Social Influence and News Consumption*

Alex Moehring Carlos Molina

February 14, 2026

Abstract

We study how social influence affects preferences for biased news through a field experiment on X (formerly Twitter). We find that (i) information about the news outlets followed by peers does not influence individuals' own news diets and (ii) social image concerns that arise when an individual believes their news diets may be observed by peers induces individuals to moderate their news diets. Using quasi-experimental variation from a policy change on X that reduced the visibility of endorsements and lowered social-image concerns, we find consistent evidence that this change increased engagement with politically-extreme outlets.

Keywords: biased news, social image, learning, polarization.
JEL Classification: C93, D72, D83, L82.

*Moehring: Mitch Daniels School of Business, Purdue University, email: moehring@purdue.edu. Molina: Department of Economics, University of South Carolina, email: carlosmolina@sc.edu. We thank Daron Acemoglu, Ben Olken, and Frank Schilbach for generous advice and support. We also thank Nicolás Ajzenman, Cevat Giray Aksoy, Ceren Baysan, Claudio Ferraz, Dean Eckles, Leopoldo Fergusson, Martin Fiszbein, Ro'ee Levy, Alexey Makarin, Jacob Moscona, and Catherine Tucker, as well as seminar participants at the Universities of British Columbia, Carleton, MIT, Georgia Tech, Northwestern, Nottingham, Pittsburgh, and South Carolina, for helpful feedback. This work received support from the George and Obie Shultz Fund at MIT, the Institute for Humane Studies, and the Bradley and Hewlett Foundations. IRB approval for this project was obtained from the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects (protocols E-3516 and E-4465) and the experiment and primary analysis were pre-registered at the American Economic Association's registry for randomized controlled trials (protocol AEARCTR-0011147).

1 Introduction

Individuals increasingly consume political news in social environments where peer choices are highly visible. Social media platforms, messaging groups, and online networks routinely expose users to what others read, share, and endorse, creating the potential for peers to shape individual news consumption. A growing concern is that interactions with peers – often in echo chambers where individuals are more likely to interact with peers with like-minded beliefs – may influence individuals' preferences for politically slanted content and ultimately polarization [Settle, 2018, Sunstein, 2018].

In this paper, we focus on two distinct channels through which social influence can impact preferences for biased news. The first is the *peer information channel*: an individual's beliefs about the news diets of their peers may directly influence that individual's own news diet.[1] For example, individuals may seek to consume a news diet that is ideologically congruent with their peers.[2] The second is the *social image concerns channel*: individuals may alter their news diets when they expect peers to observe these choices.[3] The extent to which these channels affect an individual's demand for slanted news content is ex-ante ambiguous. For example, these channels could lead to more (less) demand for biased content for a focal individual in the specific case where these channels stimulate conformity – the need to resemble their peers – and the focal individual has peers with more (less) extreme preferences for biased news.

There is limited causal evidence on the role of social influence in shaping preferences for biased news, in part due to substantial identification challenges. These challenges include individuals forming their social networks endogenously as well as the reverse causality problem: a focal individual can influence their peers *and* be influenced by them. Moreover, distinguishing between the distinct channels of social influence further complicates identification. In addition, because researchers rarely observe the full set of news outlets individuals consume, it is difficult to directly measure how social influence affects preferences for biased news.

[1] We define an individual's news diet to be the set of news outlets to which an individual consumes.

[2] This channel is motivated by and closely related to the literature studying the impact of social norms on economic behaviors ranging from household finance [Lindbeck, 1997], labor force participation [Bursztyn et al., 2020b], and extremism [Bursztyn et al., 2020a].

