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

Ro'ee Levy

A Data Collection and Processing

A.1 Surveys

A.1.1 Recruitment Ads

The Facebook ads recruiting participants to the baseline survey mentioned that a research survey was conducted by Yale University and that participants could win Amazon gift cards (Appendix Figure A.13). One version of the ad suggested that the survey was about politics and the other suggested that it was about American society.[1]

Most participants were recruited through ads targeting all Facebook users living in the US who are over 18 years old. Using a Facebook Pixel, the ads targeted Facebook users who were more likely to begin the survey. A subset of the ads targeted conservatives or moderate individuals who are often under-represented in Internet samples. Since the majority of participants took the survey on a mobile phone, an additional subset of ads focused on desktop users, to ensure that a large enough sample of participants will be offered an option to install the Chrome extension. A very small minority of users seemed to have a technical issue when taking the survey using the iOS operating system and therefore iOS users were excluded from the target audience once this was discovered (the sample still contains many iOS users). While the survey was open and participants could share the link or ad with anyone, the vast majority of participants probably entered the survey as a result of the ad.[2]

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

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

A.1.2 Baseline Survey

The baseline survey took place from early February to mid-March 2018. 40,504 responders took the survey and reached the screen where the intervention occurs. Of those, 37,494 are included in the final sample. Responders are excluded from the final sample for the following reasons: missing information on outlets the responder subscribes to either because the responder did not provide permissions to access that data or since the data were not collected properly in real-time (2.38%); the responder already subscribed to too many of the outlets such that it was not possible to define four potential liberal outlets and four potential conservative outlets (4.01%); technical issues with the Qualtrics survey which prevented some data from being collected (0.90%); taking the survey a second time (0.01%); responding carelessly (0.12%). Careless responders are defined as responders who completed all survey questions until the intervention exceptionally quickly (in under three minutes where the median time was eleven minutes) and responders who did not answer at least one-half of the closed-ended, non-required questions, or who did not answer any question on the final page before the intervention. Finally, to slightly reduce the number of outlets, alternative outlets which are defined as potential outlets for fewer than 20 participants are excluded from the experiment, along with the participants for which these outlets were defined as potential outlets. This removes fewer than 0.1% of participants from the baseline sample.

A.1.3 Endline Survey

Participants were invited to the endline survey between mid-April and early June 2018. Participants were mostly recruited to the survey using emails and Facebook ads.[3] To match endline survey responses with baseline survey responses, participants were asked to log in to the endline survey through Facebook or supply an email address. I match endline responses based on the following criteria: email address the survey invitation was sent to, Facebook id, email address entered in the survey, combination of zip code, first and last name if the combination is unique, and combination of first and last name if the combination is unique. 98.73% of endline responses were matched with baseline responders.

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

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

A.2 Facebook Data on Subscriptions and Posts Shared

I collect data on outlets participants subscribed to (pages liked) and posts they shared using a Facebook app, which provides an interface between a Facebook account and the survey.[4] The data allowed me to customize the survey by ensuring participants are not offered outlets they already subscribed to and including questions about the potential outlets. The app was approved through the standard Facebook review process.

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

I match posts participants shared with leading outlets based on the Facebook page which published the original post. If a post is not matched with any Facebook page, I determine the slant of the post based on the domain of a link included in the post. For outlets offered in the experiment, I expand the list of domains in the Bakshy, Messing and Adamic (2015) dataset to decrease measurement error. For each outlet, I create a list of relevant domains by checking which domains were shared by the Facebook page associated with the outlet and including the most dominant domains and any other domain related to the outlet. For example, I associate both "huffpost.com" and "huffingtonpost.com" with HuffPost.

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

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

A.3 Extension Data

I collect data on the Facebook feed and browsing behavior using the Google Chrome browser extension. Participants who took the survey on a computer using Chrome were offered to install the extension in the baseline survey before the intervention. In exchange for installing the extension for at least 48 hours, participants could choose between receiving a $5 gift card, participating in a lottery for a $200 gift card, or receiving a copy of the study results.

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

A.3.1 Browsing Behavior

I observe news sites that participants visited when the extension was installed. News sites visited are matched to outlets based on their domain. A news site is determined to have been visited through Facebook if the website visited appeared in the participant's Facebook feed in the 20 minutes proceeding to the visit.[5] I exclude URLs that were visited for less than one second before another URL was visited. If a URL is visited more than once within a 20-minute window, only the first visit is included. When estimating an effect on browsing behavior, I control for baseline browsing behavior in the eight weeks before the intervention, when the data exists.

A.3.2 Facebook Feed

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

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

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

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

In my data, I cannot precisely identify whether a post is sponsored or organic. Instead, I use two techniques to identify ads. First, I assume that any post seen by at least two participants who did not subscribe to the post's page is sponsored. Second, I assume that any post that appeared more than twice in a participant's feed for at least two participants is sponsored. Facebook's algorithm usually does not show the same post many times to the same user, however, advertisers can

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

choose to maximize impressions and thus may show the same post repetitively. When determining whether a post is sponsored, I assume that two posts from the same page with the same text are the same post, even if they have a different id, since advertisers can use two separate posts to run identical advertisements.[6]

While these criteria are far from perfect, they do seem to identify many ads. For example, based on my classification, the top ten words that are most likely to appear in posts identified as ads are: "get, now, free, new, today, just, time, one, us, help". In contrast, the top ten words most likely to appear in organic posts are: "trump, president, one, people, new, school, just, gun, like, now."[7]

A.4 Leading News Outlets

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

I exclude from the dataset the following popular websites which are not directly related to news: Amazon, Barack Obama, The White House, Twitter, Vimeo, Wikipedia, and YouTube. I also exclude MSN and AOL since these sites are aggregators of a wide variety of content, they may serve as homepages, and they are often visited for reasons not related to news consumption (Peterson, Shared and Iyengar, 2019). I merge websites that appear twice in the dataset, with and without a web reference, into one entry. For example, washingtonexaminer.com and www.washingtonexaminer.com are merged, with the slant defined as the mean slant of the two entries. After processing the data, the list of leading outlets contains 487 websites.

A.5 Comscore Data

The Comscore Web Behavior Database Panel is a subset of Comscore's opt-in Media Matrix Panel, which is weighted to represent the US Internet population. Each observation includes a unique machine (computer) id, which I assume represents an individual, although it is possible that multiple individuals use the same machine. When combining data for multiple years, I assign each individual the zip code in the last year for which data exists.

When classifying the referral channel through which a news site was visited, the referring channel is defined as social if the referring domain is one of the following: "facebook.com",

[6] I make this assumption when the text is at least 20 characters long.
[7] The terms exclude stop words along with the words http, can, said, see.

"live.com", "t.co", "reddit.com", "pinterest.com", "youtube.com", "linkedin.com", "twitter.com", "tumblr.com", "instagram.com". I classify any referral domain that includes the word google (e.g. "google.com" or "google.co.uk") as a search domain along with the following domains: "yahoo.com", "bing.com", "ask.com", "duckduckgo.com", "searchencrypt.com", "searchlock.com", "searchincognito.com", "search.com", "searchprivacy.co", "safesear.ch", "myprivatesearch.com", "netfind.com". I classify a site as visited directly is there is no referral domain or if the referral domain is the same domain as the domain visited.

B Additional Details on Empirical Strategy

B.1 Segregation Measures

This section describes the isolation and segregation measures in more detail, along with other measures which are presented in Appendix Tables A.7 and A.9.

B.1.1 Isolation

Isolation is the difference between the mean share of conservatives that conservatives are exposed to in the outlets they visit and the mean share of conservatives that liberals are exposed to. Exposure to conservatives in a website is defined as the share of conservatives browsing the site among all the site's visitors. Isolation can be calculated using the following formula:

$$Isol = \sum_{i \in \{C_i\}} WeightAmongCons_i * ConsExposure_i - \sum_{i \in \{L_i\}} WeightAmongLib_i * ConsExposure_i$$

where $WeightAmongCons_i$ is the share of outlets visited by individual i among all outlets visited by conservatives, $\{C_i\}$ is the set of conservative individuals, $\{L_i\}$ is the set of liberal individuals, and $ConsExposure_i$ is exposure to conservatives by individual i. Exposure can be calculated as the average share of conservatives among all outlets visited by individual i. To prevent a small sample bias, the average share does not include the visits by i:

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

where $Visits_{ij}$ is the number of visits of individual i to outlet j and $Visits_i$ is total visits by individual i, so $\frac{Visits_{ij}}{Visits_i}$ is the weight of outlet j for individual i. $Visits_j$ is total visits to site j and $Cons_j$ is total conservative visits to site j, so $\frac{Cons_j - Visits_{ij}}{Visits_j - Visits_{ij}}$ is the share of conservatives visiting outlet j excluding individual i.

B.1.2 Segregation

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

$$Seg = \sqrt{2} * std.dev(\bar{Slant}_i)$$

where $\bar{Slant}_i$ is the mean slant of outlets visited by individual i. The slant of outlet j is based on Bakshy, Messing and Adamic (2015) and normalized to the unit interval (by adding one and dividing by two).

B.1.3 Absolute Value of Slant

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

$$AbsSlant = \sum_i \frac{|\bar{Slant}_i|}{N}$$

where $\bar{Slant}_i$ is the mean slant of outlets visited by individual i and N is the number of individuals in the sample. The slant of outlet j is based on Bakshy, Messing and Adamic (2015) such that a middle-of-the-road outlet has a slant of zero, a completely conservative outlet has a slant of 1 and a completely liberal outlet has a slant of -1.

B.1.4 Congruence

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

$$Congruence = \sum_i \frac{(\bar{Slant}_i * IdeoLeaning_i)}{N}$$

where $\bar{Slant}_i$ has the same definition as in the previous measure and $IdeoLeaning$ is defined as 1 for a conservative participant and -1 for a liberal participant. N is the number of individuals in the sample for which an ideological leaning can be defined.

B.1.5 Share of Counter-Attitudinal News

To determine the share of counter-attitudinal news, I divide news sites into five quintiles: very liberal, liberal, moderate, conservative, and very conservative (Bakshy, Messing and Adamic, 2015). I define pro-attitudinal news as conservative and very conservative news consumed by a conservative, or liberal and very liberal news consumed by a liberal. Counter-attitudinal news is

conservative and very conservative news consumed by a liberal, or liberal and very liberal news consumed by a conservative. Finally, the share of counter-attitudinal news is defined as the share of counter-attitudinal news among all pro- and counter-attitudinal news.

$$ShareCounter = \sum_i \frac{\frac{\sum_j (IdeoLeaning_i \neq SlantGroupOutlet_j)}{\sum_j SlantGroupOutlet_j \in \{-1,1\}}}{N}$$

where $SlantGroupOutlet_j = 1$ if outlet j is conservative or very conservative and $SlantGroupOutlet_j = -1$ if outlet j is liberal or very liberal (moderate outlets are excluded).

B.2 Pre-Analysis Plan

The main outcome and hypotheses tested in this study were pre-registered in the AEA RCT Registry.[8] The analysis deviates from the pre-analysis plan in two important ways. First, I use equal weights when constructing the indices, while the plan states that the weights will be determined by the inverse of the covariance between the outcome measures (Anderson, 2008). This method is not used since it generates negative weights. With negative weights, the interpretation of an index is less clear. For example, the question on President Trump's approval rating received a negative weight which means that *ceteris paribus*, a participant who has a more favorable opinion on Trump would be considered more liberal.

Appendix Table A.18a estimates the effect on the political opinions index using equal weights in column (1) and inverse-covariance weights in column (2). This method does not cleanly generate weights for individuals with missing outcomes. In column (3), weights from column (2) are renormalized to sum to one for participants with missing outcomes, an index is then created for each participant by weighting the standardized outcomes, and finally, the index is standardized with respect to the control group. Since the inverse-covariance method generates negative weights, columns (4) and (5) repeat the analysis with negative weights replaced with zero and the weights renormalized accordingly. While there is some variation in the results, the most straightforward comparison is between columns (1) and (5). These columns focus on the same participants and do not include negative weights. In column (5), the effect of the conservative treatment is slightly larger but still small in magnitude and not statistically significant. Appendix Table A.18b shows that the effect on affective polarization is robust to using inverse-covariance weights.

The second important deviation from the pre-analysis plan is that the polarization index originally included five attitudinal measures and three behavioral measures, while only the attitudinal measures are analyzed in this paper. The behavioral measures were based on a question in the endline survey asking participants whether they would "like" or share a post stating that "In seeking truth, you have to get both sides of a story." The primary behavioral outcome is composed of an index of the following measures: did participants state they will share the post, did

[8] AEA RCT Registry Trial 0002713.

participants state they would "like" the post, did participants actually share the post. However, it was not possible to analyze the actual behavior of a large share of participants partly due to the unexpected Cambridge Analytica scandal, which led many individuals to revoke access to the posts they share. Furthermore, the behavioral measure turned out not to measure polarization well. While a measure of polarization should typically be correlated with partisanship, there was almost no correlation between being partisan and the behavioral outcomes.[9]

Column (1) of Appendix Table A.19 shows that the effect is still significant when using all eight variables in the polarization index.[10] Column (3) measures the effect only on the behavioral outcomes (for most participants I cannot observed whether posts were shared so this index is mostly based on the self-reported survey answers). The effect of the treatments is small and not statistically significant. While this result does not change the conclusions regarding affective polarization, it is interesting to note that exposure to counter-attitudinal outlets does not affect participants' self-reported willingness to share or like a post on seeking both sides of a story.

When processing and analyzing the data, I made various other minor changes compared to the pre-analysis plan, include the following. In the plan, I stated that I will estimate the results excluding the first two days after the intervention. Instead, I estimate the results for each week or month separately. The plan states that the regression will control for the randomization block and for whether the participant used the iOS operating system. I exclude the iOS variable for simplicity (this does not affect the primary endline survey results). I do not control for the randomization blocks (strata) since due to attrition, some blocks have only one or two respondents instead of the original three respondents. Instead, I control for ideological leaning. When controlling for the block, I am only able to analyze a subset of participants. The results for that subset are essentially the same with and without controlling for strata. I do not report raw or adjusted p-values for each index component of the political opinions and affective polarization measures, as I do not focus on the individual components. Instead, I present each component visually in appendix figures.

In the pre-analysis plan, ideological leaning is defined first by self-reported ideology and then by party affiliation. I prefer using party affiliation as the main variable defining ideological leaning to make the study comparable to other papers, which tend to focus on party affiliation (Druckman and Levendusky, 2019). The results are robust to the original definition. In contrast to the plan, I do not present several demographic variables in the balance table since they suffer from post-treatment bias and do not impute them since I already have rich survey and social media data. Finally, the pre-analysis plan states that a political knowledge index will be created. Since I do not focus on political knowledge, I instead analyze separately the effect on each political knowl-

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

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

edge primary outcome in Appendix C.6. While the results are easier to interpret when analyzed separately, an index would not change the qualitative conclusions of the section.

B.3 Controls

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

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

When estimating the effect on affective polarization, I also control for the baseline values of the *feeling thermometer* and *difficult perspective* measures (defined in Section II.D.2).

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

C Additional Analysis

C.1 Survey Purpose

At the end of the baseline survey, participants were presented with the following question: "If you had to guess, what would you say is the primary purpose of this study?" Appendix Table A.20 shows the most common phrases that participants mentioned according to their treatment assignment. Unsurprisingly, participants understood that the study is on media and politics, as

most questions focused on these topics and the consent form stated that this is the topic of the study. Among the most common phrases, there are not many substantial differences between the treatments.

Appendix Table A.21 presents the phrases with the largest differential usage between the treatment arms and the control group. While participants in both the pro- and counter-attitudinal treatments mentioned terms such as "echo chamber" and "social media" more often than the control group, probably due to the text of the intervention encouraging participants to like Facebook pages, the differences between the two treatment arms in the usage of these terms is small. When comparing the pro- and counter-attitudinal treatments to each other, almost no substantial differences stand out. One exception is that a small share of participants in the counter-attitudinal treatment thought the purpose of the survey was to get them to like liberal Facebook pages. These participants probably were not pleased with the experimenter trying to "push liberal" content (that was not the actual purpose of the experiment, of course) and therefore it is unlikely that they expressed opinions aligned with these outlets to make an impression on the experimenter. In any case, while these phrases represent a relatively large difference between the treatments, they are not mentioned often.

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

C.2 Analysis of the Content that Participants Engaged With

In this section, I show that the most common content participants engaged with as a result of the intervention is political. I analyze the posts from the subscribed outlets that participants were exposed to in their feed, links in the posts that they visited, and posts they shared using three methods. First, I show the most common phrases mentioned in the posts. Second, I define certain terms as political and analyze the share of political posts. Third, I analyze the section and outlet where each article appeared based on the URLs appearing in the posts.

