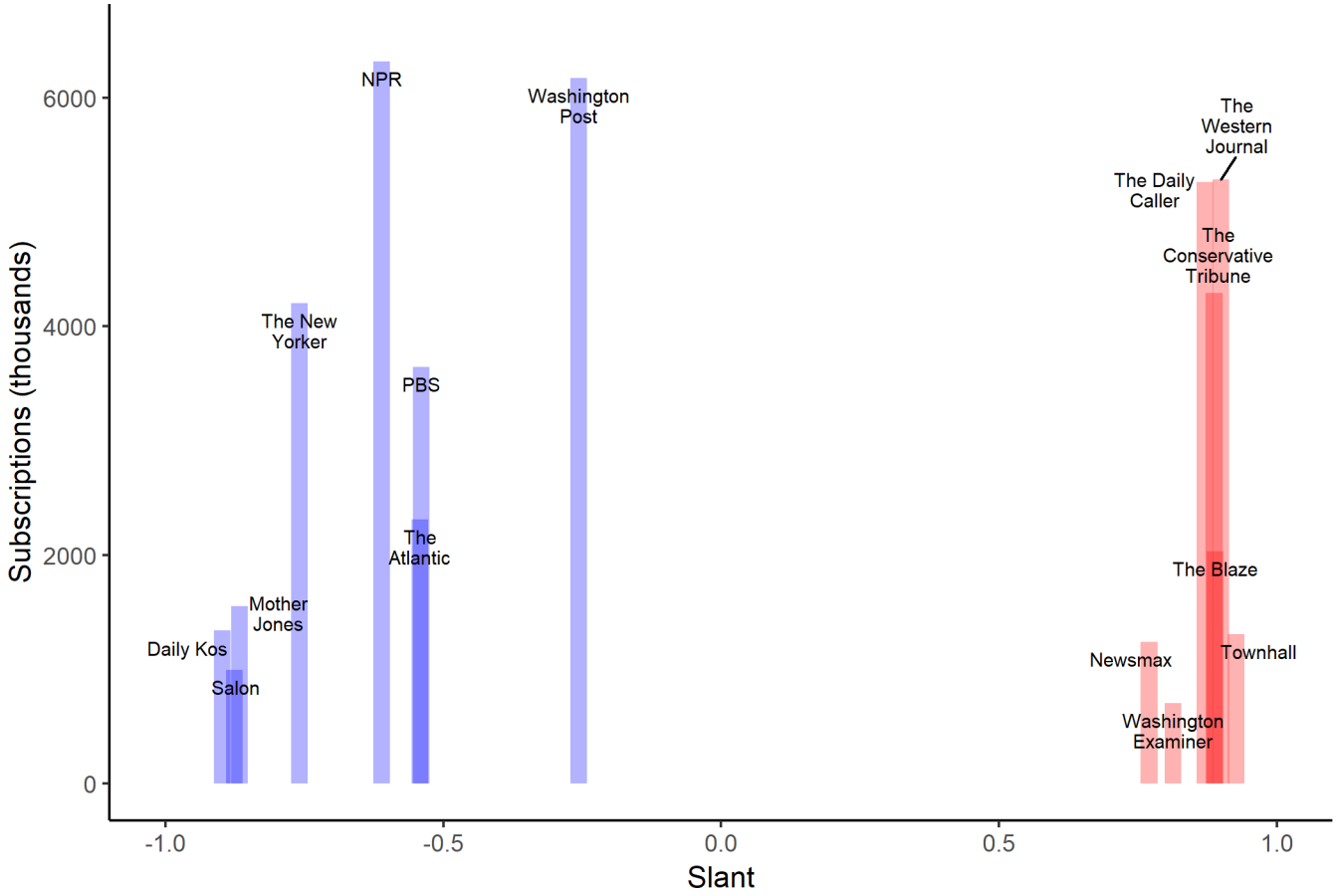
E Additional Figures and Tables

Figure A.1: Recruitment Ads

(a) Political Ad

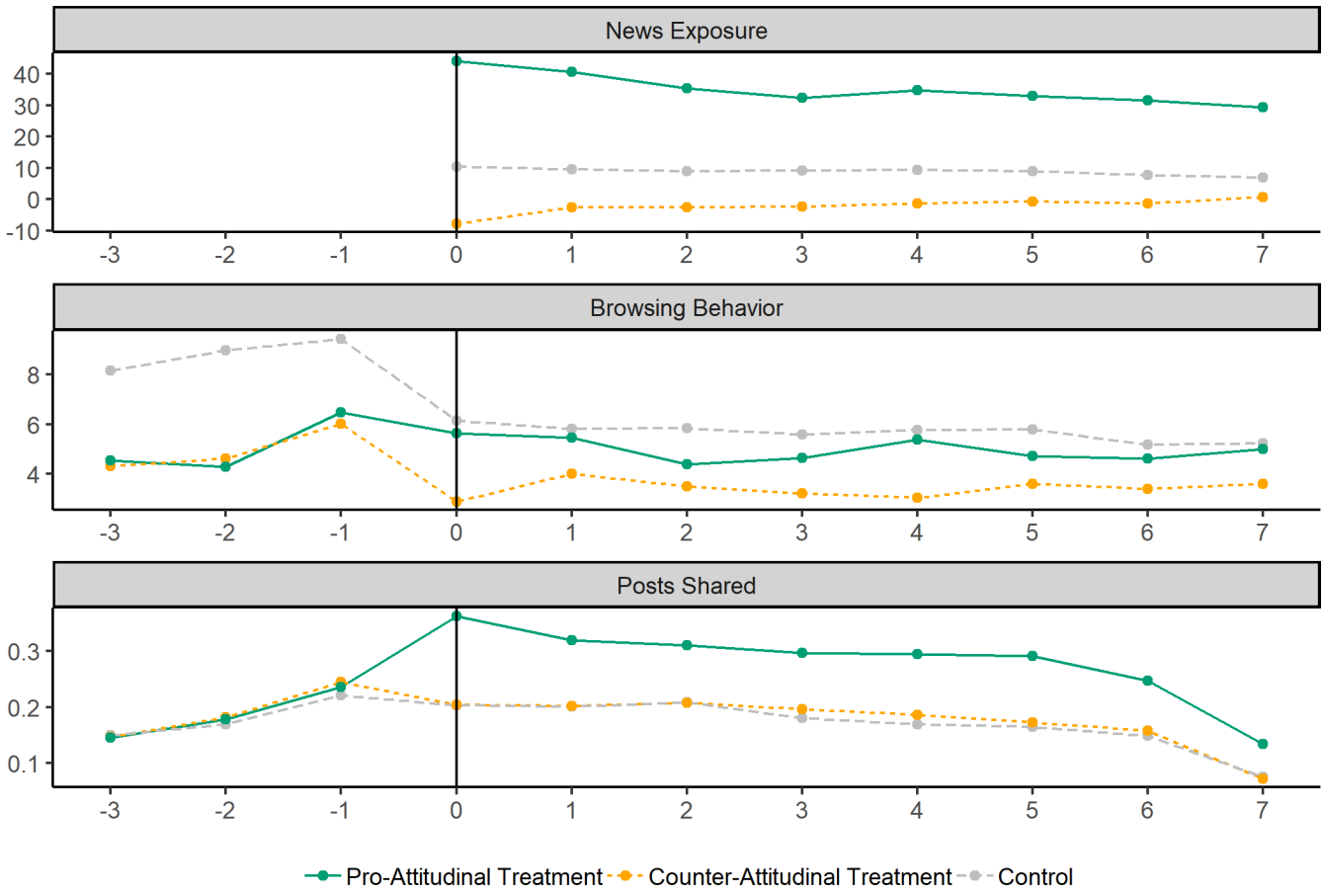
(b) General Ad

Figure A.2: Alternative Assigned Outlets



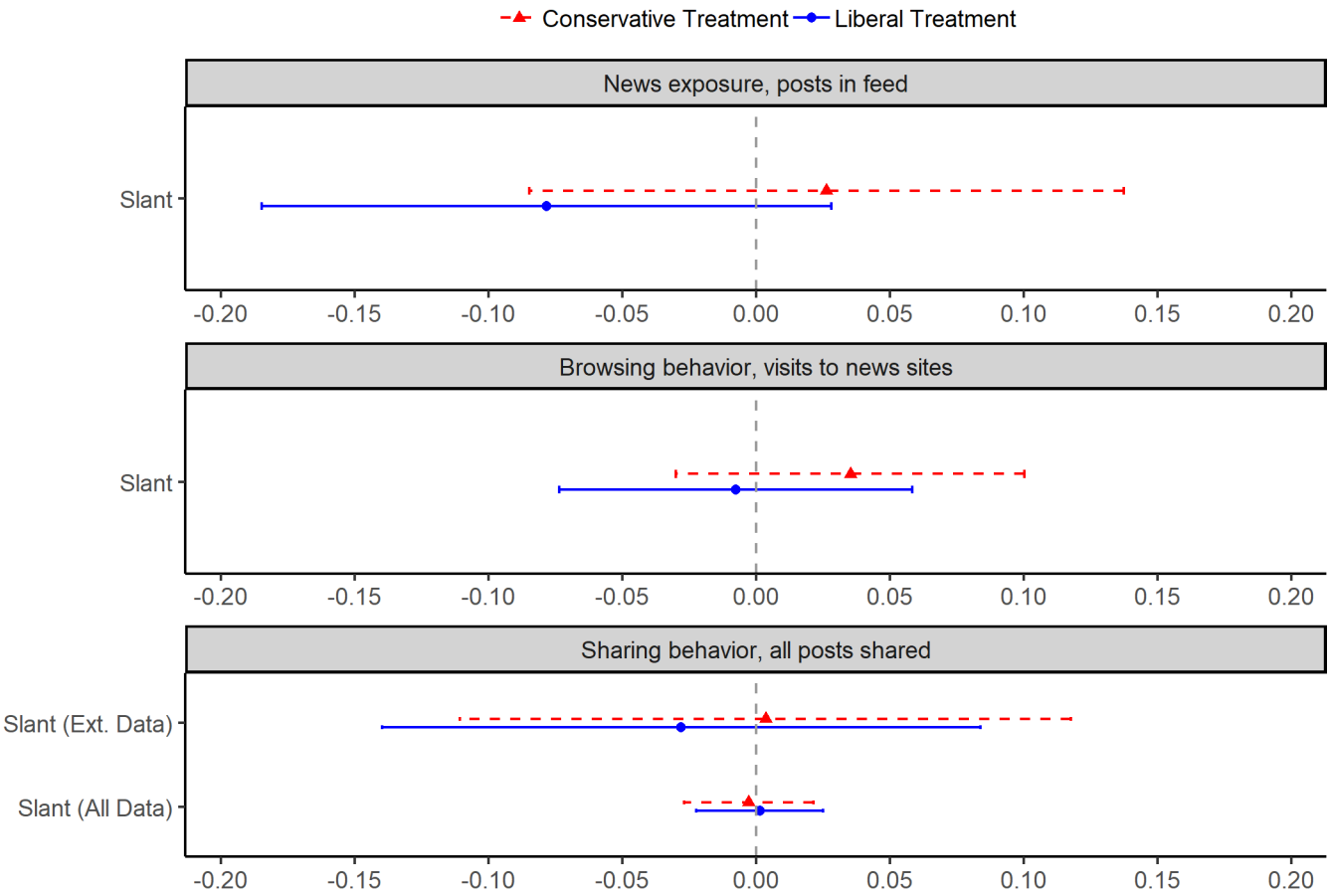
This figure displays the alternative outlets offered in the experiment if a participant already subscribed to one of the primary outlets. The x-axis is the slant of the outlets, as determined by Bakshy et al. (2015), and the y-axis is the total number of individuals who have subscribed to each outlet in April, 2018.

Figure A.3: Effect of the Treatment on the Difference between the assigned Pro-Attitudinal and Counter-Attitudinal Outlets Over Time



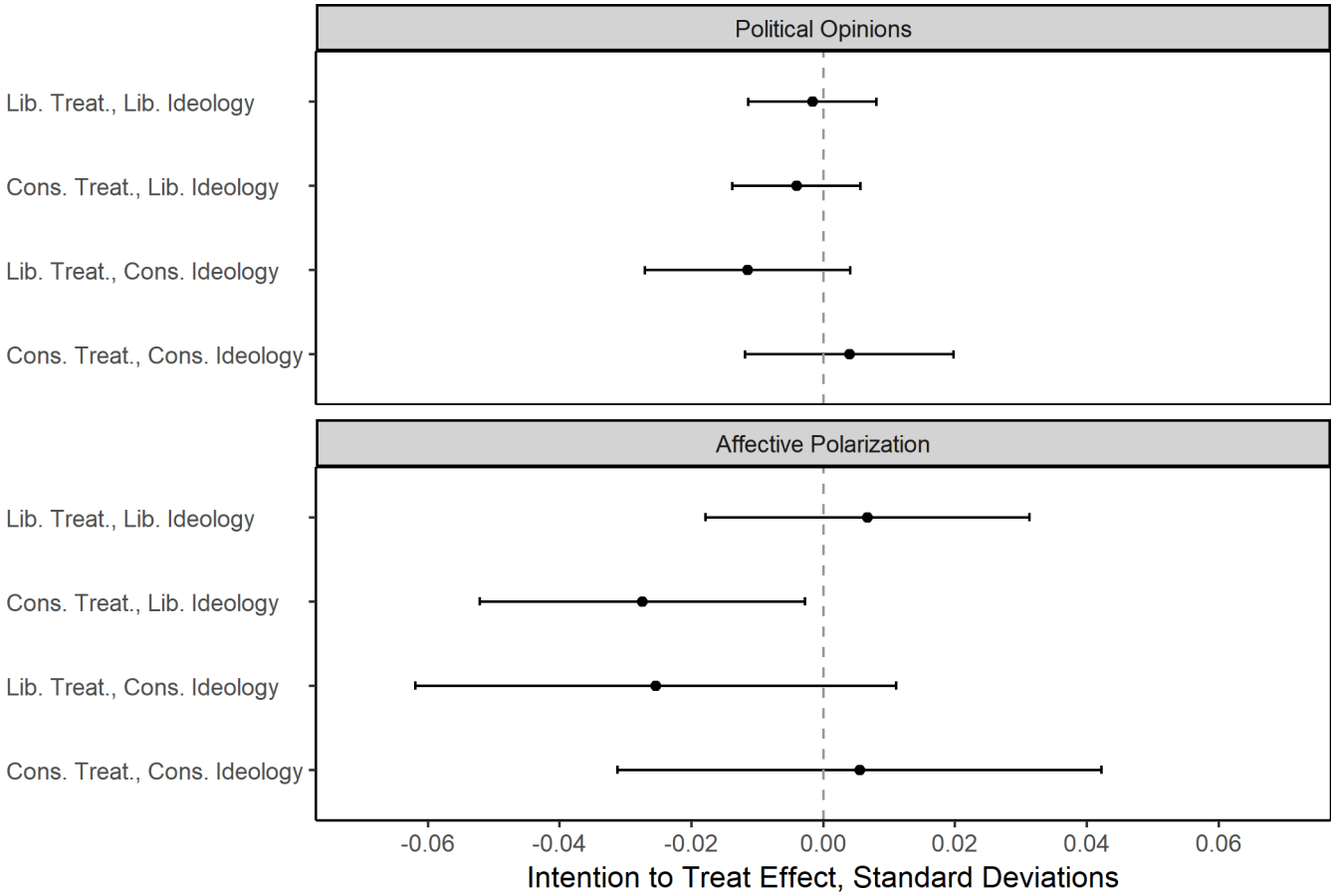
This figure shows the affect of the treatment over time. Each panel shows the difference between the pro-attitudinal and counter-attitudinal outlet.
The first panel shows the difference in the number of posts observed in the participants Facebook feed by treatment. The panel does not include data from the pre-period since I do not observe the Facebook feed before the intervention. The second panel shows the difference in the number of visits to the assigned pro-attitudinal and counter-attitudinal websites by treatment, and the third panel shows the difference in the number of posts shared by treatment.
The first two panels are based on participants who installed the extension for at least eight weeks and third panel is based on participants who provided access to their shared posts for at least eight weeks.