[3] This channel is motivated by a growing literature in economics that studies the importance of social image concerns in driving economic behaviors including political actions [DellaVigna et al., 2016], educational investments [Fryer Jr and Torelli, 2010, Bursztyn and Jensen, 2015], and labor decisions [Bursztyn et al., 2017]

We design a novel field experiment on X, formerly known as Twitter, to separately identify the extent to which peer information and social image concerns impact the demand for polarizing news content. We share with all participants a summary of their news diet. This summary contains both the quantity of news outlets a participant follows in addition to the average ideological slant of the news outlets they follow. We then cross-randomize participants into two experimental conditions. Participants in the *peer information* condition received an additional summary of the ideological slant of the news diets of their peers. In the *disclosure* condition, treated participants are incentivized to reveal to their peers the summary of their news diets via a post on the platform. Control participants receive an incentive to share a placebo message that contains no private information. All individuals are given the opportunity to change their news diets before sharing their news diets or the placebo message with their peers.

Our results suggest that social influence on social media platforms is not a major contributor to the demand for slanted news content. The peer information treatment succeeds at inducing variation in participants' beliefs about the slant of the news diets of their peers. However, this update in beliefs does not translate into changes in the participants' own news diets. Therefore, we conclude the peer information channel has limited impact on the ideological slant of the news individuals consume. The disclosure treatment, on the other hand, does induce participants to alter their news diets. We find that participants moderate their news diets when they believe their news diets will be observed by their peers, suggesting the social image concerns channel also does not contribute to the demand for slanted news outlets. Our findings therefore suggest that social influence is unlikely to be a primary driver of polarized news consumption.

Comparing the treatment and control groups in the disclosure condition yields the following four results. First, the incentive to publicize the news diet in the disclosure condition has the expected effect on sharing behavior. The treatment group is 8.2 percentage points (standard error of 0.8) more likely to share the summary of their news diet with their peers.

Second, the disclosure treatment has a large impact on participants' news diets. The treatment increased the probability that a participant makes at least one change to their news diet by 21.4% (standard error of 4.6%) and boosted the total number of outlets followed by 28.6% (standard error of 6.0%) relative to the control mean.

Third participants in the treatment group are 40.9% more likely to shift their news diet toward the ideological center and 31.6% more likely to move it toward their peers (standard errors of 8.2% and 7.8%, respectively). There is important heterogeneity based

on the initial slant of a participant's news diet and the average slant of the news diet of their peers. We find that participants move their news diets strongly toward both their peers and the ideological center when both are in the same direction. When a participant's peers have more extreme news diets than the participant, we find the participant moves their news diet toward the center. Therefore, it appears that participants predominantly want to be perceived as having a neutral news diet and that social image concerns moderate participant news diets.

Fourth, we find that the effects of the disclosure treatment are stable over time and persist for months after the experiment. In addition, we find that the disclosure treatment increases engagement with news outlets in the form of reposts, likes, and posts mentioning a news source. That the disclosure treatment has important effects on long-term news diets and engagement rejects the hypothesis that participants in this treatment adjust their news diet temporarily during the experiment and then revert to their original news diets afterwards.

The peer information treatment explores whether the peer information channel shapes preferences over news outlets. The peer information treatment *does* change participants' beliefs about the slant of the news diets of their peers: those in the peer treatment group are 32.3% (standard error of 6.2%) more likely to update their beliefs about their peers relative to the control mean and they update their beliefs in the direction of the signal we provide. Although the treatment successfully induces variation in participants' beliefs about the news diets of their peers, we find no evidence that this translates into changes in news diets. The treatment group did not differentially change the number or slant of outlets followed. These effects are statistically indistinguishable from zero and small in magnitude.

As complementary evidence, we exploit a different source of variation to estimate the importance of the peer information channel. We provide participants in the treatment group with an unbiased but noisy estimate of the slant of their peers' news diets based on a random sample of peers due to an API rate limit that prohibits collecting more exhaustive data in real time. This sampling induced noise generates additional exogenous variation that we use to identify how left vs. right surprises affect participants' beliefs about the news diets of their peers. We find that individuals who receive signals with right-leaning noise about which news sources their peers follow significantly update their posterior beliefs about the slant of their peers' news diets to the right relative to those who receive a signal with left-leaning noise. We again find that participant beliefs about the news diets of their

peers has a negligible effect on participants' own news diets using this alternative strategy.