An important challenge in this analysis is that the posts affected by the treatment cannot be cleanly identified. For example, participants in the control group visited the news sites of their potential counter-attitudinal outlets approximately 1.70 times in the two weeks following the intervention, while participants in the counter-attitudinal treatment visited these websites approximately 1.34

additional times (as shown in Figure 6). While the participants were affected by the treatment, I cannot identify which of their visits to counter-attitudinal news sites would have occurred in a counterfactual with no intervention. I focus on posts affected directly by the intervention by analyzing only posts shared by pages participants subscribed to in the experiment (excluding suspected ads). While this decreases the likelihood of including posts that participants would have engaged with without the intervention, it does not cover the entire effect of the intervention. For example, participants often visited the websites of the offered outlets indirectly, even when they did not observe the specific link to an article in their feed (as shown in Figure A.5).

Throughout this section, I focus on the eight weeks following the intervention to increase the number of data points. To reduce variability in the text analyzed, I include in the analysis only posts from the eight primary outlets and first two alternative outlet that were offered to participants. This excludes less than 3% of posts participants were exposed to.

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

C.2.1 Most Common Phrases

Appendix Table A.22 shows the most common phrases mentioned in posts participants were exposed to in their feed, in posts with links participants visited, and in posts shared by participants. I first remove punctuation, terms that appear in only one outlet, media-related terms or terms that were likely to be covered mostly by specific outlets (e.g., "write" or "New York"), and then stem the words appearing in the posts.[11]

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

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

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

Stormy Daniels and conservatives in the conservative treatment are more likely to share articles mentioning Hillary Clinton.

C.2.2 Share of Posts Mentioning Political Words

Focusing on the most common words allows us to understand which topics were most prominent but does not provide a complete analysis of the posts, especially if there is a lot of variability in the posts' content. In this subsection, I use a simple measure to determine a lower bound for the share of political posts. I define a post as political if it contains terms related to political figures ("biden, bolton, carson, clinton, devos, kushner, manafort, mccabe, mcconnell, michael cohen, obama, pelosi, pence, pruitt, tillerson, trump"), political parties ("conservative, democrat, dnc, gop, liberal, republican, the left, the right"), political institutions ("congress, elect, politic, senate, vote, white house") or political issues ("ar 15, daca, gun control, gun law, gun right, immigration, mass shooting, nra, parkland, sanctuary city, sanctuary state, school shooting, tax cut, walkout"). I search for the terms in the post's text, its URL, and any commentary on the post if it is shared.[12]

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

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

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

[12]Specifically, for shared posts I search for political terms in the message, description, and link fields.

C.2.3 Outlets and Sections

Instead of determining the posts' topics based on words in the post, I can analyze the content participants engaged with using the outlets' own classification of their articles. Most outlets classify articles into sections, such as News, Business, and Arts, and mention the sections on their website, the website's HTML, or the URL. I determine the section associated with a post based on analyzing the website associated with the post. This method is not perfect. MSNBC usually does not classify articles and videos into sections and Slate often creates short links for its URLs which were no longer available when I determined the link's section. Still, the advantage of this method is that it relies on internal decisions by the outlets, who should know their content best.

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

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

The differences between the posts shared by participants are more stark. For example, Appendix Figure A.17 shows that conservatives shared HuffPost articles in the parenting, women, or queer voices sections, while among posts shared by liberals, these sections form a very small minority.[13] Still, within each outlet, the dominant sections among posts shared are typically the political or national news sections, even in the counter-attitudinal treatment.

C.3 Heterogeneous Effects

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

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

I use two measures of being in an echo chamber. The variable *Echo Chamber* is whether the answer to "Thinking about the opinions you see people post about government and politics on Facebook,

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

how often are they in line with your own views" is above or equals the median. *Seen Counter Att.* is whether the share of potential counter-attitudinal outlets, among all potential outlets, participants reported seeing in their feed in baseline is above or equals the median.

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

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

In addition to the pre-registered tests, I explore the effect of several additional moderators. *Most News Social Media* is whether participants reported getting most of their news about government and politics through social networking sites. Participants have *High News Subscriptions* if their baseline number of subscriptions to pages of news outlets on Facebook is above or equals the median. Participants are considered *Exposed to Outlets* if their self-reported exposure to posts from the eight potential outlets in baseline is above or equals the median. Participants are considered to be *Familiar with Slant* if the distance between their perceived slant of the potential outlets and the average perceived slant by participants with the same self-reported ideology is below the median. Participants are considered to *Follow the News* if their answer to "how often do you pay attention to what's going on in government and politics?" is above the median. Participants are considered to have a *High Feeling Thermometer Difference* if the difference between their feeling toward their own party and the opposing party is above or equals the median. Finally, participants are considered *Conservative* if their ideological leaning is conservative, *Older* if their age is above or equal to the median age, and *Female* if they identify in Facebook as female.

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

Appendix Figures A.18 and A.19 estimate heterogeneous effects on subscribing to outlets, exposure to posts from outlets, and visiting the outlets' websites. Each row represents a separate regression estimating the effect of interacting the pro- or counter-attitudinal treatment with the specified variable, where the reference group is the control group. A higher value means individ-

uals were more likely to engage with the pro- or counter-attitudinal potential outlets as a result of the pro- or counter-attitudinal treatment, respectively.

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

Appendix Figure A.20 estimates heterogeneous effects on the primary endline survey outcomes. The left panel shows that the effect on political opinions is mostly homogeneous (i.e., most participants were not persuaded by the treatments). The right panel of Appendix Figure A.20 does not show strong heterogeneous effects on affective polarization according to most covariates tested.[14] The strongest heterogeneous effect found is based on the baseline feeling thermometer measure for affective polarization. The effect on affective polarization is weaker among participants who were more polarized in baseline. However, this result is significant at the 10% level and the results are not adjusted for multiple hypothesis testing, and therefore more research is needed to explore heterogeneity in affective polarization.

C.4 Reweighting for National Representativeness

C.4.1 Data Sources

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

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

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

C.4.2 Analysis

In this section, I reweight the sample to match the national population using the entropy weighting procedure (Hainmueller, 2012). I match the following subset of control covariates: self-reported ideology (mean value on a scale of 1-7), the share of participants identifying as Democrats, Republicans, and Independents, the difference between the participants feeling toward their party and the opposing party, age, and the share of females. For the feeling thermometer, self-reported ideology, age, and gender covariates, missing variables are first replaced with the mean value (less than 5% of observations are missing for each of these variables). When analyzing the effects of the pro- and counter-attitudinal treatments, I compare the sample to the US population for which an ideological leaning can be defined and use those means to reweight the sample.[15]

Appendix Tables A.23 and A.24 show that reweighting the sample does not change the main conclusions of the study. The effect on the slant of posts participants were exposed to increases slightly. The effect on sites visited, posts shared, political opinion, and affective polarization remain essentially the same, although the confidence intervals are wider. These tables should be interpreted with caution. It is likely that even after reweighting, the sample is still different than the national population on unobservables or covariates not used when reweighting the sample. Still, the tables show that it is unlikely that an effect on affective polarization is only found because the survey sample is more liberal or more polarized than the rest of the population.

C.5 Predicted Treatment Effect for the Full Baseline Sample

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

I first estimate heterogeneous effects on the slant of posts observed, the slant of news sites visited, the political opinions index, and the affective polarization index. The effects on media engagement are estimated in the extension subsample and the effects on self-reported opinions and attitudes are measured in the endline survey subsample.[16] I exclude the control group in these estimates so the interpretation is the effect of the conservative treatment on conservative media consumption and conservative opinions, compared to the liberal treatment, or the effect of the pro-attitudinal

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

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

treatment on polarization, compared to the counter-attitudinal treatment. I estimate heterogeneous effects using causal forests (Wager and Athey, 2018). The intuition behind causal forests is that one part of the sample is used to determine how to split each tree and another part is used to estimate heterogeneity. If the same sample was used for both processes, heterogeneity would be overestimated due to overfitting, as the sample would be split according to the covariates that happen to predict heterogeneous effects in this particular sample.

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

Appendix Table A.25 shows that the results predicted among the entire baseline sample are very similar to the results found among the subsamples of participants who completed the endline survey or installed the Chrome extension for at least two weeks. Based on the analysis of heterogeneity throughout this paper, the fact that the effects on opinions and attitudes are stable is not surprising, as the effects on the primary outcomes are generally homogeneous and the differences between participants in the baseline and endline surveys are not dramatic.

While these results are reassuring, two caveats should be noted. First, I control for many observable variables, but there could be unobservables differentiating the subsamples. Second, when estimating heterogeneous effect in the extension subsample, I cannot control for one important difference between the groups - the device with which the survey was taken - since participants could only install the extension when taking the survey on a computer using Google Chrome.

C.6 Effects on Knowledge

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

Appendix Table A.26 shows that the treatments had little to no effect on the knowledge outcomes. The coefficients of interest are the effects of the liberal treatment on liberal outcomes and conservative treatment on conservative outcomes. Most of the coefficients are small in magnitude and not statistically significant.

Appendix Table A.27 tests whether there is no substantial effect on knowledge because the treatment did not affect exposure to the topics the endline survey focused on. The table uses the extension data to estimate the effects of the treatments on posts appearing in the participants' social media and shows that the intervention affected all four self–reported familiarity outcomes (Michael Cohen, Stephon Clark, Louis Farrakhan, and the Hillary Clinton speech).[17]

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

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

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

C.7.1 Decomposition Calculations

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

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

$$S_C * A_C * U_C$$

where S_C is the mean number of new subscriptions to the offered counter-attitudinal outlets; A_C is the effect of the algorithm determining the share of posts in the feed from the subscribed counter-attitudinal outlets among all the posts in the feed (formally defined later in this section); and U_C

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

is the total number of posts observed in the feed in the counter-attitudinal treatment. I define the number of posts observed in the pro-attitudinal treatment as:

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

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

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

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

where $TotalSub_i$ and $TotalPosts_i$ are the number of offered outlets the participant subscribed to and the total number of posts observed, respectively. These regressions are presented in columns (1) and (2) of Appendix Table A.28. X_i controls for Facebook usage before the intervention to increase precision.

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

$$SharePosts_{ij} = A_C * Sub_{ij} + A_\Delta * Sub_{ij} \times Pro_{ij} + \delta * Pro_{ij} + \nu_{ij} \quad (1)$$

where $SharePosts_{ij}$ is the share of posts participant i observed from group j, Sub_{ij} is the number of outlets participant i subscribed to from group j. Pro_{ij} is whether the outlets in the group are pro-attitudinal. I instrument for Sub_{ij} and $Sub_{ij} \times Pro_{ij}$ with $Offer_{ij}$ and $Offer_{ij} \times Pro_{ij}$, where $Offer_{ij}$ is whether participant i is was offered outlets from group j in the intervention. This regression is presented in column (3) of Appendix Table A.28. Conceptually, it can be easier to think of this regression as two separate regressions. One regression includes only the potential counter-attitudinal outlets, and measure the effect of subscribing to an outlet on exposure to the outlet (A_C). I exploit the fact that for some participants the counter-attitudinal outlets were offered and for others they were not offered. In a second regression, I repeat this exercise for the potential pro-attitudinal outlets. A_Δ is the difference between the coefficients.

C.7.2 Alternative Decompositions

Appendix Figure A.21 presents the decomposition exercise using several alternative estimations. The x-axis is the gap in exposure to posts from the pro- and counter-attitudinal outlets, in the two weeks following the intervention. Most of these specifications lead to similar results, although I am often underpowered to detect precise effects. The first row of the figure is the primary specification shown in Figure 10. The second row adds fixed effects for the potential outlets defined for each participant. This assures that the estimates are derived from comparing participants who could have been offered the same set of outlets. The rest of the decompositions are described below.

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

Exclude Suspected Ads In the primary decomposition, I assume that Facebook's algorithm determines whether participants observe posts from outlets they subscribe to. This typically holds for organic posts. However, participants also observe sponsored posts (ads) which are different in several important aspects. First, they can appear in a user's feed even if she did not subscribe to the outlet. Second, the placement of sponsored posts can be determined by the advertiser. For example, an outlet can promote posts to a subset of users who subscribed to its Facebook page. This means that part of the effect attributed to the algorithm may result from the behavior of advertisers.[18] When excluding suspected ads, the gap between exposure to pro- and counter-attitudinal outlets slightly decreases. This suggests that ads target users whose ideology matches the outlet they subscribe to. Still, even when ads are excluded, the gap between the two groups of outlets remains large and the decomposition does not change substantially.

Reweight Based on Compliance The effect of the algorithm is estimated using two IV estimators, and thus its causal interpretation relies on the assumption that there is no essential heterogeneity (Heckman, Urzua and Vytlacil, 2006). Otherwise, the difference between exposure to posts, conditional on subscriptions, in the pro- and counter-attitudinal treatments might be due to the combination of heterogeneity in the effect of subscribing to outlets and selection into compliance, and not due to differing effects of subscribing to pro- and counter-attitudinal outlets. In

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

the fifth row panel of Appendix Figure A.21, I re-weight the IV estimators, such that participants predicted to comply receive a lower weight. I first calculate the probabilities of compliance with the pro- and counter-attitudinal treatments, by regression compliance on the following covariates using a logit regression: age, female, self-reported ideology, party (dummy variables for Democrat, Republican, and Independent), and the difference between the participant's feelings toward her party and the opposing party. I then predict the probability of compliance for each participant and define the participant's weight as the inverse of the predicted probability.

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

Reweight to Match Population Demographics In the sixth row of the figure, I reweight the participants to match population means on the same set of variables mentioned in the previous section using the entropy weighting procedure. Reweighting decreases the gap in the number of posts observed. When analyzing the results separately for conservatives and liberals, I find that the algorithm's tendency to increase exposure to matching news outlets is driven by the liberals in my sample (I am underpowered to estimate this result precisely) and that could explain the decreased gap in exposure when reweighting the results.[19] Still, there remains a substantial gap in exposure to pro- and counter-attitudinal posts even after reweighing the participants.

Excluding Facebook Usage The effect on Facebook usage is only marginally significant. In the seventh row of Appendix Figure A.21, I assume that the exposure gap only stems from subscriptions and the platform algorithm, and exclude the usage dimension. For this decomposition, I change the calculation of A in Appendix Equation 1, and instead of estimating the effect on the share of posts in the feed, I estimate the effect on the number of posts observed by participant i from outlets in group j.

Decomposition Over Time In the final two rows of Appendix Figure A.21, I decompose the gap in exposure for the first and second week after the intervention. I use the same estimate

[19]The difference between liberal and conservatives could be due to the ideology of participants or differences in the outlets offered.

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

C.7.3 Differential Exposure to Articles Within an Outlet

To estimate whether participants were exposed to news more likely to match their opinions within an outlet, I focus on the subset of articles that were shared on Facebook or Twitter by at least one member of Congress in January to November 2018. I define the slant of an article according to the mean first dimension of the DW-Nominate score of congress members who shared the article (Jeffrey et al., 2020).[20]

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

D Interpretation

How should we interpret the fact that the intervention affected attitudes toward parties, while political opinions remained stable? In this section, I compare two frameworks explaining affective

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

polarization and examine which is most consistent with the data.

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

$$\gamma_i = \sum_{k \in \{1..K\}} w_{ik}\theta_{ik} \tag{2}$$

where $w_{ik} \in \{0, 1\}$ is the weight consumer i places on belief k when determining her political opinion. A weight can be thought of as the priority the consumer places on a specific belief. For example, a consumer's support for a climate bill can depend on two beliefs: the consumer's belief on whether the bill will decrease or increase emissions and the belief on whether the bill will increase or decrease electricity prices. A liberal may place a positive weight only on the effect on emissions and a conservative may place a positive weight only on the effect on prices.[21] A political party uses the same framework and its opinion is a weighted average of various beliefs.