Figure A.4: Effect of the Treatment on Slant of News Consumption, Excluding the Potential Outlets of Each Individual



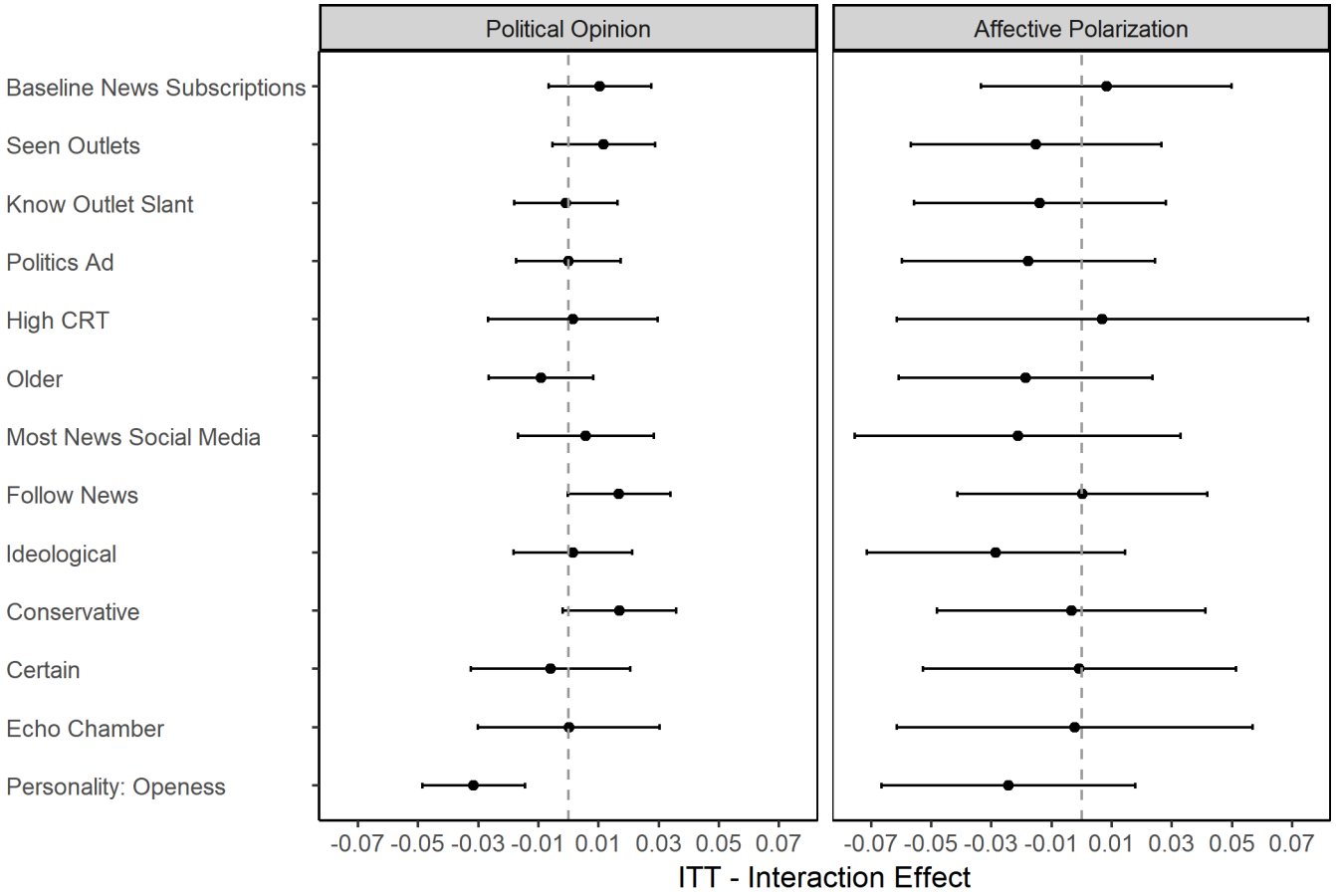
This figure shows the effect of the liberal and conservative treatments on the mean slant, in standard deviations, of all news individuals engaged with, excluding the four potential liberal outlets and the four potential conservative outlets defined for each participant. Each row in the figure is estimated 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 first three outcomes include participants which installed the browser extension for at least two weeks, and the last outcome includes a larger sample of all participants who provided permissions to access their posts for at least two weeks. Error bars reflect 90 percent confidence intervals.