While our experiment demonstrates that social image concerns can moderate demand for slanted news outlets, a concern remains that these effects may not generalize to platform-wide interventions. We therefore complement our experimental results with an analysis of a natural experiment that exploits a major platform-level policy change on X. In June 2024, X modified its interface such that the list of users who "liked" a post became private, thereby removing the public visibility of this engagement action. This policy change closely mirrors the disclosure randomization in our experiment: it exogenously removes the social image concern associated with liking a post. Our experimental results would suggest this would remove the moderating force of social image concerns on individuals' news diets. Using a comprehensive dataset of more than nine million posts published by news outlets on the platform before and after the change, we show that making likes private led to a large and statistically significant increase in the quantity of likes, and this was especially true for posts from politically slanted outlets. This result provides consistent evidence in a real-world environment that social image concerns play an important role in shaping engagement with biased news. Taken together, our experimental and quasi-experimental evidence indicates that social image concerns systematically discourage the consumption and endorsement of extreme or partisan news online.

Related Literature

This paper contributes to several strands of the existing literature. First, we contribute to the literature that examines consumer preferences for politically slanted news outlets, including both theoretical underpinnings for such preferences [Suen, 2004, Gentzkow and Shapiro, 2006, Mullainathan and Shleifer, 2005] and empirical evidence that documents that consumers often prefer slanted news outlets Gentzkow and Shapiro [2010], Chopra et al. [2022, 2023]. In addition, this work is related to the empirical literature studying how the news consumption of an individual's peers impacts the focal individual's consumption [Aral and Zhao, 2019, Messing and Westwood, 2014]. We contribute to this literature by demonstrating that social influence plays an important role in discouraging the consumption of biased news, mainly due to social image concerns. We find that the peer information channel does not play a significant role in influencing news diets: our treatments influence participants' beliefs about which news outlets their peers follow, but they do not translate into changes in participants' own news diets.

Our results shed light on the role that information technologies, particularly social

media, play in explaining preferences for biased news and ultimately shaping political attitudes [Gentzkow and Shapiro, 2011, Enikolopov et al., 2020, Boxell et al., 2022]. Much of the recent literature investigates the causal effect of exposure to pro-attitudinal (as opposed to counter-attitudinal) news content on polarizing attitudes based on the idea that social media algorithms expose individuals to more pro-attitudinal content to maximize engagement [Guess and Coppock, 2020, Levy, 2021, Broockman and Kalla, 2022, Casas et al., 2022]. Our paper contributes to this literature by focusing on a different channel: that social media influences preferences for biased news (and thus polarization) by facilitating interactions within echo chambers, where individuals are more likely to interact with peers who share similar ideological beliefs. Our paper is among the first to provide causal evidence that social influence on social media drives preferences for politically slanted news. We find little evidence that social influence causes individuals to consume more politically slanted news outlets. Instead, our paper reveals a mechanism that could mitigate individuals' preferences for politically slanted outlets. As social media platforms amplify the visibility of individual interactions, social image concerns become more prevalent. Our findings suggest that individuals moderate their news diets due to social image concerns to signal a preference for a balanced news diet to their peers.

Finally, our paper advances the literature studying the effects of social image concerns [Bursztyn and Jensen, 2017], particularly how they relate to political attitudes [Gerber et al., 2008, Funk, 2010, DellaVigna et al., 2016]. We provide the first evidence on how social image concerns impact the demand for biased news. We also contribute by studying these concerns in the context of social media, and connect the recent psychology literature that highlights the important role of social media in exacerbating image concerns and the understudied results of this effect on online behavior [Fardouly and Vartanian, 2016].