Outlet j receives signal s_{jk} on the state of the world: $s_{jk} \sim N(\theta^*_k, \frac{1}{h_{jk}})$, where θ^*_k is the true state of the world and h_{jk} is the precision of the signal received. Media outlets act as delegates for their consumers by covering issues according to the weights their consumers place on them.[22] Therefore, pro-attitudinal outlets cover issues more when $w_{own} > w_{opposing}$ and counter-attitudinal outlets cover issues more when $w_{opposing} > w_{own}$, where w_{own} are the weights used by the individual's own party and $w_{opposing}$ are the weights used by the opposing party. Indeed, Figure 2 suggests that there is substantial differentiation in the topics news outlets cover. Returning to the climate change example, posts from the outlets offered in the experiment also demonstrate this differential coverage: for every post from a conservative outlet mentioning the words "environment" or "climate," 1.28 posts mentioned the word "economy," while for liberal outlets, the ratio was 0.43.[23]

I assume that consumers exposed to a new outlet update their beliefs in the direction of the outlet. This type of movement is expected if media outlets are biased in their reporting and consumers are naive and do not completely take the bias into account (DellaVigna and Kaplan, 2007).[24]

[21] In the Pew Research Center Political Survey from January 2019, 74% of Democrats stated that the environment should be a top priority for President Trump and Congress in 2019, compared to only 31% of Republicans. On the other hand, 79% of Republicans said the economy should be a top priority, compared to 64% of Democrats (the sample includes respondents leaning toward the Democratic and Republican parties). As a clarifying example for the framework, I intentionally focus on a broad issue, support for climate change policy. Some of the questions forming the political opinions index focus on more specific topics, but the same logic holds. For example, the favorability of the March for Our Lives Movement could depend on participants' belief on whether banning certain weapons will decrease gun violence and their belief on whether the movement will prevent most gun owners from purchasing their preferred guns.

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

[23] This calculation is based on the ratio between the number of times the words "economy", "climate" and "environment" appeared in the messages of all posts shared by the eight primary outlets and first two alternative outlets between February 15, 2018, and December 31, 2018. Duplicate posts with the same message are excluded.

[24] An alternative explanation for why consumers' posteriors move toward the opposing party when exposed to

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

$$A_{ip} = g(\gamma_p - \hat{\gamma}_{ip}) \tag{3}$$

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

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

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

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

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

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

Alternatively, the attitude of a consumer toward a party may depend on whether the political opinion of a party is reasonable according to the party's weights. Hence, the benchmark opinion is the opinion the party would hold according to the consumer's beliefs regarding the state of the world, weighted by the weights party p places on those beliefs. In other words, affective polarization increases when consumers cannot rationalize the parties' political opinions and perceive that

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

the party is not adhering to its values.[25] The change in affective polarization following an update to the consumer's beliefs is:

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

Note that the result is identical to Appendix Equation 4 besides the weights placed on beliefs. Therefore, if the consumer and the party place the same weight on beliefs ($w_{pk} = w_{ik}$), there is no difference between the two theories. However, with heterogeneous weights, political opinions and affective polarization may be differentially affected. In the climate bill example, a liberal who believes the climate bill will mitigate emissions and *decrease* consumer prices will support the bill. The consumer will have a negative attitude toward a party opposing the bill since even if the party places a zero weight on decreasing emissions, it should still support the bill. If the liberal is exposed to conservative outlets and learns that the bill is likely to increase prices, she may still support the bill since she places a positive weight only on mitigating emissions but will develop a less negative attitude toward a party that places a positive weight on consumer prices and thus opposes the bill.[26]

This theory is consistent with the results of the experiment if the consumers updated beliefs on which they place zero weights, but at least one of the parties places positive weights.[27] This would result in consumers' political opinions remaining constant, but attitudes toward parties changing.[28]

To further test these theories, I analyze the effect of the experiment on participants' attitudes toward the opposing party. If affective polarization is simply a function of political distance, attitudes toward parties will be affected when consumer i updates beliefs on which she places positive weights (Appendix Equation 4). Therefore, attitudes toward both parties are more likely to be af-

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

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

[27] It is plausible that as a result of the experiment consumers updated beliefs on which they place zeros weights since they are less likely to have been exposed to counter-attitudinal outlets covering these beliefs. Thus, they are expected to have weaker priors regarding those beliefs. Indeed, Appendix Figure A.4 shows that participants assigned to the counter-attitudinal treatment were more likely to say that they modified their views in the past two months because of something they saw on social media, compared to participants assigned to the pro-attitudinal treatment.

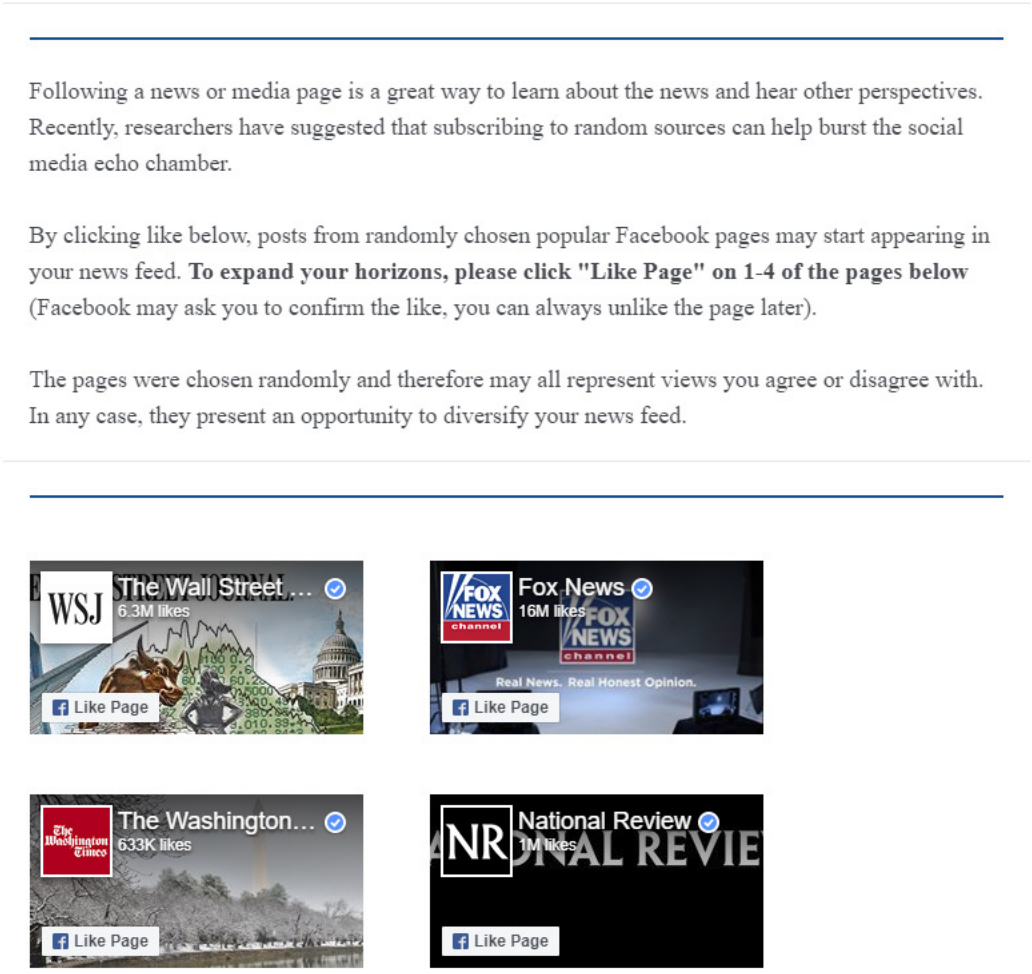
[28] The stability of political opinions relies on a strong assumption that consumers place zero weights on some beliefs or that they determine their political opinions based on lexicographic orderings of beliefs. This assumption is plausible in certain cases. For example, individuals who do not believe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

fected by pro-attitudinal outlets that cover these beliefs. On the contrary, if affective polarization is a function of unreasonable opinions, attitudes toward party p will be affected more by beliefs on which p places positive weights (Appendix Equation 5). As a result, pro-attitudinal outlets are more likely to affect attitudes toward one's own party, while counter-attitudinal outlets are more likely to affect attitudes toward the opposing party. Appendix Table A.17 shows that attitudes toward the opposing party are indeed more likely to be affected by exposure to counter-attitudinal outlets, consistent with the theory that affective polarization is due to opinions that are perceived to be unreasonable.

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

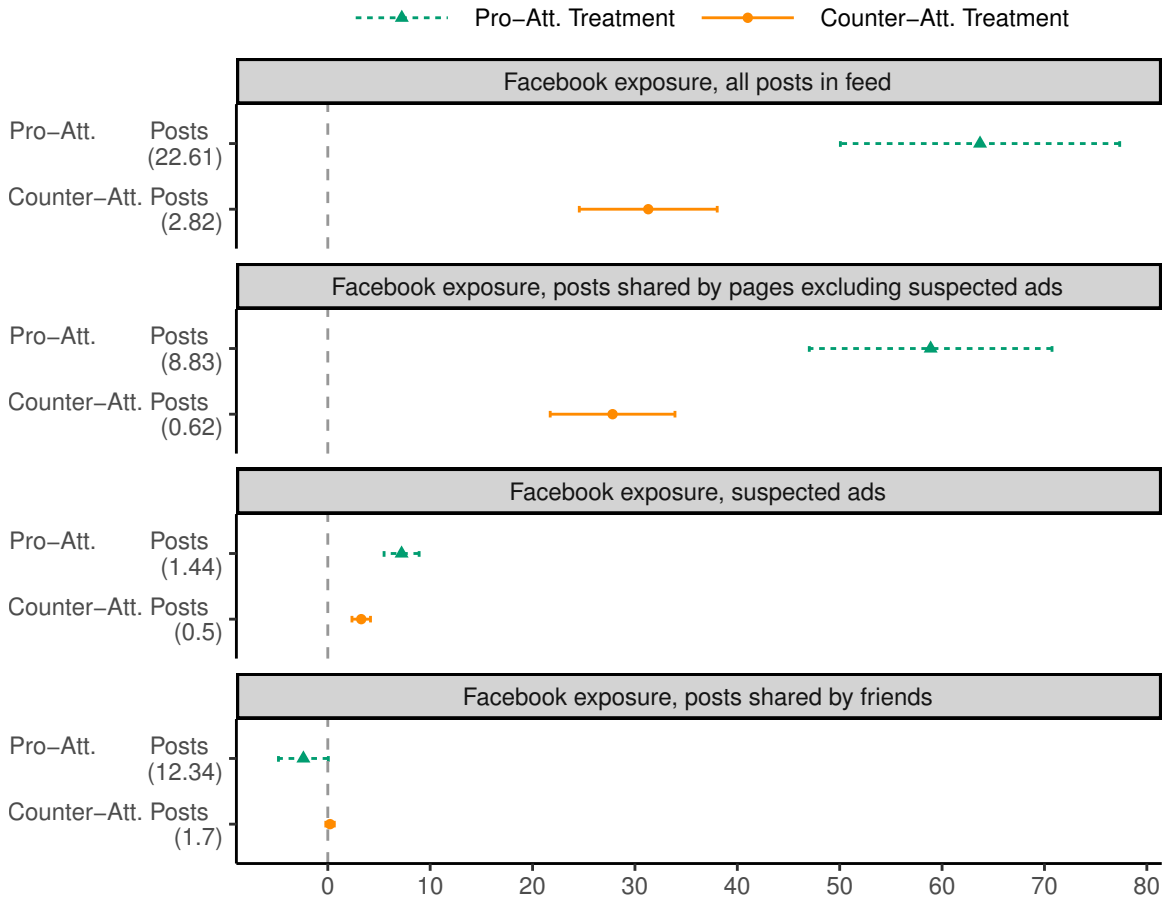
E Additional Figures and Tables

Figure A.1: Example for the Conservative Treatment Intervention



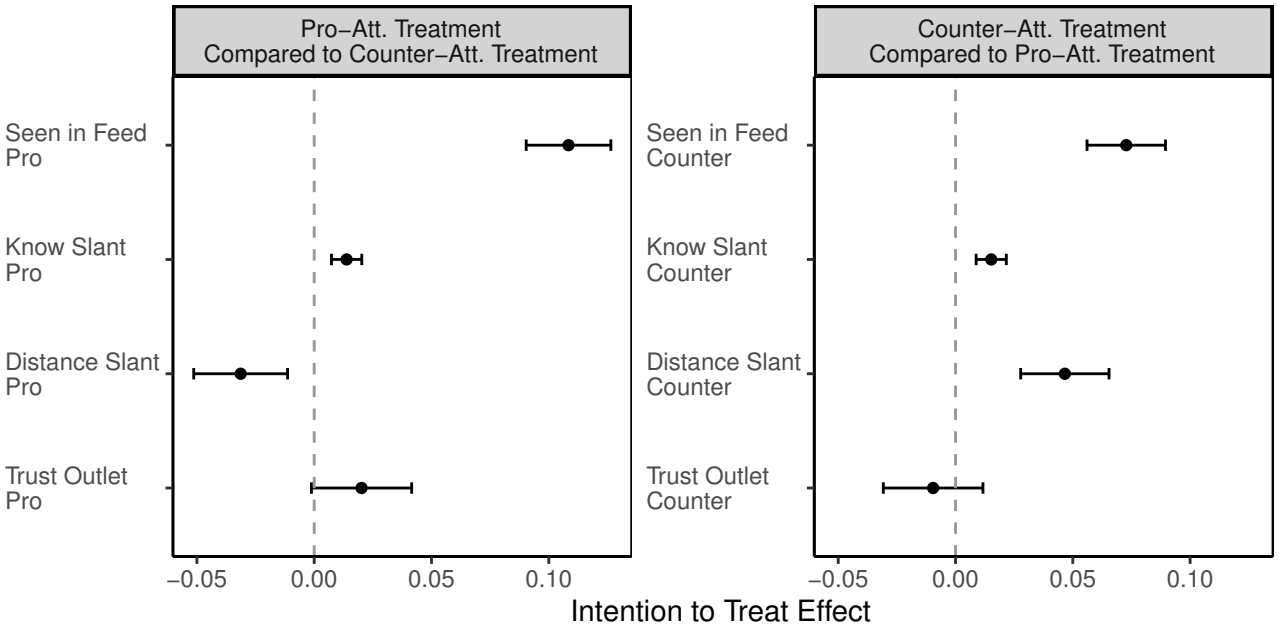
This figure shows a survey page asking participants to subscribe to four conservative outlets. Participants randomly assigned to the conservative treatment, who have not already subscribed to the four primary outlets, were presented with an intervention similar to this figure. The "Like Page" buttons were generated using Facebook's Page Plugin. The image in the background of each button was automatically updated according to the outlet's Facebook page, and the order of the outlets was determined randomly.

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



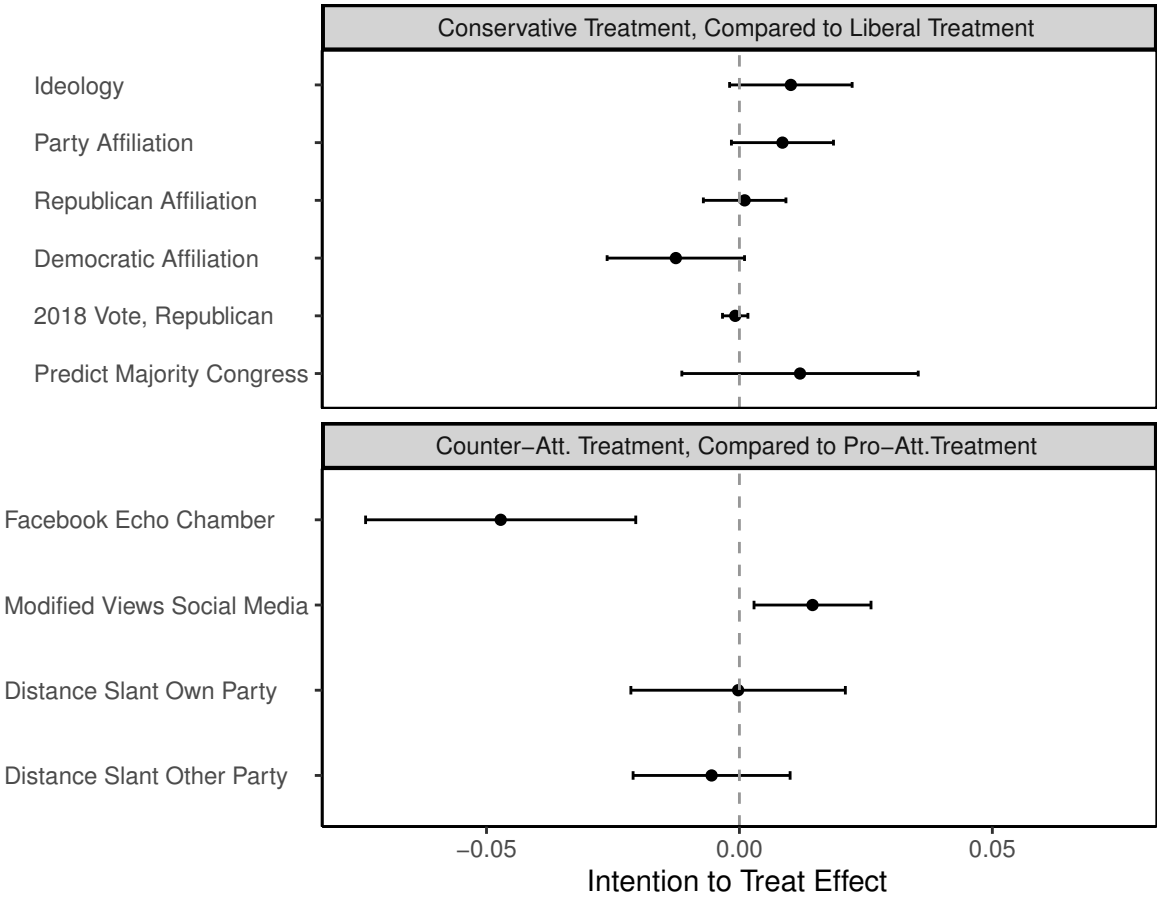
This figure shows the effect of the pro-attitudinal and counter-attitudinal treatments on exposure to posts from the potential outlets in the two weeks following the intervention. The control group mean for each outcome is in parenthesis. The first panel showing total exposure is identical to the second panel of Figure 6. The second panel shows the effect on posts shared by Facebook pages organically. This includes all posts shared by the potential outlets, or other Facebook pages referring to the potential outlets, besides posts which are likely to be sponsored (ads). The third panel shows the effect on exposure to suspected ads related to the outlets. The fourth panel shows the effect on posts shared by Facebook friends. Appendix A.3 explains how ads were identified. Error bars reflect 90 percent confidence intervals.