Figure A.5: Effect of the Treatment on Political Opinions, by Baseline Ideology



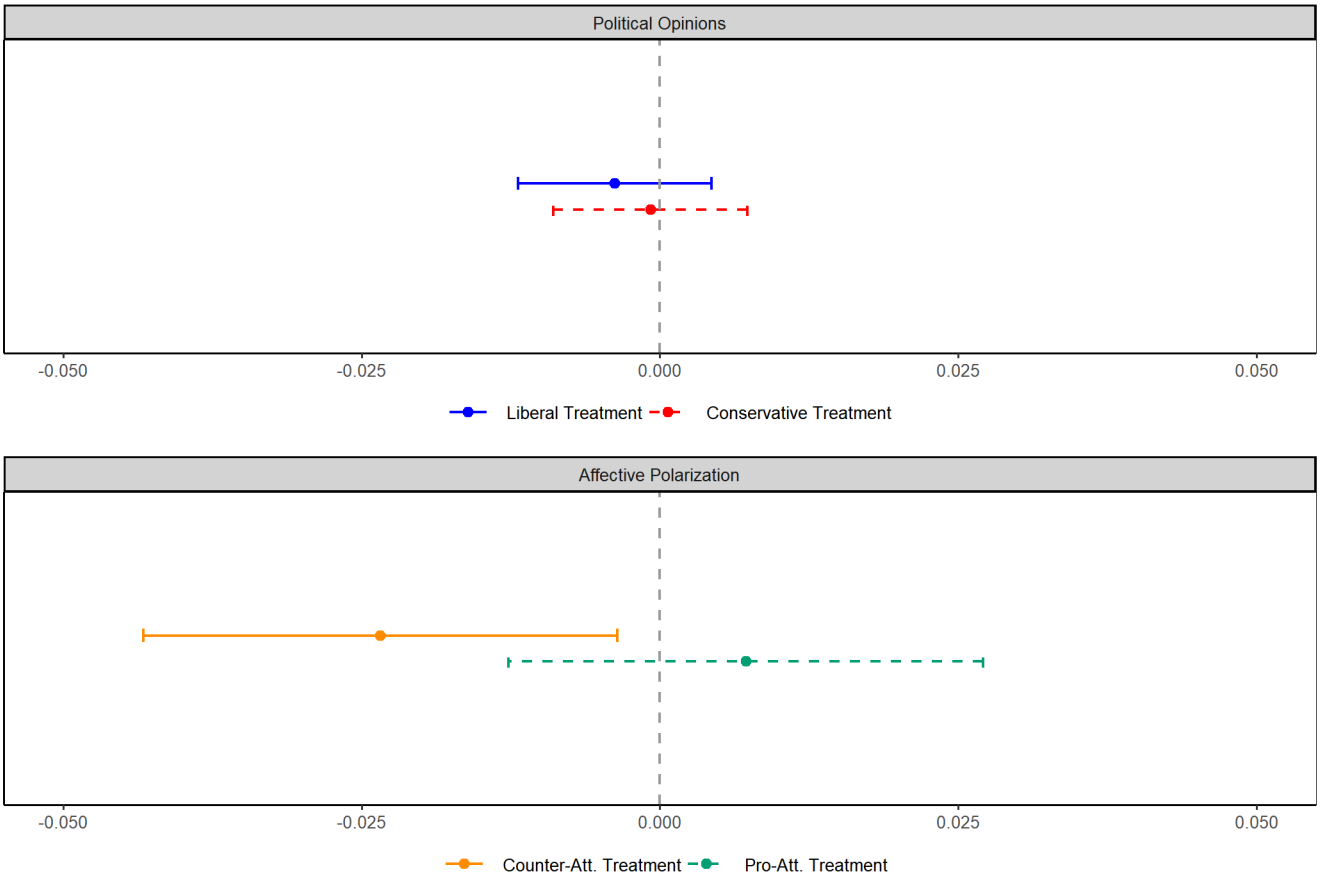
This figure shows the effect of the treatment among ideological subgroups based on the following model:
$Y_i = \beta_1 T_i^L I_i^L + \beta_2 T_i^L I_i^C + \beta_3 T_i^C I_i^L + \beta_4 T_i^C I_i^C + I_i + \alpha X_i + \varepsilon_i$
where: T_i^C, T_i^L are binary indicators for the conservative and liberal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

Figure A.6: Heterogeneous Effects of Treatment on Political Beliefs



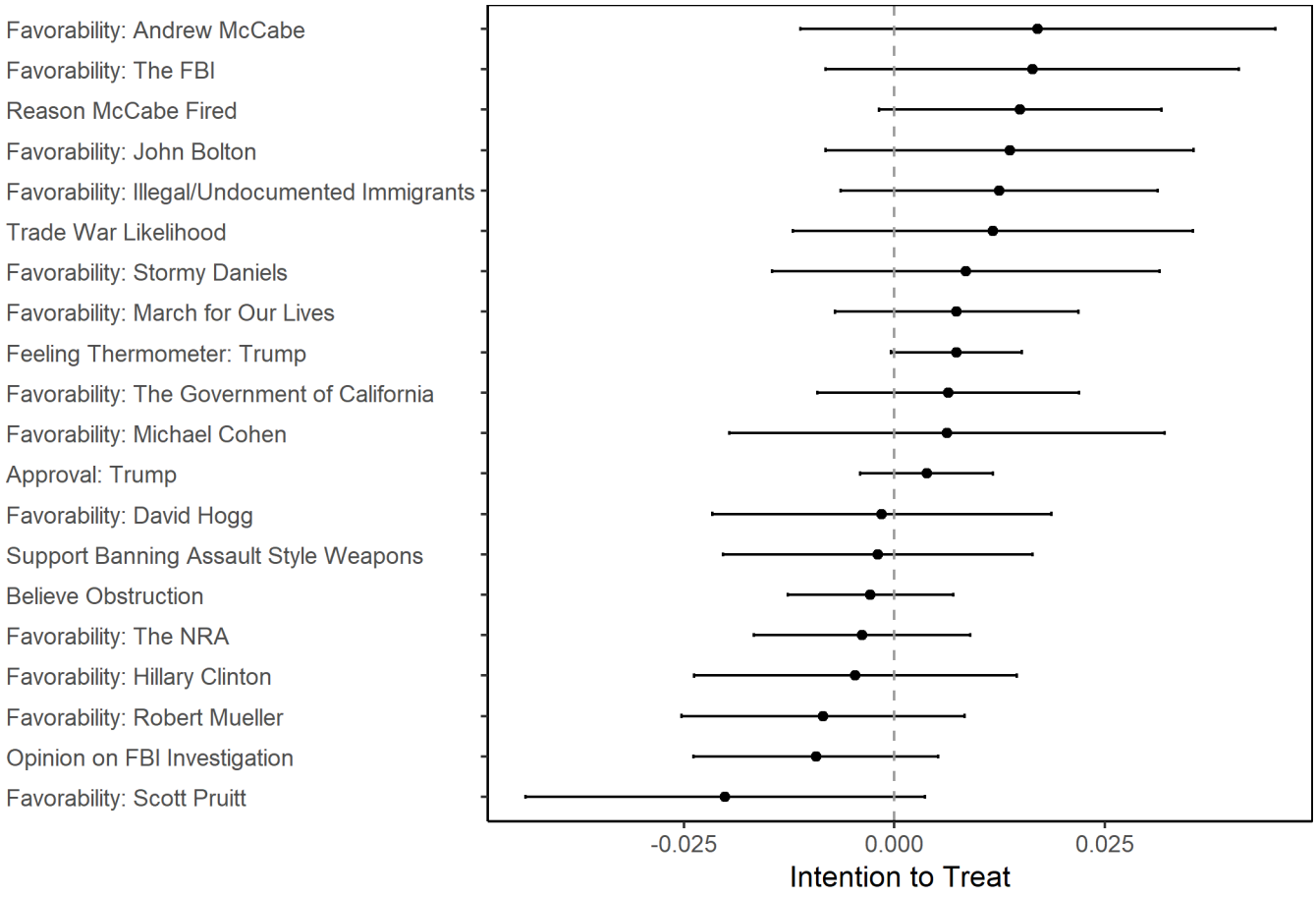
In the Political Opinion Panel each row represents the β_2 coefficient in the following separate regression:
$Y_i = \beta_1 T_i^C + \beta_2 T_i^C \times Var + \beta_3 Var + \alpha X_i + \varepsilon_i$,
where the dependent variable is the political opinion index, the independent variable is the full interaction of the conservative treatment and the variable analyzed in the row and 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 Affective Polarization Panel each row presents the β_2 coefficient in the following regression:
$Y_i = \beta_1 T_i^P + \beta_2 T_i^P \times Var + \beta_3 Var + \alpha X_i + \varepsilon_i$,
where the dependent variable is the affective polarization index, the independent variable is the full interaction of the pro-attitudinal treatment and the variable analyzed in the row and and 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. Error bars reflect 90 percent confidence intervals.