Outline

The rest of the paper is organized as follows. Section 2 describes the experimental design, empirical strategies, outcome variables of interest, and other details of the field experiment. Section 3 presents the basic statistics of our sample. Sections 4 and 5 describe the main results regarding the role of social image concerns and peer information in shaping individual preferences for biased news, respectively. Section 6 analyzes a policy change on X that provides a natural experiment varying the extent to which social image concerns impact likes on the platform. Section 7 concludes, while the Supplementary Material contains several additional descriptions and robustness checks.

2 Experimental Design and Data

2.1 Experiment Overview

Figure A-1 summarizes the design of the experiment. We used X ads to recruit American adults between March and June 2023. A total of 951,470 unique users saw the ads, which received 12,940 clicks. Appendix A.1 details our ads and recruitment.

Users who clicked on the ads were directed to the survey landing page, which contained an overview of the experiment, described the incentives to participate in the study, requested the participants' X handle or username, and presented the consent form. A total of 5,190 individuals who consented to participate and had a public account were invited to begin the baseline survey.[4] The baseline survey collected demographics and a wide range of pre-treatment covariates including political engagement, political ideology of the focal participant and their peers, political knowledge, and affective polarization. The survey also elicited self-reported beliefs about the political bias of the news outlets that the participants were following, as well as those followed by their peers.[5] We randomized the 4,546 participants who completed the baseline survey into the treatment arms we describe below.

While participants completed the baseline survey, we used the platform's API to collect the list of accounts they followed and those that followed them. We also collected the same information for a random sample of the participant's followers.[6]

Using these data, we provide all participants with a summary of their news diet based on the outlets they follow on X immediately following the baseline survey. This summary reports both the number of news outlets the participant follows (relative to the average user of X) and the average political slant of these outlets.[7] Figures A-2A and A-2B illustrates the design of the infographics conveying this information to participants, using examples

[4]Users with private X accounts were ineligible to participate because we are unable to observe their news diets and construct the treatments for them.

[5]Unless otherwise specified, we use the term *peers* to broadly refer to the set of accounts connected to a participant's account via a follow—either accounts that the participant follows or those that follow the participant. In Section 5.3.2, we analyze alternative definitions of a participant's peers and examine the robustness of the main results under these alternatives.

[6]We only retrieved the network of a subset of five followers in real-time due to restrictions on the number of queries that can be made to the X API. After the experiment, we retrieve the network of a larger random subset of followers (approximately 20). We discuss this in more detail in Section 2.4, including how we exploit this limitation to generate additional exogenous variation to study the role of peer information.

[7] To compute these measures, we constructed a comprehensive census of news outlets on X described in Section 2.3. We cross-reference this census, which also includes a measure of political bias for each outlet, with the list of accounts that each participant follows to identify the aforementioned metrics.

of two well-known politicians.

We then cross-randomized participants into the *peer information* and *disclosure* conditions. Participants assigned to the peer information treatment group received an additional summary that contains an estimate of the average slant of the news diets of the participant's peers. We refer to this as the *peer signal*, which we estimate by taking the average slant of the news outlets followed by the random set of the participant's followers for whom we collect information at baseline.[8] Figures A-2C and A-2D illustrates the design of the infographics containing the peer signal.

After reviewing the respective infographics (a single infographic in the case of the peer information control group), participants were then provided with instructions regarding the disclosure condition. Those in the treatment group were incentivized to post on the platform an infographic revealing the summary of their own news diet. Importantly, the instructions clearly stated that participants would not be asked to share the summary until after they had had the opportunity to modify which news outlets they followed. The control group received a similar message, but were instead incentivized to post a placebo message containing only a referral link to promote the study with no personal information.

Participants were then given the opportunity to change the news outlets they followed. We provided a list of six news outlets that the participant did not already follow, three outlets that would shift the slant of the participant's news diet to the left and three that would shift it to the right. We also explained how following any of these sources would affect their average political slant.[9] Figure A-3 presents an example of this information. In addition to these six outlets, participants could click a button to open a page showing how following or unfollowing any news outlet would impact the slant of their news diet.