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



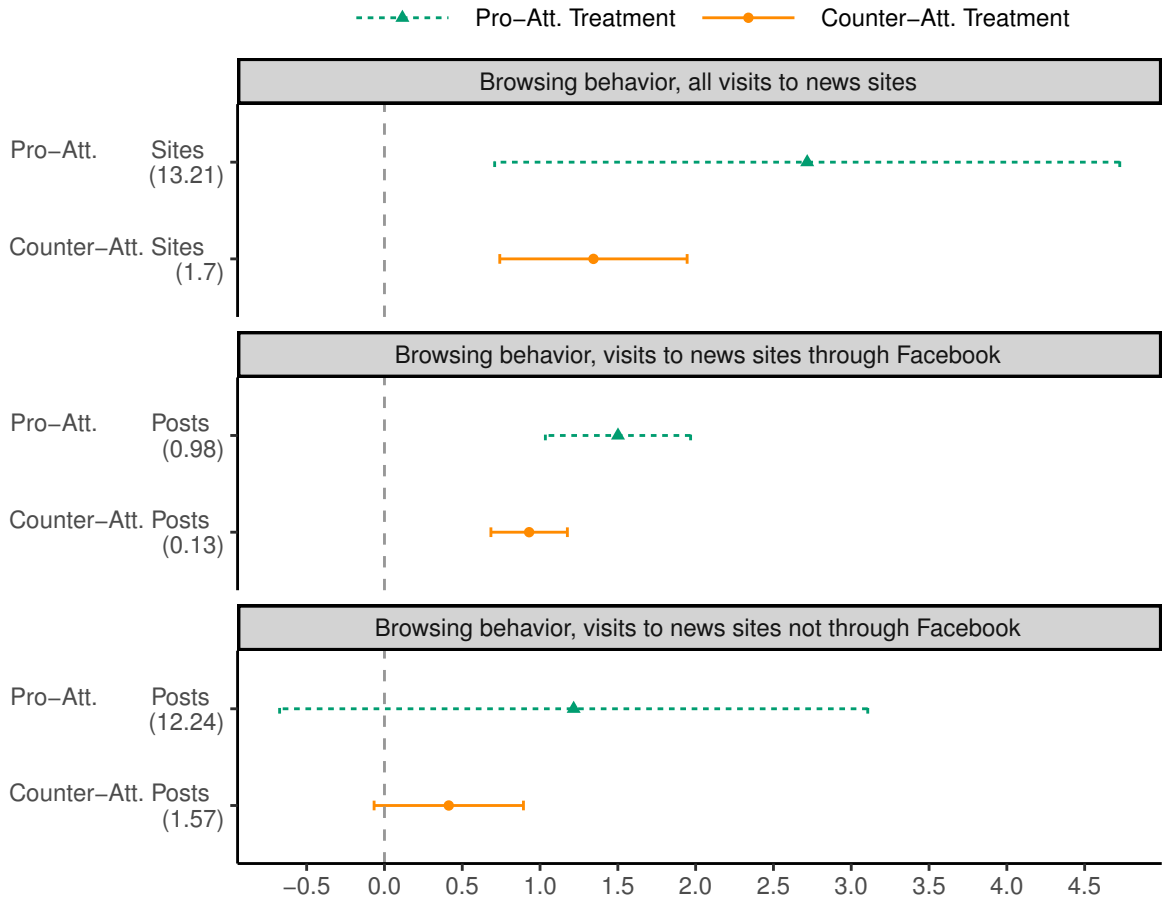
This figure shows the effects of the treatments on attitudes toward the potential outlets. Each row represents a regression pooling the opinions of participants in the endline survey on the eight potential outlets defined for each participant. *Seen in Feed* is whether the participant reported seeing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

Figure A.4: Effects of the Treatments on Additional Survey Outcomes



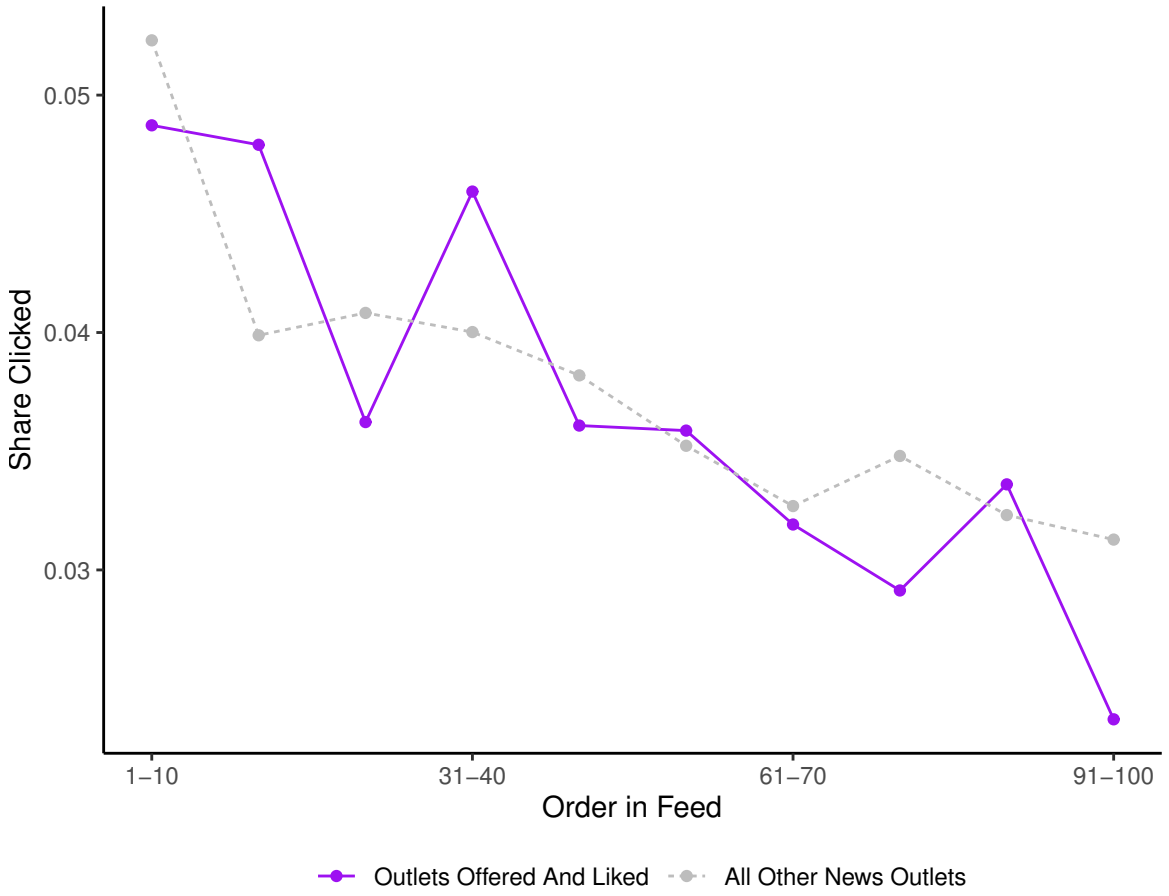
This figure shows the effect of the experiment on additional endline survey outcomes. *Ideology* is self-reported on a 7-point scale. *Party Affiliation* is the party the participant identifies with on a 7-point scale. *Republican/Democrat Affiliation* is whether the participant is a strong Republican/Democrat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

Figure A.5: Effects of the Pro- and Counter-Attitudinal Treatments on News Sites Visited, by Source



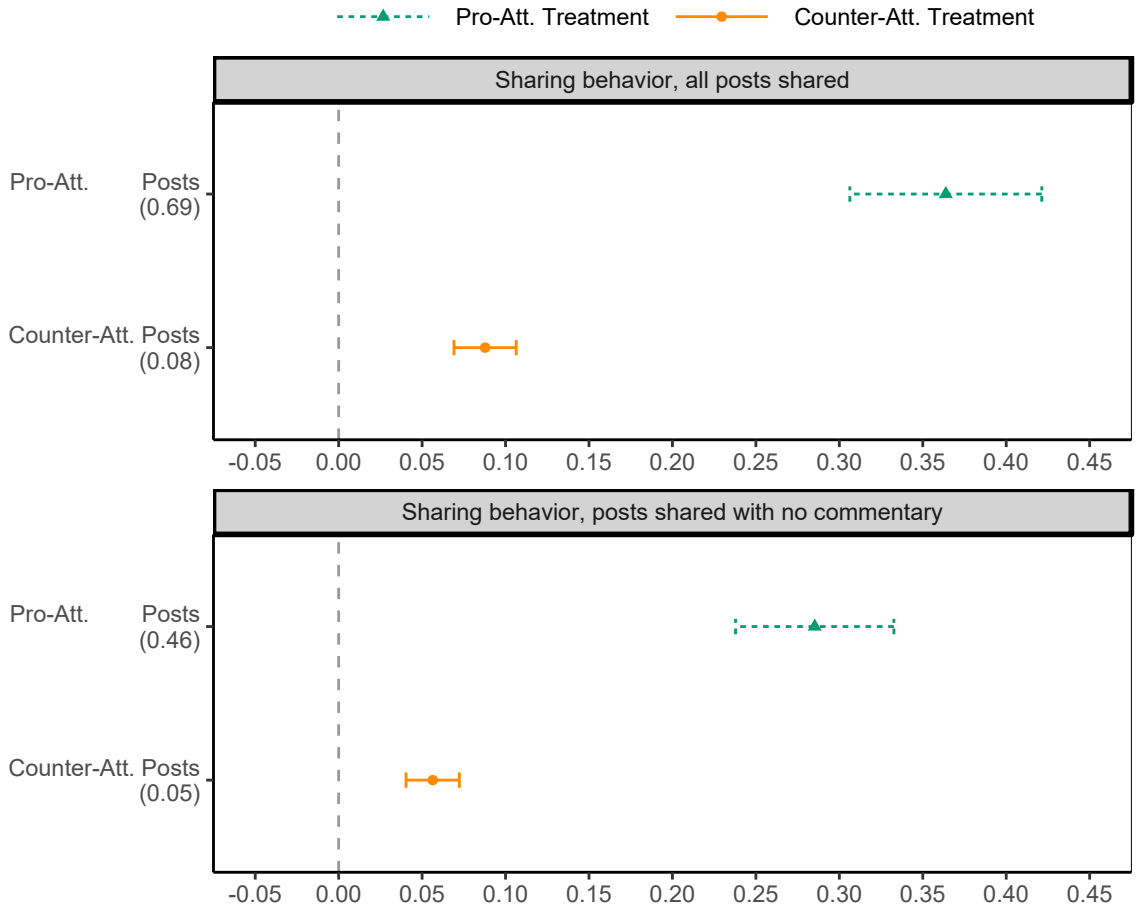
This figure shows the effect of the pro- and counter-attitudinal treatments on total visits to the potential outlets' websites in the two weeks following the intervention. The control group mean for each outcome is in parenthesis. The first panel showing total visits is identical to the third panel of Figure 6. The second panel shows the effect on visits to websites that could be matched with a URL appearing in a Facebook post. The third panel shows the effect on all other visits. Appendix Section A.2 explains how posts were matched with visits to news sites. Error bars reflect 90 percent confidence intervals.

Figure A.6: Share of Links Visited by Order in Feed



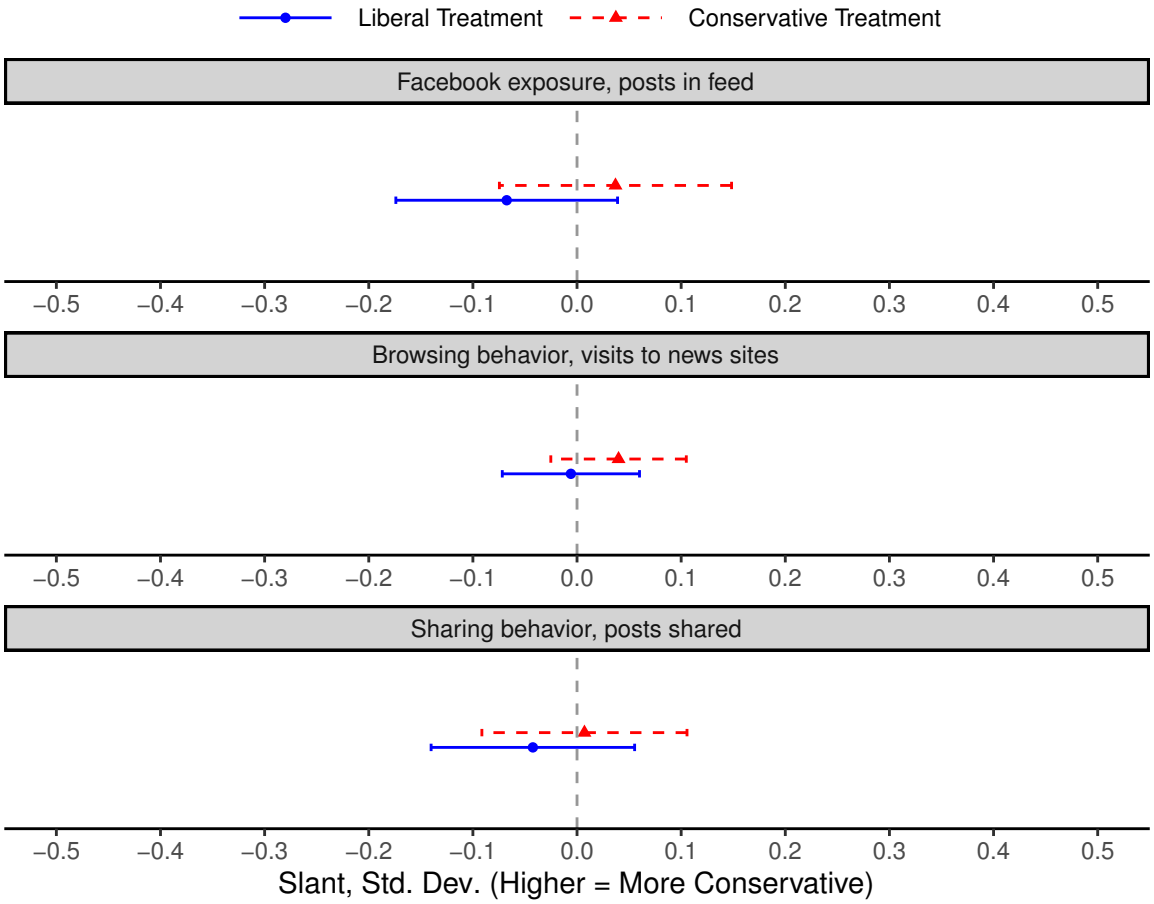
This figure shows the share of links which were visited by participants. The data include all posts with links from the pages of leading news outlets, excluding suspected ads, in the two weeks following the intervention. To determine the order of posts, a Facebook feed session is defined to begin when a participant views a post on Facebook at least 30 minutes after viewing a previous post. To smooth the results, posts are grouped into groups of ten based on their order. Appendix A.2 explains how posts were matched with visits to news sites and Appendix A.3 explains how suspected ads were identified.