Figure A.7: Effect of the Treatment by Treatment Arm



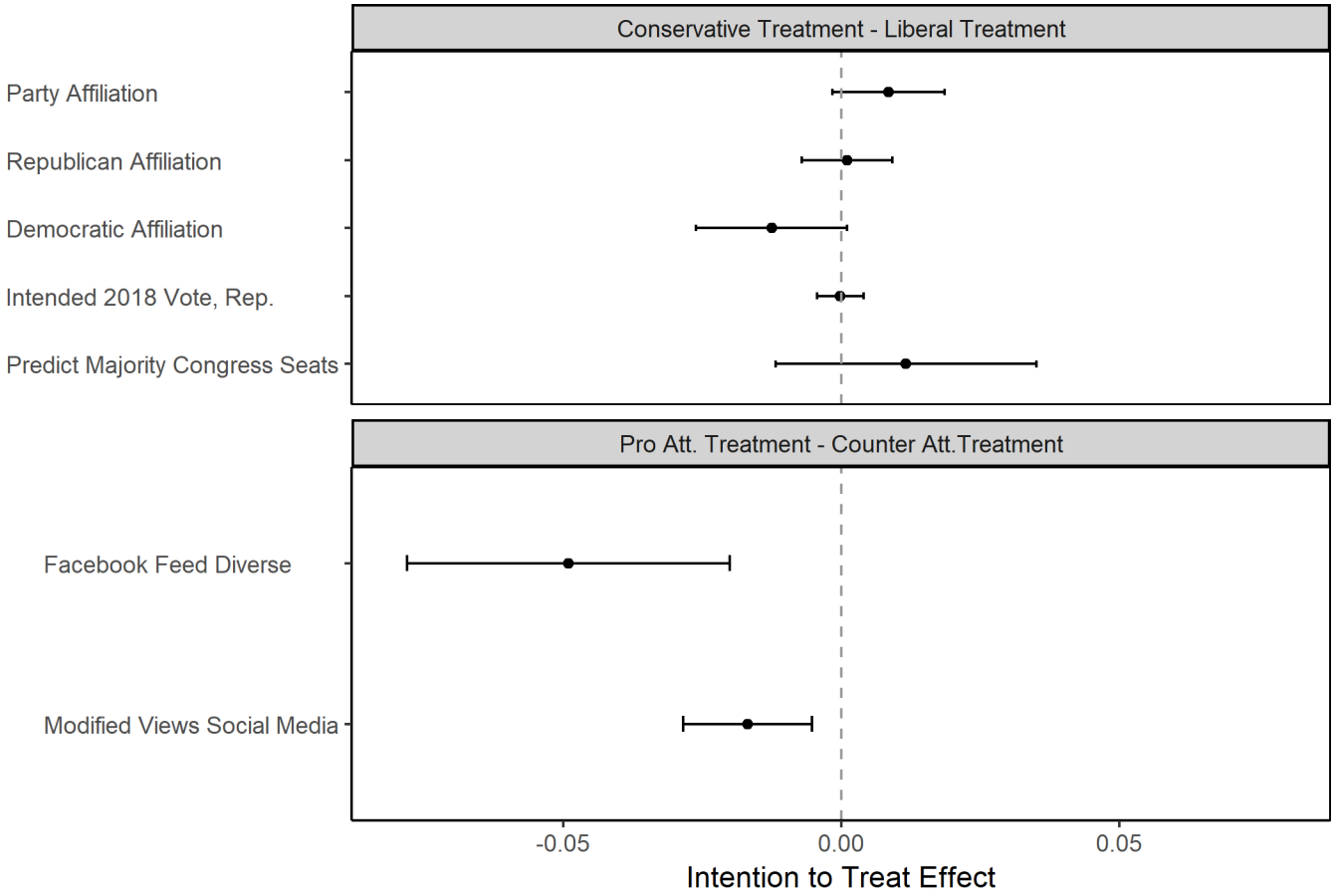
This figure shows the effect of each treatment arm on the political opinions index and affective polarization index. The indices are described in section 3.4.2. The specification and controls are described in more detail in Section 3.6. Error bars reflect 90 percent confidence intervals.

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



This figure shows the effect of the conservative treatment, compared to the liberal treatment on each outcome composing the political opinions index. Each row represents a separate regression as specified in Section 3.6. All the dependent variables have been standardized by subtracting each variable by the control group mean and dividing by the control group's standard deviation. Each variable is defined such that a higher value is associated with a more conservative outcome. The *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 how respondents feel towards 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). An answer of only improper action is coded as one, improper action along with damaging credibility or revenge is coded as 0, and an answer of only damaging credibility or revenge is coded as -1. *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.

Figure A.9: 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. *Republican/Democrat Affiliation* is coded as 1 if the participant lean 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, Rep.* is whether the participant would have voted for The Republican Party candidate in their district, if the election was held today, among participants intending to vote for the Republican or Democratic Party candidate. *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).
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. *Facebook Feed Diverse* is the answer to the question "Thinking about the opinions you see about government and politics on Facebook, how often are they in line with your own views?", where the possible answers are "Always or nearly all the time" (coded as 0), "Most of the time" (1), "Some of the time" (2), "Not too often" (3).
Non binary outcomes are standardize 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.

Table A.1: List of Outlets Offered and Subscriptions

Outlet	Group	Offered	Subscribed	Share
The Washington Times	Conservative	12365	3166	0.26
The National Review	Conservative	12056	2868	0.24
The Wall Street Journal	Conservative	11805	3914	0.33
Fox News	Conservative	10841	1387	0.13
The Daily Caller	Conservative	1470	309	0.21
The Western Journal	Conservative	509	146	0.29
Washington Examiner	Conservative	607	131	0.22
Townhall	Conservative	135	37	0.27
The Conservative Tribune	Conservative	72	31	0.43
The Blaze	Conservative	80	24	0.30
Newsmax	Conservative	32	13	0.41
Slate	Liberal	11737	2938	0.25
MSNBC	Liberal	11687	2716	0.23
HuffPost	Liberal	10642	2304	0.22
The New York Times	Liberal	10143	3285	0.32
Washington Post	Liberal	2825	1296	0.46
Salon	Liberal	1668	572	0.34
Daily Kos	Liberal	661	222	0.34
NPR	Liberal	119	66	0.55
Mother Jones	Liberal	150	58	0.39
The Atlantic	Liberal	203	111	0.55
The New Yorker	Liberal	105	65	0.62
PBS	Liberal	40	23	0.57

This table shows the list of all outlets included in the experiment. *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.

Table A.2: Balance Table by Assignment to the Liberal and Conservative Treatments, Among Participants who Completed the Follow-up Survey

	Mean					Difference		
Variable	All	US	Control	Liberal Treat.	Cons. Treat.	Control - Lib.	Control - Cons.	Cons. - Lib.
Baseline Survey								
Ideology (-3, 3)	-0.70	0.15	-0.71	-0.70	-0.69	-0.010	-0.023	0.013
Republican	0.16	0.25	0.16	0.16	0.16	0.001	0.001	0.001
Independent	0.36	0.29	0.35	0.37	0.36	-0.016*	-0.007	-0.009
Democrat	0.40	0.3	0.41	0.39	0.40	0.014	0.008	0.006
Vote Support Clinton	0.55		0.55	0.55	0.55	-0.002	-0.000	-0.002
Vote Support Trump	0.25		0.25	0.24	0.25	0.009	-0.002	0.012
Feeling Therm, Rep.	27.5	43.1	27.6	27.4	27.6	0.224	-0.051	0.275
Feeling Therm, Dem.	47.8	48.7	48.1	47.7	47.5	0.400	0.684	-0.284
Difficult Pers., Rep. (1, 5)	3.18		3.20	3.16	3.19	0.040	0.005	0.035
Difficult Pers., Dem. (1, 5)	2.35		2.34	2.35	2.37	-0.006	-0.034	0.028
Most News Radio	0.09	0.08	0.09	0.09	0.09	-0.002	-0.001	-0.001
Most News Web	0.49	0.21	0.50	0.49	0.48	0.012	0.019**	-0.007
Most News TV	0.19	0.53	0.18	0.18	0.20	0.007	-0.015**	0.022***
Most News Social	0.17	0.13	0.17	0.18	0.17	-0.013*	0.000	-0.014*
Device								
Mobile	0.63		0.63	0.64	0.62	-0.007	0.007	-0.015
Facebook								
Female	0.52	0.51	0.52	0.52	0.52	-0.006	-0.004	-0.003
Age	49.6	47.5	49.6	49.1	50.0	0.489	-0.425	0.914**
Total Subscriptions	472		478	475	463	3.084	15.094	-12.011
News Outlets Subscriptions	8.65		8.61	8.64	8.70	-0.027	-0.088	0.061
News Outlets Slant (-1, 1)	-0.22		-0.22	-0.22	-0.22	0.001	-0.005	0.006
Access Posts, Pre-Treatment	0.98		0.98	0.98	0.97	0.000	0.004	-0.004
N	17,634		6,116	5,763	5,755			
F-Test						0.914	0.717	1.161
P-value						(0.602)	(0.871)	(0.249)

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 4.

Table A.3: Balance Table by Assignment to the Pro-attitudinal and Counter-attitudinal Treatment, among Participants who Completed the Follow-up Survey

	Mean					Difference		
Variable	All	US	Control	Pro-Att.	Counter-Att.	Control - Pro.	Control - Counter.	Pro. - Counter.
Baseline Survey								
Ideology, Abs. Value (0, 3)	1.80	1.22	1.84	1.85	1.84	-0.002	0.004	0.005
Republican	0.16	0.25	0.16	0.16	0.16	0.002	0.000	-0.002
Independent	0.36	0.29	0.35	0.36	0.35	-0.018**	-0.005	0.013
Democrat	0.40	0.3	0.42	0.41	0.41	0.014	0.009	-0.005
Vote Support Clinton	0.55		0.57	0.57	0.57	-0.001	-0.001	0.001
Vote Support Trump	0.25		0.26	0.26	0.25	0.001	0.007	0.006
Feeling Therm, Difference	50.1	39.2	50.7	50.0	49.7	0.767	1.038*	0.271
Difficult Pers., Difference	1.96		1.98	1.94	1.95	0.045	0.035	-0.010
Most News Radio	0.09	0.08	0.09	0.09	0.09	0.000	-0.002	-0.003
Most News Web	0.49	0.21	0.50	0.48	0.48	0.019**	0.012	-0.006
Most News TV	0.19	0.53	0.18	0.19	0.18	-0.010	-0.001	0.009
Most News Social	0.17	0.13	0.17	0.17	0.18	-0.004	-0.010	-0.006
Device								
Mobile	0.63		0.63	0.63	0.62	-0.005	0.006	0.011
Facebook								
Female	0.52	0.51	0.52	0.52	0.53	-0.005	-0.009	-0.004
Age	49.6	47.5	49.8	49.8	49.7	0.050	0.096	0.046
Total Subscriptions	472		474	468	472	6.730	2.277	-4.453
News Outlets Subscriptions	8.65		8.70	8.72	8.83	-0.027	-0.129	-0.102
News Outlets Slant (-1, 1)	-0.22		-0.22	-0.21	-0.23	-0.011	0.007	0.018
Access Posts, Pre-Treatment	0.98		0.98	0.98	0.97	-0.000	0.004	0.004
N	17,634		6,116	5,760	5,758			
F-Test						0.817	0.808	0.907
P-value						(0.729)	(0.741)	(0.600)

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 4 and 5.

Table A.4: 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	-0.30	-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
Republican	0.17	0.15	0.21	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
Democrat	0.40	0.44	0.33	0.45	0.34	0.47	0.29	0.40	0.37
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	0.18	0.35	0.15	0.37	0.25	0.28
Feeling Therm, Rep.	29.3	25.8	33.5	23.3	33.8	21.7	36.5	27.6	30.4
Feeling Therm, Dem.	47.9	51.0	43.0	53.6	42.3	54.6	39.0	48.4	45.2
Difficult Pers., Rep. (1, 5)	3.15	3.24	3.00	3.34	2.96	3.38	2.86	3.18	3.10
Difficult Pers., Dem. (1, 5)	2.38	2.22	2.57	2.09	2.62	2.05	2.74	2.31	2.47
Most News Radio	0.09	0.09	0.09	0.08	0.09	0.08	0.09	0.09	0.09
Most News Web	0.48	0.48	0.46	0.49	0.47	0.48	0.47	0.48	0.47
Most News TV	0.20	0.20	0.21	0.19	0.20	0.19	0.20	0.20	0.21
Most News Social	0.17	0.17	0.17	0.19	0.17	0.18	0.17	0.18	0.17
Mobile	0.67	0.66	0.70	0.66	0.67	0.67	0.68	0.65	0.69
Female	0.52	0.56	0.47	0.60	0.46	0.58	0.46	0.56	0.47
Age	48.7	49.4	47.9	48.2	49.0	48.5	48.1	49.0	48.5
Total Subscriptions	476	497	433	525	431	516	431	507	432
News Outlets Subscriptions	8.52	9.21	7.81	9.16	8.12	9.18	7.81	9.05	7.92
N	12,105	6,734	5,380	5,512	6,596	6,279	6,216	6,312	6,181

This table presents descriptive statistics on compliance by treatment arm. For an explanation on each variable see Table 4.