Figure A.7: Effects of the Pro- and Counter-Attitudinal Treatments on Number of Posts Shared, Access Posts Subsample



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

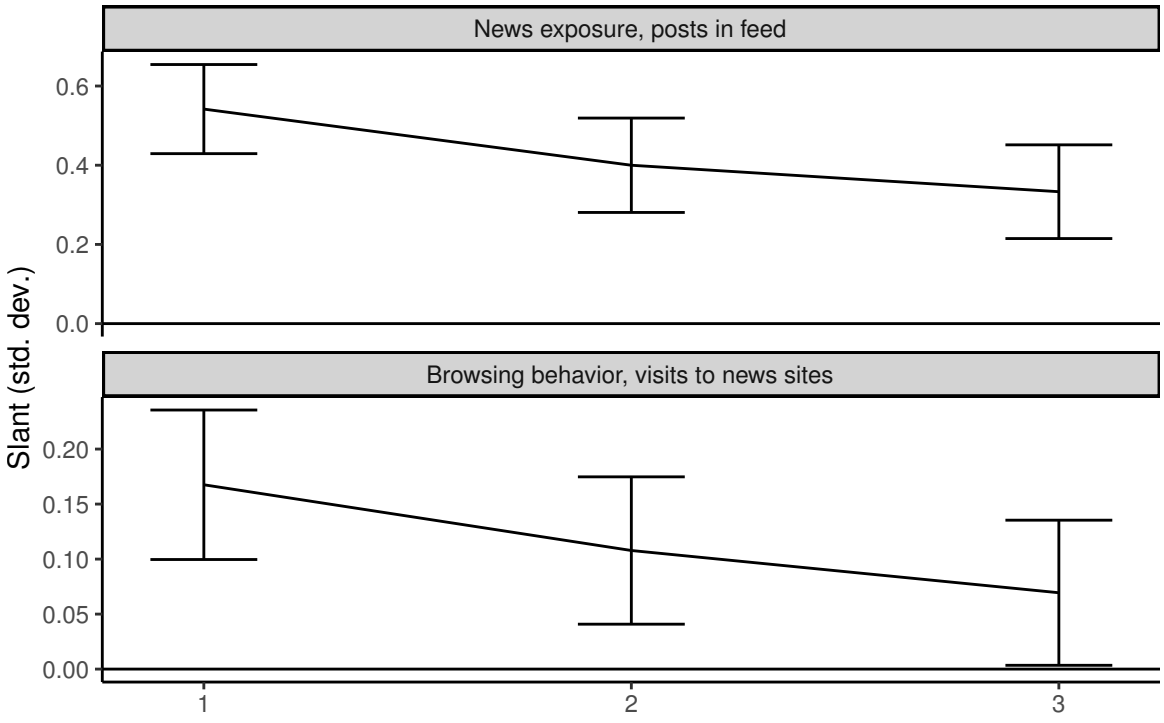
Figure A.8: Effect of the Liberal and Conservative Treatments on Slant, Excluding Each Participant's Eight Potential Experimental Outlets



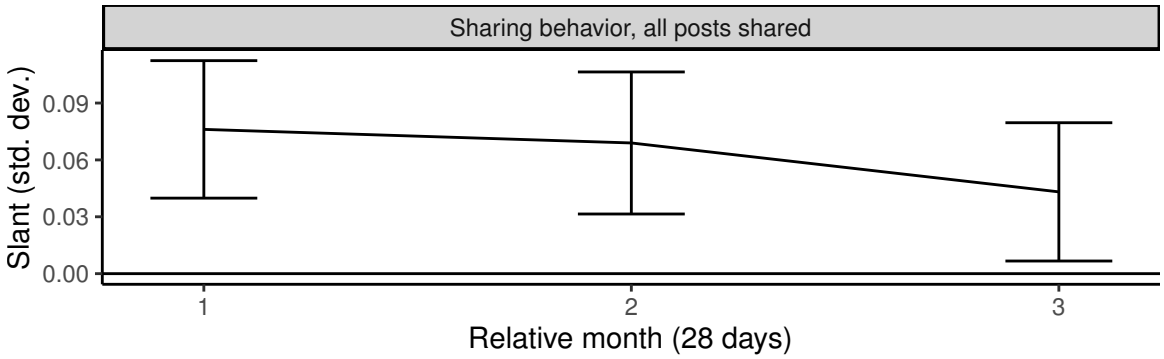
This figure shows the effect of the liberal and conservative treatments on the mean slant, in standard deviations, of all news participants engaged with, excluding the four potential liberal outlets and the four potential conservative outlets defined for each participant. The regressions control for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

Figure A.9: Effects of the Conservative Treatment on Mean Slant by Month, Compared to the Liberal Treatment

(a) News Exposure and Browsing Behavior

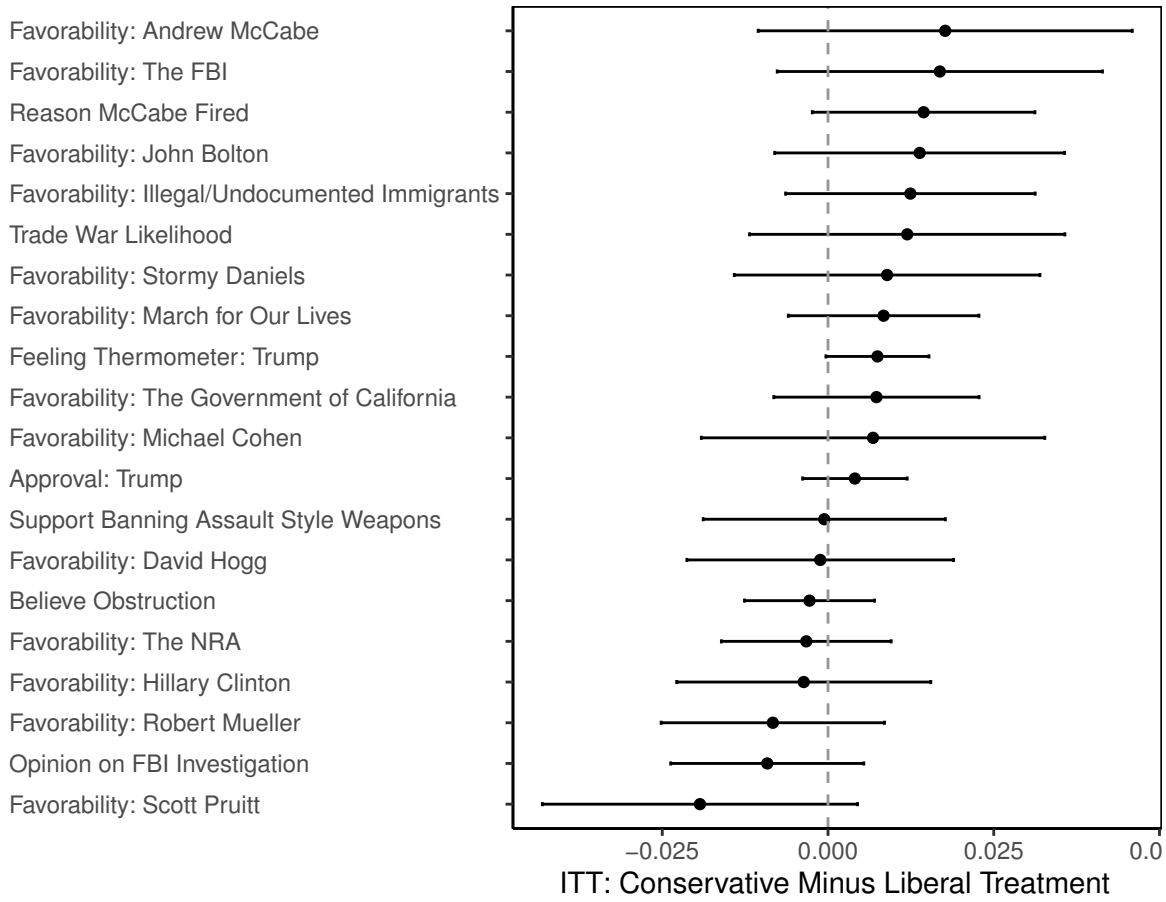


(b) Sharing Behavior



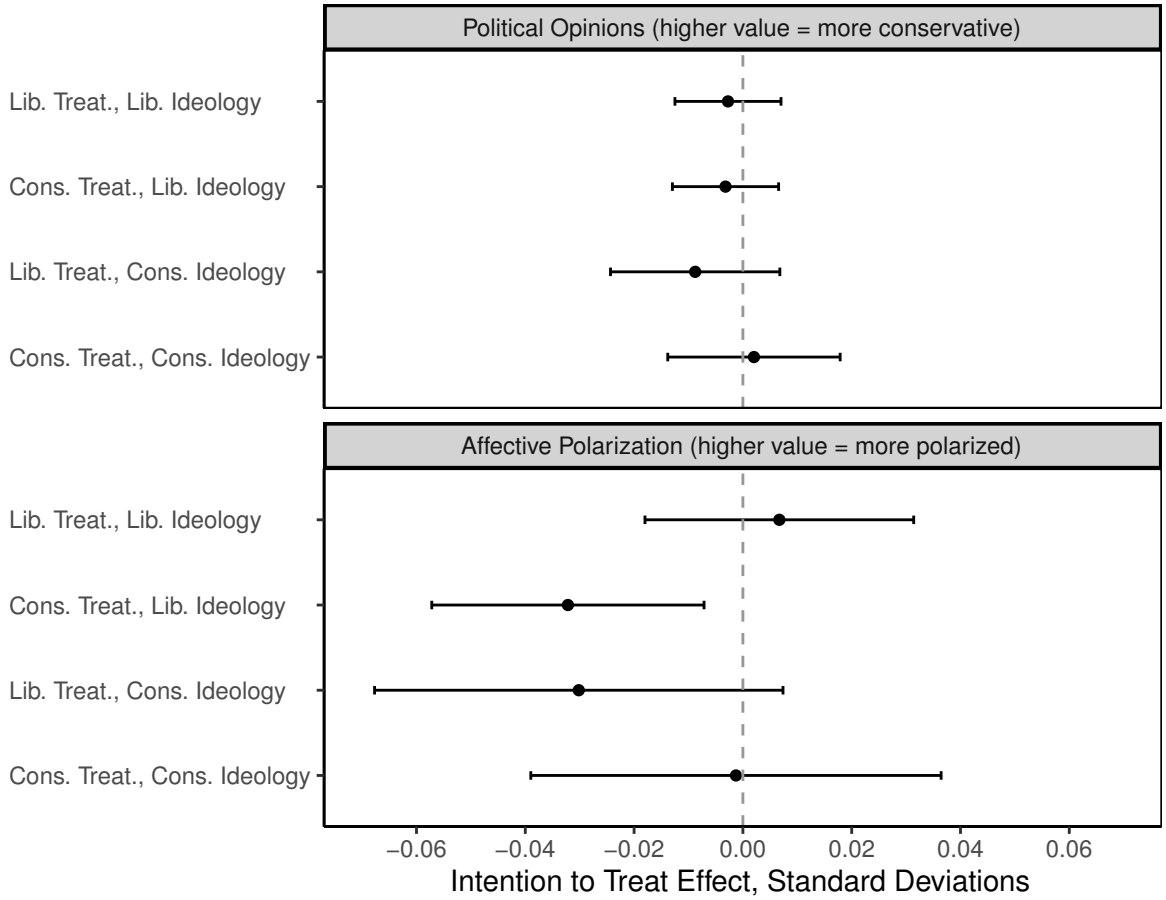
These figures show the difference between the effect of the liberal and conservative treatments on the mean slant over time. Each panel presents a series of regressions, where the dependent variable is the slant of outlets in a specific month. In the x-axis, relative month 1 is defined as 28 days immediately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

Figure A.10: Effects on Components of the Political Opinion Index



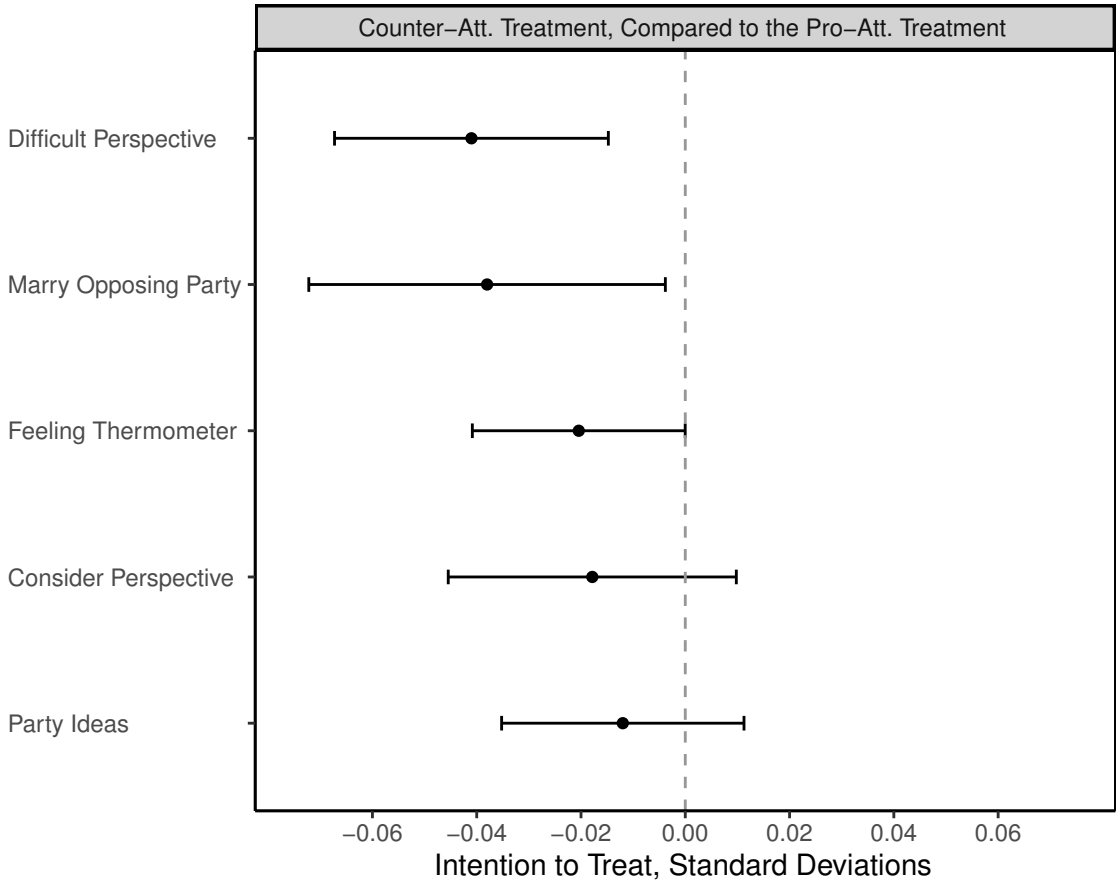
This figure shows the effect of the conservative treatment, compared to the liberal treatment on outcomes composing the political opinions index. Each row represents a separate regression as specified in Section II.E. Outcomes are defined such that a higher value is associated with a more conservative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

Figure A.11: Effects of the Treatments on Primary Outcomes, by Ideological Leaning



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

Figure A.12: Effects of the Treatments on Components of the Affective Polarization Index



This figure shows the effect of the counter-attitudinal treatment on the measures composing the affective polarization index, compared to the pro-attitudinal treatment. Each row presents the result of a regression estimating the effect of the treatment on one dependent variable where a higher value is associated with a more polarized outcome. *Difficult Perspective* and *Consider Perspective* measure political empathy. The former is the difference in how difficult it is to see things from each party's point of view, and the latter is the difference in how important it is to consider the perspective of each party. *Marry Opposing Party* is how participants would feel if their son/daughter married someone from the opposing party. *Feeling Thermometer* is the difference in how warm participants feel toward each party. *Party Ideas* is the difference in how many good ideas each party is perceived to have. The outcomes are described in more detail in Section II.D.2 and the regressions are specified in Section II.E. Error bars reflect 90 percent confidence intervals.