Table A.5: Primary Outcomes, Controlling for Covariates

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

	(1)	(2)	(3)
Conservative Treatment	0.010	−0.002	−0.002
	(0.018)	(0.006)	(0.005)
Liberal Treatment	−0.008	−0.008	−0.006
	(0.018)	(0.006)	(0.005)
Cons. Treatment - Lib. Treatment	0.018	0.006	0.004
	(0.019)	(0.006)	(0.005)
Common Controls		X	X
Baseline Political Opinions Controls			X
Observations	17,643	17,643	17,643

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

	(1)	(2)	(3)
Counter-Att. Treatment	−0.050***	−0.033**	−0.023*
	(0.019)	(0.015)	(0.012)
Pro-Att. Treatment	−0.016	0.0002	0.007
	(0.019)	(0.015)	(0.012)
Pro-Att. Treat. - Counter-Att. Treatment	0.033*	0.033**	0.031**
Common Controls		X	X
Baseline Affective Polarization Controls			X
Observations	16,894	16,894	16,894

These table present the effect 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 square and gender. Column (3) also controls for baseline question similar to endline questions composing each index. The specification and controls are described in more detail in Section 3.6. *p<0.1 **p<0.05 ***p<0.01

Table A.6: Effect of Conservative and Liberals Switching Facebook Feeds

FB Slant, Std. Dev.	0.040
	(0.036)
First Stage F	50.9
Control Difference in Slant: Conservative - Liberal	-1.193
Effect of Switching Feeds	-0.048
Control Difference in Pol. Opinoins: Conservative - Liberal	1.875
Effect of Switching Feed, Share of Control Group	-0.026
Observations	1,123

The table shows how the political opinions index would have changed if the slant of posts in the Facebook feed of a liberal participant would have been been similar to the posts in the feeds of conservative (or vice versa). The calculation is based on the following steps. *FB Slant, Std. Dev.* shows the effect of the slant of posts in the Facebook feed on the political opinion index. The regressions are IV regressions with the treatment as the instrument and controlling for the covariates specified in section 3.6. *Control Difference in Slant* is the difference in the slant of posts in the Facebook feed of liberals and conservatives in the control group. *Effect of Switching Feeds* multiples the effect found with the difference to show how much opinions would have change if the feeds of liberals and conservatives had the same slant. *Control Difference in Slant* compares this effect to the actual difference between the opinions of conservatives and liberals. *Effect of Switching Feed, Share of Control Group* divides the effect of switching feed by the difference in the control group to calculate the final figure discussed in section 5.1. The sample is all posts between the baseline and endline surveys for participants who installed the extension for at least two weeks. *p<0.1 **p<0.05 ***p<0.01

Table A.7: Effect of a Balanced Facebook Feed

	Affective Polarization: Std. Dev.	Feeling Thermometer: Degrees
	(1)	(2)
FB Counter-Att. Share	−0.564*	−11.756
	(0.292)	(7.933)
Control: Counter-Att. Share	0.166	0.166
Effect of Balanced Facebook Feed	-0.187	-3.909
Observations	1,071	1,030

The table shows how affective polarization would have changed if the Facebook feed was balanced. The calculation is based on the following steps. *FB Counter-Att. Share* shows the effect of the share of counter-attitudinal news on the affective polarization index and the feeling thermometer measure. The regressions are IV regression with the treatment as the instrument and controlling for the covariates specified in section 3.6. *Control: Counter-Att. Share* shows that in the control group, approximately 17% of posts were counter-attitudinal. *Effect of Balanced Facebook Feed* shows how affective polarization and the feeling thermometer outcome would have decreased if the share of counter-attitudinal posts increased to 50%. The sample is all posts between the baseline and endline surveys for participants who installed the extension for at least two weeks. *p<0.1 **p<0.05 ***p<0.01

Table A.8: Effect of a Facebook Feed Equating the Congruence Scale of Online News Consumed

	Browsing Congruence Scale, Std. Dev.	Affective Polarization, Std. Dev.	Feeling Thermometer, Degrees
	(1)	(2)	(3)
FB Congruence Scale, Std. Dev.	0.332***	0.115*	2.115
	(0.070)	(0.059)	(1.611)
Control Group Diff. in Congruence:			
Other Sources - FB	-0.18		
Effect Required to Equate Congruence	-0.542		
Effect of Equating Congruence		-0.058	-1.094
Observations	1,083	1,088	1,046

The table shows how affective polarization would have changed if news consumed through Facebook had the same congruence scale as other news consumed. The calculation is based on the following steps. Column (1) shows the effect of the congruence scale of news exposed to in Facebook on the congruence scale of news sites visited. The congruence scale measures the mean slant of all news an individual was exposed to, multiplied by (-1) for liberal participants. Columns (2) and (3) show the effect of the congruence scale of news exposed to in Facebook on the affective polarization index and the feeling thermometer outcome. All the regressions are IV regression with the treatment as the instrument and controlling for the covariates specified in section 3.6. *Control Group Diff. in Congruence* shows that in the control group there is a difference of 0.18 standard deviations in the congruence scale between news sites visited through Facebook and other news sites visited. *Effect Required to Equate Congruence* shows that the congruence scale of the Facebook feed has to increase by 0.54 to equate the congruence scale of sites visited through Facebook and other sites visited. It is calculated by dividing the second row by the first row. *Effect of Equating Congruence* shows how affective polarization and the feeling thermometer would have decreased if the congruence scale in the the feed would have increased by 0.54 and sites visited through Facebook had the same congruence scale as other sites visited.
The sample is all posts between the baseline and endline surveys for participants who installed the extension for at least two weeks. *p<0.1 **p<0.05 ***p<0.01

Table A.9: Estimations Decomposing the Segregation in News Exposure

	Subscribed OLS	Exposed IV	Usage OLS
	(1)	(2)	(3)
Offered	1.435***		
	(0.063)		
Offered * Pro	0.504***		
	(0.095)		
Subscribed		0.010***	
		(0.001)	
Subscribed * Pro		0.006***	
		(0.002)	
Counter-Att. Treatment			−19.409*
			(11.007)
Pro-Att. Treatment			0.562
			(11.626)
Outlet and Ind. FE	X	X	
Observations	11,984	11,984	1,498

This table displays the regressions used to decompose the gap in exposure between pro-attitudinal posts in the pro-attitudinal treatment and the counter attitudinal posts in the counter-attitudinal treatment. In column (1), all potential outlets (the four pro-attitudinal and four-counter-attitudinal outlets assigned to each participant) and participants are pooled. The dependent variable is whether the participant subscribed to an outlet and the independent variable is the full interaction of whether the outlet was offered in the experiment and whether the outlet is pro-attitudinal outlet for the participant subscribing to the outlet. In column (2), all potential outlets and participants are pooled in an IV regression. The dependent variable is the share of all posts from the outlet supplied to the individual, the independent variable is the full interaction of whether the participant subscribed to the outlet and whether the outlet is pro-attitudinal. Subscription to an outlet is instrumented with the outlet being offered in the experiment. In the first two columns, standard errors are clustered at the individual level. In column (3), the dependent variable is the number of posts observed by the participant and the independent variable is whether the participant was assigned to the pro-attitudinal or counter-attitudinal treatment. The regression controls for Facebook visits before the intervention. *p<0.1 **p<0.05 ***p<0.01