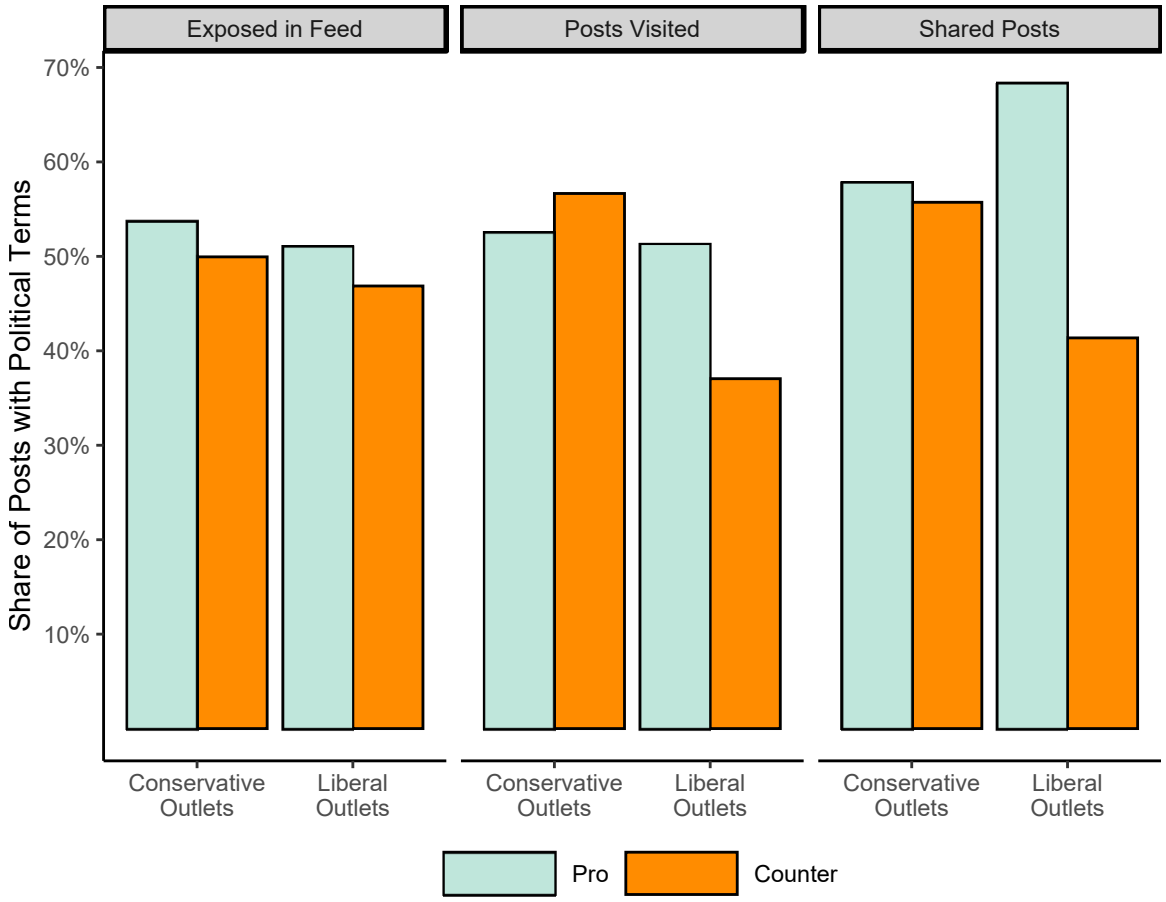
Figure A.13: Recruitment Ads

(a) Political Ad

(b) General Ad

These figures are examples of the ads used to recruit participants.

Figure A.14: Share of Posts Mentioning Political Terms



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

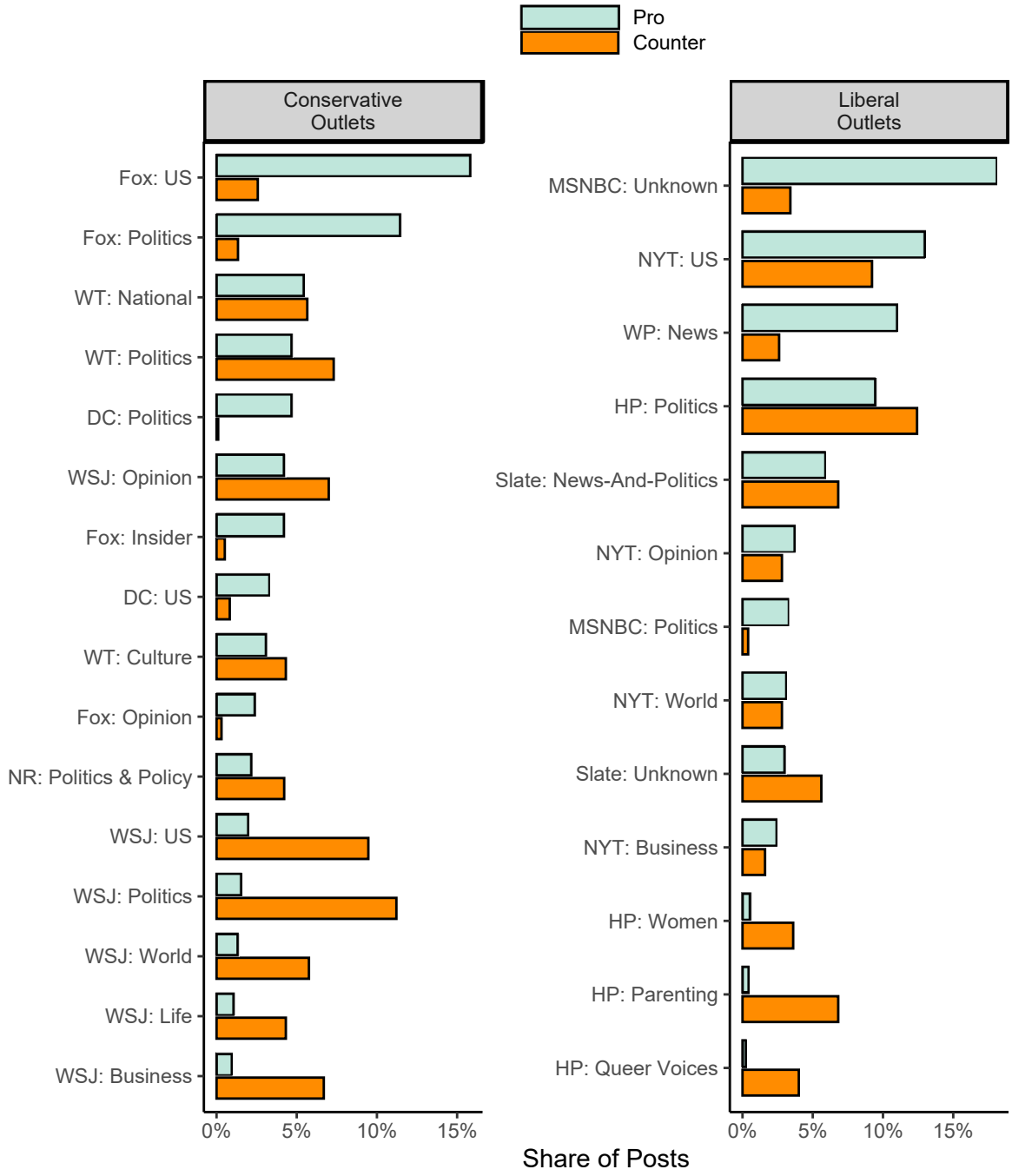
Figure A.15: Links in Posts Observed in the Feed, by Outlet and Section

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

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

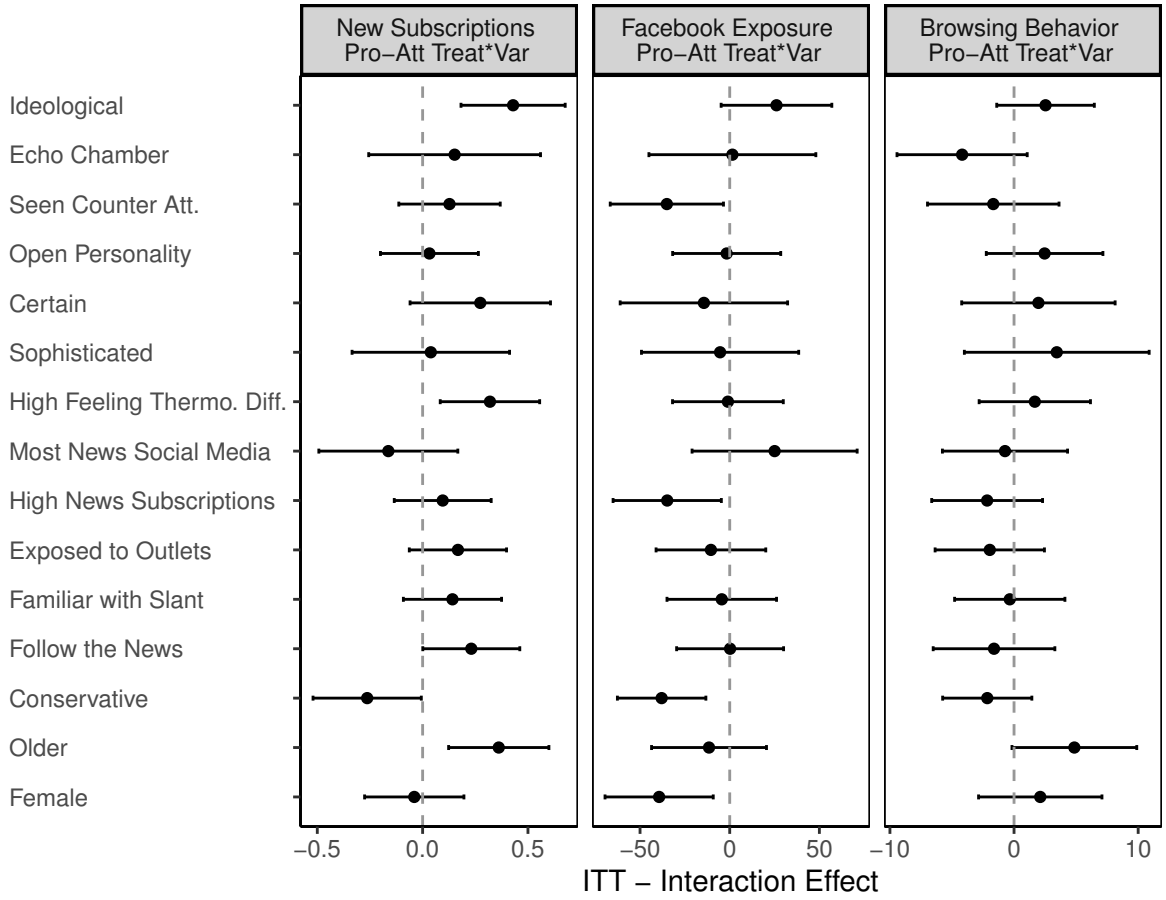
This figure shows the most common outlets and sections participants visited through links shared by the outlets they subscribed to. For more details see Figure A.15.

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



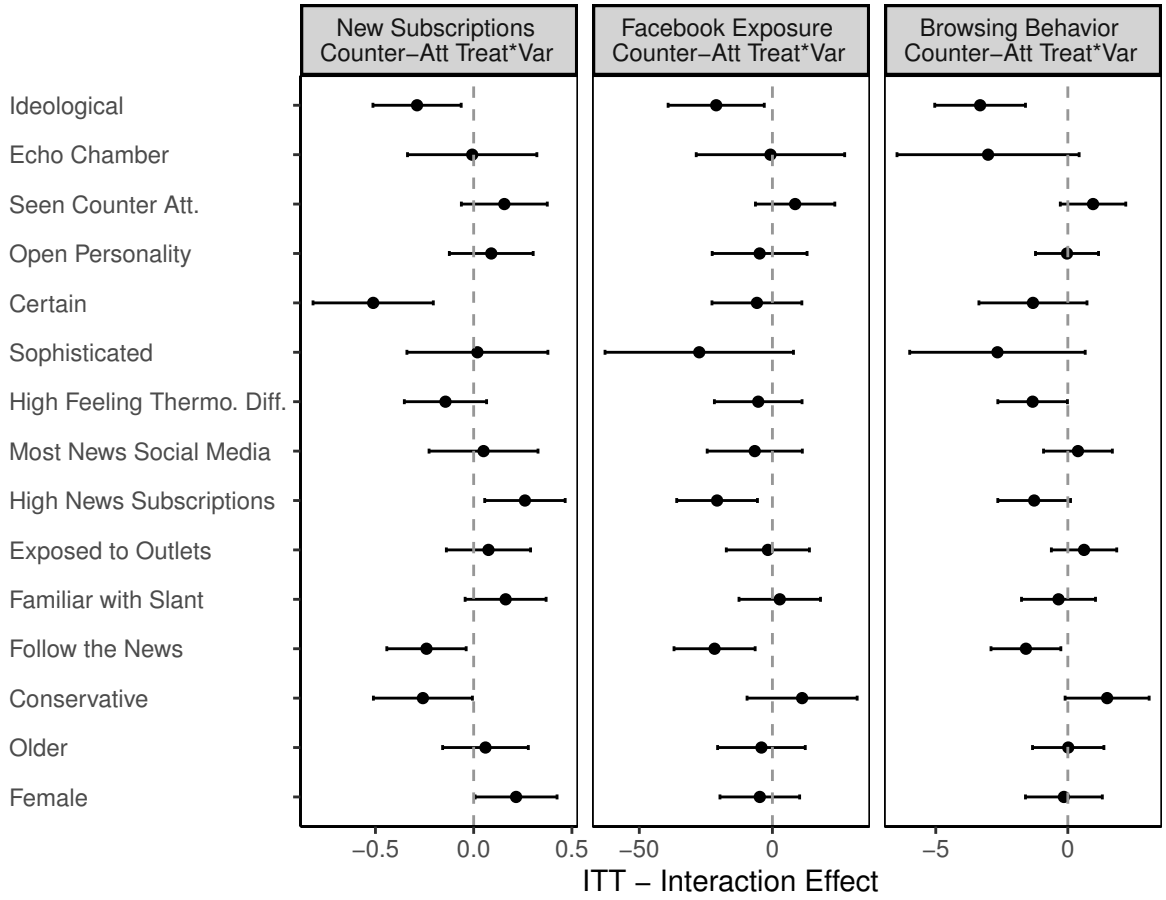
This figure shows the most common outlets and sections of the links participants shared when sharing posts from the outlets they subscribed to. For more details see Figure A.15.

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



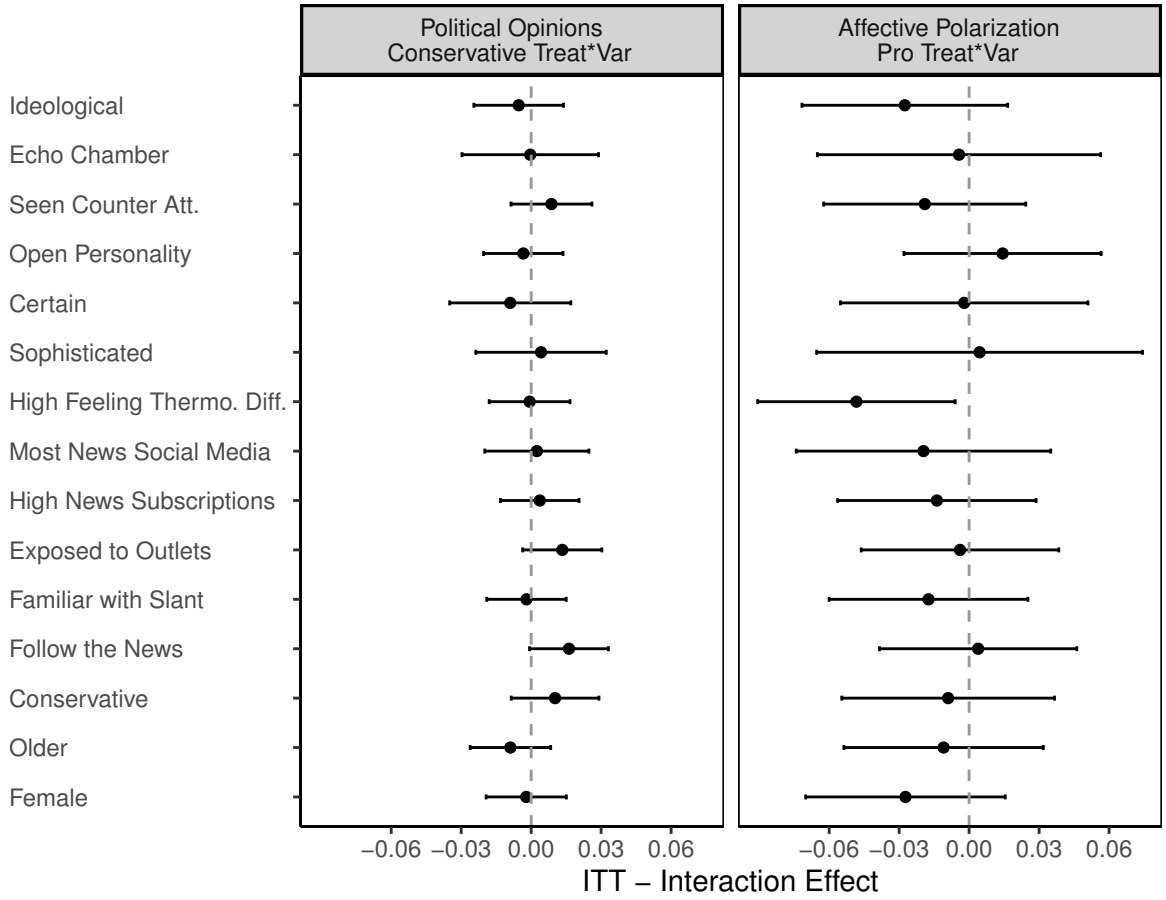
This figure shows heterogeneous effects of the pro-attitudinal treatment on engagement with the pro-attitudinal outlets. Each row presents the β coefficient in the following regression:
$Y_i = \alpha T_i^P + \beta T_i^P \times Var + \gamma Var + \delta X_i + \varepsilon_i$,
where the dependent variables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

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



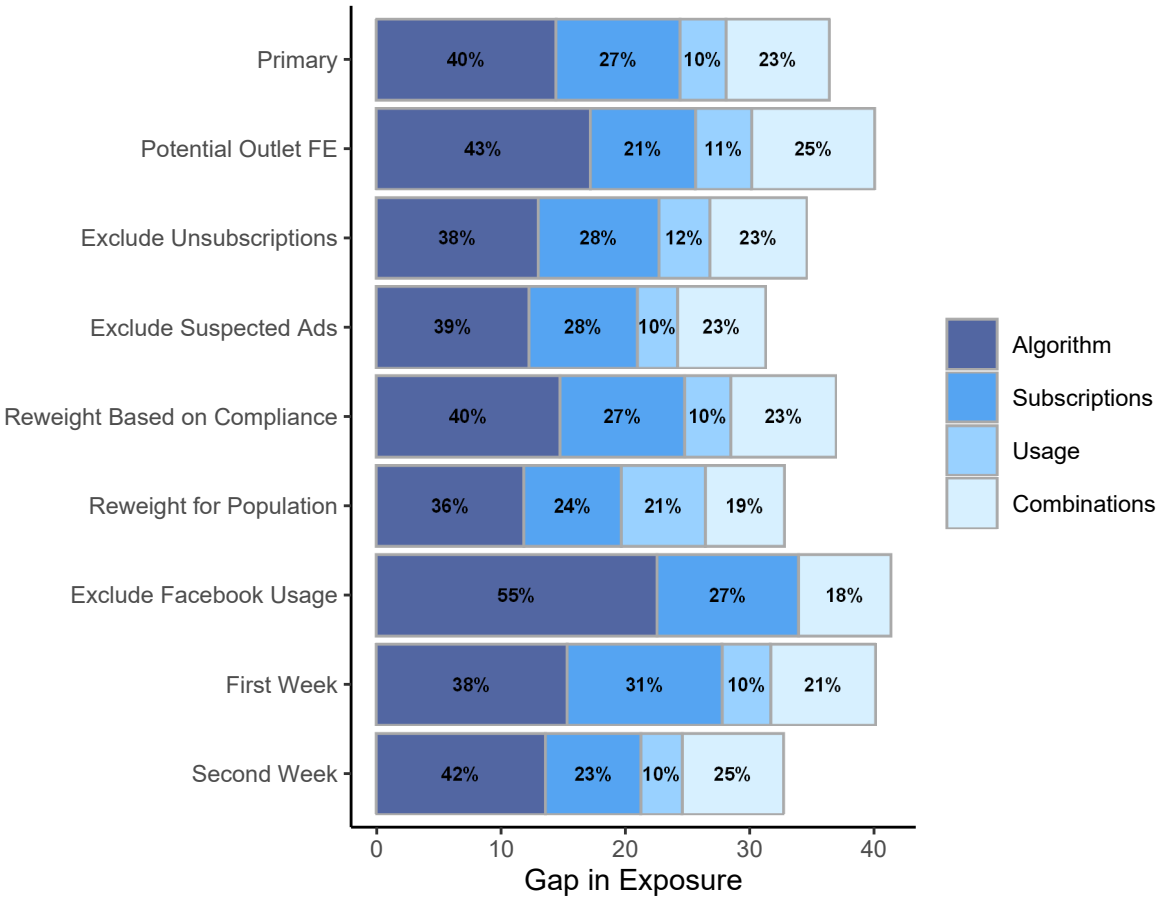
This figure shows heterogeneous effects of the counter-attitudinal treatment on engagement with the counter-attitudinal outlets. Each row presents the β coefficient in the following regression:
$Y_i = \alpha T_i^A + \beta T_i^A \times Var + \gamma Var + \delta X_i + \varepsilon_i$,
where the dependent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

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



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

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



This figure decomposes the gap between the number of posts participants were exposed to from the offered pro- and counter-attitudinal outlets. The first row repeats the main specification described in Figure 10. The second row controls for the potential outlets defined for each participant. The third row defines subscriptions as subscribing to the outlet for at least two weeks. The fourth row excludes posts that are likely to be sponsored (ads). The fifth row reweights the participants in each treatment such that the compliers resemble the entire sample. The sixth row reweights the participants such that the entire sample resembles the US population. The seventh row excludes differences in usage between the groups. The final two rows decompose the results separately for the first and second week after the intervention. Each row is described in more detail in Section C.7.2.

Table A.1: Outlets Offered

Outlet	Treatment	Slant	Potential	Offered	Sub.
The Washington Times	Conservative	0.70	37,120	12,366	3,278
The National Review	Conservative	0.90	36,168	12,057	2,953
The Wall Street Journal	Conservative	0.28	35,406	11,805	4,059
Fox News	Conservative	0.78	32,566	10,842	1,425
The Daily Caller	Conservative	0.87	4,522	1,471	323
Washington Examiner	Conservative	0.82	1,719	607	133
The Western Journal	Conservative	0.90	1,531	509	153
Townhall	Conservative	0.93	397	135	37
The Blaze	Conservative	0.89	221	80	25
The Conservative Tribune	Conservative	0.89	204	72	34
Newsmax	Conservative	0.77	114	32	14
Slate	Liberal	-0.68	35,206	11,738	3,008
MSNBC	Liberal	-0.81	35,091	11,688	2,786
HuffPost	Liberal	-0.62	31,927	10,643	2,359
The New York Times	Liberal	-0.55	30,337	10,145	3,376
Washington Post	Liberal	-0.26	8,234	2,824	1,341
Salon	Liberal	-0.88	5,119	1,668	595
Daily Kos	Liberal	-0.90	2,015	661	232
The Atlantic	Liberal	-0.54	636	203	116
Mother Jones	Liberal	-0.87	515	150	59
NPR	Liberal	-0.61	431	119	70
The New Yorker	Liberal	-0.76	317	105	65
PBS	Liberal	-0.54	134	40	23

This table shows the list of outlets included in the experiment. *Slant* is the outlet's slant, ranging from -1 to 1 (Bakshy, Messing and Adamic, 2015). *Potential* is the number of participants for whom the outlet was defined as a potential outlet. *Offered* is the number of participants who were offered to subscribe to the outlet, based on their treatment assignment. *Sub.* is the number of participants who subscribed to each outlet in the intervention. The first four liberal outlets and the first four conservative outlets are the primary outlets offered in the experiment and the rest of the outlets are the alternative outlets offered if a participant already subscribed to a primary outlet.

Table A.2: Descriptive Statistics by Sample

		Baseline Sample	Access Posts Subsample	Endline Survey Subsample	Extension Subsample
1)	Ideology (-3, 3)	-0.61	-0.61	-0.71	-0.95
2)	Ideology, Abs. Value (0, 3)	1.75	1.75	1.80	1.81
3)	Democrat	0.38	0.38	0.40	0.44
4)	Republican	0.17	0.17	0.16	0.14
5)	Independent	0.37	0.36	0.36	0.36
6)	Feeling Therm., Difference	50.22	50.27	50.32	51.08
7)	Difficult Pers., Difference	1.92	1.92	1.96	1.92
8)	Most News Social Media	0.18	0.18	0.17	0.16
9)	Took Survey Mobile	0.67	0.67	0.63	0.00
10)	Female	0.52	0.52	0.52	0.49
11)	Age	47.69	47.65	48.78	52.47
12)	Total Subscriptions	474	474	472	481
13)	News Outlets Subscriptions	8.11	8.11	8.28	8.61
14)	Compliance	0.53	0.53	0.58	0.76
15)	N	37,494	34,592	17,635	1,835

This table presents descriptive statistics by subsample. *Baseline Sample* includes all participants. *Access-Posts Subsample* includes participants who provided access to posts they shared for at least two weeks. *Endline Survey Subsample* includes participants who completed the endline survey. *Extension Subsample* includes participants who installed the browser extension for at least two weeks. *Ideology, Abs. Value* is the absolute value of self-reported ideology. *Feeling Therm., Difference* is the difference between feelings toward the 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 point of view of the opposing party and their own party. *News Outlets Subscriptions* is subscriptions to pages of leading news outlets. For all other variables, see Table 2.

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

Variable	Mean		Difference		
	Sample N=36,330	US	Control - Pro.	Control - Counter.	Pro. - Counter.
Baseline Survey					
Ideology, Abs. Value (0, 3)	1.80	1.31	0.00	-0.00	-0.00
Democrat	0.39	0.37	0.01	0.00	-0.01
Republican	0.17	0.30	0.00	-0.01	-0.01
Independent	0.36	0.29	-0.01*	0.00	0.01**
Vote Support Clinton	0.54		-0.00	-0.00	0.00
Vote Support Trump	0.27		0.00	0.00	0.00
Feeling Therm., Difference	50.22	38.44	0.36	0.41	0.05
Difficult Pers., Difference	1.92		0.03	0.02	-0.02
Facebook Echo Chamber	1.20		0.00	-0.01	-0.01
Follows News	3.36	2.48	0.01	0.01	0.01
Most News Social Media	0.17	0.12	0.00	-0.00	-0.01
Device					
Took Survey Mobile	0.67		-0.01*	-0.00	0.01*
Facebook					
Female	0.52	0.52	-0.01	-0.00	0.00
Age	47.91	47.70	0.02	0.08	0.06
Total Subscriptions	473		6.91	3.16	-3.75
News Outlets Slant, Abs. Value	0.54		-0.00	-0.00	0.00
Access Posts, Pre-Treat.	0.98		0.00	0.00	-0.00
Attrition					
Took Followup Survey	0.47		0.03***	0.03***	0.00
Access Posts, 2 Weeks	0.92		0.01	0.00	-0.00
Extension Install, 2 Weeks	0.05		0.00	-0.00	-0.00
F-Test			1.23	0.80	0.99
P-value			[0.20]	[0.75]	[0.48]

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

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

Variable	Mean			Difference		
	Sample N=17,635	US	FB Users	Control - Lib.	Control - Cons.	Cons. - Lib.
Baseline Survey						
Ideology (-3, 3)	-0.71	0.17		-0.01	-0.02	0.01
Democrat	0.40	0.35	0.30	0.01	0.01	0.01
Republican	0.16	0.28	0.21	0.00	0.00	0.00
Independent	0.36	0.32	0.35	-0.02*	-0.01	-0.01
Vote Support Clinton	0.55			-0.00	-0.00	-0.00
Vote Support Trump	0.25			0.01	-0.00	0.01
Feeling Therm., Rep.	27.54	43.06		0.20	-0.04	0.24
Feeling Therm., Dem.	47.79	48.70		0.43	0.68	-0.25
Difficult Pers., Rep. (1, 5)	3.18			0.04	0.01	0.04
Difficult Pers., Dem. (1, 5)	2.35			-0.01	-0.03	0.03
Facebook Echo Chamber	1.20		1.12	0.01	-0.01	0.01
Follows News	3.38	2.42		0.02	0.02	-0.00
Most News Social Media	0.17	0.13		-0.01**	-0.00	-0.01*
Device						
Took Survey Mobile	0.63			-0.01	0.01	-0.01
Facebook						
Female	0.52	0.52	0.55	-0.01	-0.00	-0.00
Age	48.78	47.30	42.86	0.55*	-0.31	0.86**
Total Subscriptions	472			2.37	15.27	-12.90
News Outlets Slant (-1, 1)	-0.20			0.00	-0.01	0.01
Access Posts, Pre-Treat.	0.98			0.00	0.00*	-0.00
F-Test				1.15	0.97	1.32
P-Value				[0.29]	[0.49]	[0.16]

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

Table A.5: Balance Table, Pro- and Counter-Attitudinal Treatments, Among Participants Who Completed the Endline Survey

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

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

Table A.6: Descriptive Statistics by Compliance

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

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

Table A.7: Segregation in News Engagement

(a) Comscore Data

Category	Share	Seg.	Slant, Abs.
1) All Browsing		0.190	0.264
2) Direct	49.9%	0.213	0.263
3) Social	5.1%	0.280	0.358
4) Search	37.3%	0.176	0.286
5) Other	7.6%	0.216	0.300
6) FB	4.2%	0.287	0.354
7) Not FB	95.8%	0.188	0.263

(b) Extension Data

Category	Share	Seg.	Slant, Abs.	Isol.	Cong.	Share Counter
1) Subscribed		0.361	0.554	0.513	0.519	0.118
2) FB Feed		0.211	0.373	0.219	0.320	0.196
3) Friends	48.2%	0.162	0.318	0.153	0.257	0.230
4) Pages	40.7%	0.283	0.449	0.366	0.398	0.153
5) Ads	11.2%	0.255	0.400	0.270	0.320	0.192
6) Browsing		0.197	0.329	0.165	0.260	0.218
7) Not FB	85.4%	0.197	0.324	0.143	0.250	0.222
8) FB	14.6%	0.222	0.361	0.252	0.308	0.203
9) Friends	53.3%	0.203	0.331	0.176	0.265	0.219
10) Pages	36.7%	0.297	0.439	0.429	0.395	0.154
11) Ads	10.0%	0.229	0.379	0.196	0.310	0.171
12) Shared		0.255	0.414	0.307	0.363	0.181

These tables display segregation measures for online and social media news engagement. Sub-table (a) is based on 2017-2018 Comscore data and sub-table (b) is based on data from control group participants in the extension subsample from the first eight weeks after the extension was installed. The segregation measures are defined in Appendix B.1. For more details on how Facebook data were processed and suspected ads were identified see Appendix A.3.

Table A.8: Segregation in Browsing Behavior, Additional Results

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

Category	Share	Seg.	Slant, Abs.
1) All Browsing		0.194	0.244
2) Direct	45.3%	0.217	0.252
3) Social	27.6%	0.260	0.321
4) Search	21.7%	0.147	0.252
5) Other	5.4%	0.224	0.290
6) FB	26.3%	0.264	0.325
7) Not FB	73.7%	0.186	0.236

(b) Segregation Measures Over Time, Comscore Data

Category	Share	Seg.	Slant, Abs.
1) All: 2007-2008		0.174	0.256
2) All: 2017-2018		0.190	0.264

These tables display additional results on segregation in browsing behavior. Sub-table (a) includes only individuals in the Comscore panel who visited multiple news sites through Facebook and through other sources. Sub-table (b) includes the 2007-2008 and 2017-2018 Comscore panels. The segregation measures are defined in Appendix B.1.

Table A.9: Segregation in News Engagement, Visit-Level

(a) Comscore

Category	Share	Seg.	Slant, Abs.
1) All Browsing		0.348	0.412
2) Direct	65.5%	0.359	0.424
3) Social	7.3%	0.412	0.500
4) Search	20.0%	0.264	0.352
5) Other	7.3%	0.318	0.380
6) FB	6.0%	0.422	0.513
7) Not FB	94.0%	0.342	0.406

(b) Extension Data

Category	Share	Seg.	Slant, Abs.	Isol.	Cong.	Share Counter
1) Subscribed		0.454	0.624	0.573	0.520	0.104
2) FB Feed		0.315	0.476	0.284	0.387	0.124
3) Friends	35.8%	0.290	0.434	0.197	0.325	0.154
4) Pages	55.8%	0.331	0.504	0.458	0.428	0.107
5) Ads	8.4%	0.303	0.474	0.305	0.380	0.113
6) Browsing		0.300	0.430	0.216	0.321	0.153
7) Not FB	90.3%	0.297	0.424	0.191	0.312	0.157
8) FB	9.7%	0.323	0.485	0.373	0.405	0.113
9) Friends	43.1%	0.288	0.436	0.222	0.331	0.145
10) Pages	50.2%	0.359	0.536	0.571	0.478	0.086
11) Ads	6.6%	0.233	0.410	0.168	0.332	0.120
12) Shared		0.318	0.457	0.414	0.368	0.158

These tables display segregation measures based on visit-level data instead of aggregating data first at the user-level. In these tables users who visit more websites implicitly receive more weight. Sub-table (a) is based on 2017-2018 Comscore data and sub-table (b) is based on data from control group participants in the extension subsample from the first eight weeks after the extension was installed. The segregation measures are defined in Section III.