Table A.10: Association between Consumers' Ideology and the Slant of Posts in the Facebook Feed

	Slant		Mean Slant	Slant		Mean Slant
	(1)	(2)	(3)	(4)	(5)	(6)
Conservative Ideology	0.357***	0.139***	0.126***	0.129*	−0.008	−0.024
	(0.021)	(0.010)	(0.005)	(0.067)	(0.009)	(0.018)
Level of Observation	URL	URL	Ind*Outlet	URL	URL	Ind*Outlet
Data = All Domains	X	X	X			
Data = Potential Exp. Outlets				X	X	X
Outlet FE		X	X		X	X
Observations	227,099	227,099	61,162	20,341	20,341	1,435

This table shows the correlation between ideology and the news individuals were exposed to in their Facebook Feed. In columns (1), (2), (4), (5), each observation is a link to an article which appeared in a participant's Facebook feed and the dependent variable is the slant of articles consumed. In columns (3) and (6), the data is aggregated at the participant*outlet level and the dependent variable is the mean slant of articles consumed. The slant of each article is calculated according to the DW-Nominate score of all congress members who shared the article in Facebook or Twitter. Each URL is only counted once for each congress member who shared it. The URLs are matched with Facebook URLs by first finding the redirected URL, and then trimming the protocol from the URL (e.g. https://) and any query within most URLs. To increase power, instead of focusing only on two weeks, the data spans from the intervention until the end of April 2018.
Columns (1)-(3) include all links displayed in the Facebook feed. These columns show that there is clear correlation between ideology and the links observed. Column (2) shows that this correlation holds even with domain fixed effects. 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.
Columns (4)-(6) focus only on posts from the eight potential outlets which could be offered to each participant and control for list of potential outlets. Column (4) shows that even within the outlets included in the experiment there is correlation between ideology and the slant of articles. This could be explained by both the tendency of individuals to subscribe to outlets that fit their ideology, and by Facebook's tendency to display more posts from outlets that match one's ideology. Column (5) controls for outlet fixed effects and shows that there is no correlation between consumer's ideology and the article slant within an outlet. Column (6) aggregates the data at the individual*outlet level and leads to the same conclusion. Standard errors are clustered at the individual level. *p<0.1 **p<0.05 ***p<0.01

Table A.11: Primary Outcomes Using Different Index Methods

(a) Political Opinions

	(1)	(2)	(3)	(4)	(5)
Liberal Treatment	−0.004	−0.005	−0.038	−0.004	−0.003
	(0.005)	(0.017)	(0.083)	(0.007)	(0.005)
Conservative Treatment	−0.001	0.022	−0.059	0.007	0.004
	(0.005)	(0.017)	(0.084)	(0.007)	(0.005)
Cons. - Lib. Treatment	0.003	0.027	0.010	-0.021	0.006
	(0.005)	(0.017)	(0.007)	(0.053)	(0.005)
Index Method	Standrd	Inv-Cov	Inv-Cov, All Obs	Inv-Cov	Inv-Cov, All Obs
Negative Weights	-	Include	Include	Exclude	Exclude
Observations	17,128	9,251	17,128	9,251	17,128

This table estimates the effect of the treatment on the political opinions index 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 observations for any of the measures. In column (3), inverse-covariance weights are used for all participants with non-missing outcomes, for participants with missing outcomes the weights are renormarlized 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 zero and all weights renormalized to sum to 1. The specification and controls are described in Section 3.6.

(b) Affective Polarization

	(1)	(2)	(3)
Counter-Att. Treatment	−0.023*	−0.031**	−0.021**
	(0.012)	(0.014)	(0.010)
Pro-Att. Treatment	0.007	0.005	0.002
	(0.012)	(0.014)	(0.010)
Pro-Att. Treat. - Counter-Att. Treatment	0.031**	0.036**	0.023**
	(0.012)	(0.014)	(0.010)
Index Method	Standard	Inv-Cov	Inv-Cov All Obs
Observations	16,894	13,616	16,894

This table estimates the effect of the treatment on the affective polarization index 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 observations for any of the measures. In column (3), inverse-covariance weights are used for all participants with non-missing outcomes, for participants with missing outcomes the weights are renormarlized to sum to 1, such that an outcome measure is created for all participants who have at least one non-missing outcome. The specification and controls are described in Section 3.6.

Table A.12: Effect of the Treatment on Behavioral and Attitudinal Polarization Measures

	All	Affective	Behavior
Counter-Att. Treatment	-0.023^{*}	-0.023^{*}	-0.007
	(0.014)	(0.012)	(0.018)
Pro-Att. Treatment	0.006	0.007	-0.005
	(0.014)	(0.012)	(0.018)
Pro-Att. Treat. - Counter-Att. Treatment	0.029^{**}	0.031^{**}	0.002
	(0.014)	(0.012)	(0.019)
Controls			
Observations	17,159	16,894	16,640

This tables estimates the effect of the treatment on polarization using different indices. Column (1) includes the five self-reported affective polarization outcomes along with the three polarization behavioral measures. Column (2) includes only the five self-reported affective polarization outcomes and column (3) includes only the three polarization behavioral measures. The affective polarization outcomes are described in Appendix Section 3.4 and the behavioral outcomes are described in Section C. The specification and controls are described in Section 3.6.

Table A.13: 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	0.007	−0.004	0.008	0.002	0.015*	−0.003	−0.001
	(0.006)	(0.007)	(0.006)	(0.008)	(0.005)	(0.009)	(0.009)	(0.006)
Conservative Treatment	−0.002	0.002	−0.002	0.019**	0.010*	0.0001	−0.007	0.0003
	(0.006)	(0.007)	(0.006)	(0.008)	(0.005)	(0.009)	(0.009)	(0.006)
Cons. Treat - Lib. Treat	0.00	-0.01	0.00	0.01	0.01	-0.01*	-0.00	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Expected Effect	Lib Treat	Lib Treat	Cons Treat	Cons Treat	Lib Treat	Lib Treat	Cons Treat	Cons Treat
Observations	17,643	17,439	17,643	17,472	16,177	13,879	12,149	15,663

This tables 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 square, whether the participant follows the news and whether the participant stated they know the name of their representative in congress. *p<0.1 **p<0.05 ***p<0.01

Table A.14: 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	2.919**	1.436***	0.150	0.066
	(1.275)	(0.418)	(0.157)	(0.055)
Conservative Treatment	0.527	0.178	0.340***	0.064*
	(1.024)	(0.311)	(0.125)	(0.037)
Cons. Treat - Lib. Treat	-2.39**	-1.26***	0.19	-0.00
	(1.13)	(0.38)	(0.16)	(0.06)
Expected Effect	Lib. Treat	Lib. Treat	Cons. Treat	Cons. Treat
Observations	1,212	1,212	1,212	1,212

This table estimates the effect of the treatment on topics appearing in the participant's Facebook feed. *Michael Cohen* is the number of times the expression "Michael Cohen" appeared. *Clark Shooting* is the number of times the expression "Stephon Clark". *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 square, 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. The data is not limited to the first week after the extension installed since 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 occurred. *p<0.1 **p<0.05 ***p<0.01