Table A.10: Effects of the Treatments on News Exposure, News Sites Visited and Sharing Behavior, Two Weeks Following the Intervention, Poisson Regression

	Pro-Att. Outlets Facebook Exposure	Pro-Att. Outlets Browsing Behavior	Pro-Att. Outlets Sharing Behavior	Counter-Att. Outlets Facebook Exposure	Counter-Att. Outlets Browsing Behavior	Counter-Att. Outlets Sharing Behavior
	(1)	(2)	(3)	(4)	(5)	(6)
Pro-Att. Treat.	1.34***	0.29**	0.57***	0.33**	0.19	0.17
	(0.13)	(0.14)	(0.21)	(0.16)	(0.25)	(0.31)
Counter-Att. Treat.	−0.06	−0.03	0.26	2.49***	0.54***	1.27***
	(0.13)	(0.14)	(0.21)	(0.16)	(0.19)	(0.31)
Pro-Att. exponentiated	3.82	1.33	1.77	1.39	1.22	1.18
Counter-Att. exponentiated	0.94	0.97	1.3	12.11	1.72	3.56
Observations	1,648	1,648	1,648	1,648	1,648	1,648

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

Table A.11: Effects of the Treatments on News Slant by Subsample

	News Exposure			Browsing Behavior			Shared Posts		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Liberal Treatment	−0.237***	−0.234***	−0.191***	−0.091**	−0.080**	−0.100**	−0.021*	−0.106*	−0.045
	(0.060)	(0.063)	(0.073)	(0.037)	(0.039)	(0.046)	(0.012)	(0.056)	(0.065)
Conservative Treatment	0.355***	0.365***	0.462***	0.102**	0.105**	0.107**	0.046***	0.054	0.131*
	(0.067)	(0.070)	(0.082)	(0.040)	(0.041)	(0.050)	(0.013)	(0.060)	(0.073)
Cons. Treat. - Lib. Treat.	0.59***	0.60***	0.65***	0.19***	0.19***	0.21***	0.07***	0.16***	0.18**
	(0.06)	(0.07)	(0.08)	(0.04)	(0.04)	(0.05)	(0.01)	(0.06)	(0.07)
Ext. Subsample	X			X					
Posts Subsample							X		
Ext. + Posts Subsample		X			X			X	
Ext. + Posts + Endline Subsample			X			X			X
Observations	1,556	1,433	1,010	1,785	1,652	1,166	18,328	979	685

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

Table A.12: Effects of the Treatments on Primary Outcomes, Controlling for Covariates

(a) Effect on Political Opinions

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

(b) Effect on Affective Polarization

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

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

Table A.13: Effect of the Treatments on the Affective Polarization Index, Excluding Each Index Component

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

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

Table A.14: Effects of the Treatments on Primary Outcomes, According to Outlets Offered

(a) Effect on Political Opinions, According to Outlets Offered

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

(b) Effect on Affective Polarization, According to Outlets Offered

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

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

Table A.15: Effects of the Treatments on Primary Outcomes, by Subsample

(a) Effect on Political Opinions, by Subsample

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

(b) Effect on Affective Polarization, by Subsample

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

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

Table A.16: Effect of News Exposure on Affective Polarization

(a) Causal Effect Based on Experimental Variation

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

(b) Cross-Sectional Correlation in Control Group

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

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

Table A.17: Effects of the Treatments on Attitudes Toward Each Party

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

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

Table A.18: Primary Outcomes Using Different Index Methods

(a) Political Opinions

	(1)	(2)	(3)	(4)	(5)
Liberal Treatment	−0.006	−0.008	0.001	−0.007	−0.006
	(0.005)	(0.017)	(0.015)	(0.009)	(0.007)
Conservative Treatment	−0.001	0.025	0.011	0.010	0.005
	(0.005)	(0.017)	(0.014)	(0.009)	(0.007)
Cons. - Lib. Treatment	0.005	0.033*	0.010	0.017*	0.011
	(0.005)	(0.017)	(0.010)	(0.009)	(0.007)
Controls	X	X	X	X	X
Index Method	Standard	Inv-Cov	Inv-Cov	Inv-Cov	Inv-Cov
Include Missing Outcomes	-	No	Yes	No	Yes
Replace Negative Weights With 0	-	No	No	Yes	Yes
Observations	17,635	9,434	17,635	9,434	17,635

(b) Affective Polarization

	(1)	(2)	(3)
Pro-Att. Treatment	0.005	0.004	0.001
	(0.012)	(0.017)	(0.013)
Counter-Att. Treatment	−0.028**	−0.031*	−0.035***
	(0.012)	(0.017)	(0.013)
Pro-Att. Treat. - Counter-Att. Treatment	0.033***	0.035**	0.036***
	(0.012)	(0.017)	(0.013)
Controls	X	X	X
Index Method	Standard	Inv-Cov	Inv-Cov
Include Missing Outcomes	-	No	Yes
Observations	16,896	10,059	16,896

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

Table A.19: Effects of the Treatments on Behavioral and Attitudinal Polarization Measures

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

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

Table A.20: Common Phrases Mentioned When Describing the Baseline Survey's Purpose

(a) Common Three-Word Phrases by Treatment Assignment

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

(b) Common Two-Word Phrases by Treatment Assignment

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

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

Table A.21: Phrases with Highest Differential Usage When Describing the Survey's Purpose

(a) Control Group and the Pro-Attitudinal Treatment

Expression	Share Among Phrases with the Same Length		
	Control	Pro	Counter
chamber	0.16%	0.41%	0.47%
divers	0.01%	0.13%	0.14%
echo	0.16%	0.42%	0.47%
echo chamber	0.20%	0.51%	0.58%
media echo	0.02%	0.12%	0.13%
media echo chamber	0.02%	0.15%	0.17%
open	0.01%	0.16%	0.21%
page	0.00%	0.14%	0.19%
social	1.68%	2.21%	2.08%
social media	1.91%	2.56%	2.40%

(b) Control Group and the Counter-Attitudinal Treatment

chamber	0.16%	0.41%	0.47%
divers	0.01%	0.13%	0.14%
echo	0.16%	0.42%	0.47%
echo chamber	0.20%	0.51%	0.58%
like	0.18%	0.31%	0.46%
open	0.01%	0.16%	0.21%
page	0.00%	0.14%	0.19%
percept	0.86%	0.61%	0.50%
promot	0.03%	0.09%	0.15%
willing	0.01%	0.05%	0.10%

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

connect polit	0.04%	0.07%	0.02%
like	0.18%	0.31%	0.46%
peopl identifi	0.02%	0.04%	0.01%
percept media polit	0.03%	0.04%	0
polit	10.62%	10.41%	9.67%
push	0.03%	0.07%	0.14%
push liber	0.02%	0.03%	0.09%
rang	0.02%	0.01%	0.04%
seem like	0.01%	0	0.03%
social media bias	0.03%	0.07%	0.01%

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

Table A.22: Most Common Two-Words Phrases Appearing in Posts

(a) Post Participants were Exposed to in their Feed

Exposed in Feed, Conservative Outlets		Exposed in Feed, Liberal Outlets	
Pro	Counter	Pro	Counter
donald trump (10.68%)	presid trump (5.15%)	presid trump (8.33%)	presid trump (7.56%)
presid donald (8.97%)	donald trump (5.09%)	donald trump (4.07%)	donald trump (4.86%)
presid trump (3.79%)	presid donald (2.92%)	white hous (3.20%)	white hous (2.66%)
white hous (2.92%)	white hous (2.58%)	stormi daniel (1.93%)	presid donald (2.16%)
high school (2.30%)	high school (1.56%)	presid donald (1.63%)	stormi daniel (2.16%)
hillari clinton (1.56%)	trump administr (1.44%)	high school (1.14%)	michael cohen (1.23%)
gun control (1.53%)	gun control (1.19%)	special counsel (1.02%)	high school (1.20%)
school shoot (1.39%)	school shoot (1.05%)	unit state (1.01%)	unit state (0.99%)
trump administr (1.33%)	special counsel (0.91%)	michael cohen (0.98%)	special counsel (0.95%)
attorney general (1.22%)	hillari clinton (0.85%)	school shoot (0.97%)	gun violenc (0.91%)

(b) Post With Links Visited by Participants

Posts Visited, Conservative Outlets		Posts Visited, Liberal Outlets	
presid trump (5.07%)	presid trump (5.33%)	presid trump (5.19%)	donald trump (3.01%)
donald trump (4.06%)	donald trump (3.18%)	donald trump (4.35%)	presid trump (3.01%)
white hous (2.84%)	white hous (3.18%)	white hous (2.12%)	day befor (0.90%)
presid donald (2.03%)	gun control (2.05%)	high school (1.17%)	former fbi (0.90%)
high school (1.83%)	hillari clinton (1.74%)	presid donald (1.06%)	high school (0.90%)
gun control (1.62%)	second amend (1.54%)	school shoot (0.78%)	someon els (0.90%)
north korea (1.42%)	presid donald (1.33%)	special counsel (0.73%)	white hous (0.90%)
attorney general (1.22%)	robert mueller (1.23%)	unit state (0.73%)	anoth child (0.60%)
hillari clinton (1.22%)	special counsel (1.23%)	michael cohen (0.67%)	anyon els (0.60%)
justic depart (1.22%)	trump administr (1.13%)	robert mueller (0.67%)	black student (0.60%)

(c) Posts Shared by Participants

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

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

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

	News Exposure		Browsing Behavior		Shared Posts	
	(1)	(2)	(3)	(4)	(5)	(6)
Liberal Treatment	-0.237^{***}	-0.337^{***}	-0.091^{**}	-0.059	-0.021^{*}	-0.011
	(0.060)	(0.094)	(0.037)	(0.052)	(0.012)	(0.019)
Conservative Treatment	0.355^{***}	0.419^{***}	0.102^{**}	0.148^{**}	0.046^{***}	0.067^{***}
	(0.067)	(0.099)	(0.040)	(0.067)	(0.013)	(0.019)
Cons. Treat. - Lib. Treat.	0.59^{***}	0.76^{***}	0.19^{***}	0.21^{***}	0.07^{***}	0.08^{***}
	(0.06)	(0.09)	(0.04)	(0.07)	(0.01)	(0.02)
Reweigted		X		X		X
Observations	1,556	1,556	1,785	1,785	18,328	18,328

This table estimates the effects of the treatments on the slant of posts observed in the Facebook feed, websites visited and posts shared. Columns (1), (3), and (5) show the estimates in the extension or access posts subsamples using equal weights. These columns are the same as columns (1), (4), and (7) in Appendix Table A.11. Columns (2), (4), and (6) reweight the subsamples to match the US population based on the following covariates: self-reported ideology, the share of participants identifying as Democrats, Republicans, and Independents, the difference between the participants' feelings toward their party and the opposing party, age, and the share of females. This analysis is discussed in Appendix C.4. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.24: Effects of the Treatments on Primary Outcomes, Reweighted to Match the US Population

(a) Political Opinions

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

(b) Affective Polarization

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

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

Table A.25: Predicted Effect in Full Baseline Sample

		(1)	(2)	(3)
Outcome	Treatment	Main Effect Estimated	Predicted Effect in Subsample	Predicted Effect in Baseline Sample
News exposure, posts slant (std. dev.)	Conservative treatment, compared to liberal treatment	0.592	0.545	0.571
Browsing behavior, news sites slant (std. dev.)	Conservative treatment, compared to liberal treatment	0.193	0.204	0.218
Political opinions index	Conservative treatment, compared to liberal treatment	0.005	0.003	0.003
Affective polarization index	Pro-Attitudinal treatment, compared to counter-attitudinal treatment	0.033	0.026	0.027

This table predicts the main effects estimated in the paper for the entire baseline sample. Column (1) shows the main effect estimated in each subsample. These effects are shown in columns (1) and (4) of Appendix Table A.11 and column (3) of Appendix Table A.12. For columns (2) and (3), I first estimate heterogeneous effects in the endline survey and extension subsamples using causal forests with many survey and Facebook covariates as explained in Section C.5. Column (2) predicts the treatment effect within the subsample using out-of-bag prediction. Column (3) predicts the effect for the entire baseline sample.

Table A.26: Effects of the Treatments on Self-reported Familiarity and Accurate Political Knowledge Outcomes

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

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

Table A.27: Effects of the Treatments on Exposure to Words in the Facebook Feed

	Michael Cohen	Clark Shooting	Louis Farrakhan	Clinton Speech
	(1)	(2)	(3)	(4)
Liberal Treatment	2.558***	1.172***	0.161	0.041
	(0.820)	(0.350)	(0.116)	(0.041)
Conservative Treatment	0.554	0.080	0.398***	0.077**
	(0.531)	(0.260)	(0.103)	(0.032)
Cons. Treat - Lib. Treat	-2.00**	-1.09***	0.24*	0.04
	(0.81)	(0.31)	(0.13)	(0.04)
Controls	X	X	X	X
Expected Effect	Lib. Treat	Lib. Treat	Cons. Treat	Cons. Treat
Observations	1,730	1,730	1,730	1,730

This table estimates the effects of the treatments on topics appearing in participants' Facebook feeds. *Michael Cohen, Clark Shooting*, and *Louis Farrakhan* are the number of times the terms "michael cohen", "stephon clark", and "louis farrakhan" appeared, respectively. *Clinton Speech* is the number of times the word Clinton appeared along with the word vote and either the word India or the word husband. All regressions control for party affiliation, ideology, vote, age, age squared, gender, whether the participant follows the news, and whether the participant stated they know the name of their representative in congress. Data are from the extension subsample from the first eight weeks following the intervention. Robust standard errors. *p<0.1 **p<0.05 ***p<0.01

Table A.28: Estimations Decomposing the Segregation in News Exposure

	Subscriptions	FB Usage: Total Posts Observed	Platform Algorithm: Share of Posts
	OLS	OLS	IV
	(1)	(2)	(3)
Pro-Att. Treatment	0.505***	248.765*	
	(0.086)	(150.666)	
Subscriptions			0.966***
			(0.093)
Subscriptions * Pro-Att.			0.460***
			(0.162)
Unit	Participant	Participant	Participant by Outlet Group
Baseline Controls		X	
Mean in Counter-Att. Treatment	1.535	2043.019	0.851
Observations	1,059	1,059	2,117

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

References

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

Angrist, Joshua D., and Ivan Fernandez-Val. 2013. "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework." In *Advances in Economics and Econometrics - Tenth World Congress*. , ed. Daron Acemoglu, Manuel Arellano and Eddie Dekel, 401–433.

Aronow, Peter M., and Allison Carnegie. 2013. "Beyond LATE: Estimation of the Average Treatment Effect with an Instrumental Variable." *Political Analysis*, 21(4): 492–506.

Bakshy, Eytan, Solomon Messing, and Lada A Adamic. 2015. "Exposure to Ideologically Diverse News and Opinion on Facebook." *Science*, 348(6239): 1130–1132.

Chan, Jimmy, and Wing Suen. 2008. "A Spatial Theory of News Consumption and Electoral Competition." *Review of Economic Studies*, 75(3): 699–728.

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

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

Flaxman, Seth R, Goel Sharad, and Justin M Rao. 2016. "Filter Bubbles, Echo Chambers, and Online News Consumption." *Public Opinion Quarterly*, 80: 298–320.

Gosling, Samuel D., Peter J. Rentfrow, and William B. Swann. 2003. "A Very Brief Measure of the Big-Five Personality Domains." *Journal of Research in Personality*, 37(6): 504–528.

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

Heckman, James J., Sergio Urzua, and Edward J. Vytlacil. 2006. "Understanding Instrumental Variables in Models With Essential Heterogeneity." *The Review of Economics and Statistics*, 88(3): 389–432.

Jeffrey, Lewis, B. Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet. 2020. "Voteview: Congressional Roll-Call Votes Database."

Peterson, Erik, Goes Shared, and Shanto Iyengar. 2019. "Partisan Selective Exposure in Online News Consumption: Evidence from the 2016 Presidential Campaign." *Political Science Research and Methods*, 1–17.

Rogowski, Jon C., and Joseph L. Sutherland. 2016. "How Ideology Fuels Affective Polarization." *Political Behavior*, 38(2): 485–508.

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

Stone, Daniel F. 2020. "Just a Big Misunderstanding? Bias and Affective Polarization." *International Economic Review*, 61(1): 189–217.

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

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