Once participants had an opportunity to adjust their news diets, we re-scraped the network of their X account and presented them with an updated news diet summary. All participants then had the option to select one of two buttons – "Share Referral Link" or "Share News Diet Summary" – which would automatically draft posts for them to review and share, thus facilitating compliance with the incentives of the disclosure condition. At this stage, no differential information or reminders about the incentives were provided

[8]Section 5.3.2 describes the reason we use information on the participant's followers to construct this peer signal and presents robustness to alternative definitions.

[9]In an effort to make these suggestions relevant for the participant, we randomized whether the suggested outlets came from the most popular outlets or from a personalized collaborative-filtering based recommendation algorithm. Appendix A.2 describes these algorithms in more detail and assesses the robustness of our key findings to the algorithm used.

across treatment groups. Figure A-4 illustrates the design of the drafted posts shown to participants who clicked on these buttons.

Finally, participants were invited to complete an endline questionnaire that re-elicited beliefs about political ideology, the political bias of the news outlets they follow on X, and the political bias of the news outlets followed by their peers. In the months after the intervention, we tracked the activity of participants and their peers using the X API, including both changes in the participant's network (and therefore news diets) and the participant's engagement through posts, likes, and reposts.

2.2 Incentives to Participate

The consent page at the start of the intervention informed participants there was an opportunity to enter a lottery for a $200 prize upon completion of the endline survey. It also noted that participants might be selected during the survey to complete an additional task, which would enter them into a bonus lottery (see details below). There were no other financial incentives to participate in the study. Participants were also informed that by completing the survey they would receive information about the political bias of the news content they regularly encounter on X and noted that their participation would contribute to an academic research project.

We used the bonus lottery to incentivize participants to post either their news diet summary, if assigned to the disclosure treatment group, or the placebo message containing the referral link and no information about their news diet, if assigned to the disclosure control group. Participants were only eligible for participation in the bonus lottery if they complied with the treatment by posting the relevant post. We randomized the bonus lottery amount to either $100 or $200.

Participants were made aware that the probability of winning each of the two lotteries was independent–compliance with the bonus task did not affect the likelihood of winning the participation lottery. We elicited participants' beliefs about their chances of winning both lotteries in the endline survey and found that the median participant expected to earn a total of $2.50 in the study, with approximately $2 coming from the participation lottery and the remaining $0.5 from the bonus lottery. Participants who complied with the bonus task (17%) assigned an expected value of $3 to this lottery.

2.3 News Outlets Dataset

We construct a census of U.S. political news outlets with accounts on X. As explained in Section 2.1, this dataset is central to our experimental design: we combine it with information on the accounts each participant follows (or the accounts followed by each participant's peers) on X to determine both the number of news outlets they follow and the average slant of those outlets. These measures are then used to generate the infographics shown to participants during the experiment.

Appendix A.3 provides a detailed description of the construction of this census and here we present a high-level summary. First, we create a comprehensive set of potential news outlets by harmonizing the lists compiled by Athey et al. [2021] and Braghieri et al. [2024].

Second, we link each outlet to an external measure of political slant using the measures provided in Robertson et al. [2018], which captures each outlet's relative propensity to be shared on X by Republicans versus Democrats. The slant measure ranges from -1 to 1: an outlet shared exclusively by Democrats (Republicans) has a slant of -1 (1) and a moderate outlet (shared equally by Democrats and Republicans) has a slant close to 0. We also calculated a score representing the share of each outlet's articles that cover hard news (e.g., political, economic, or societal developments) rather than soft news (e.g., entertainment or sports). Throughout we use the adjusted slant $s_j = s_j^{raw} \times s_j^{hard}$ where s_j^{raw} is the raw slant score for outlet j coming directly from Robertson et al. [2018] and s_j^{hard} is the hard score we estimate. This adjustment accounts for the fact that some outlets are preferred by partisans but carry less hard news and therefore are less politically slanted than others.[10]

Third, we map these domains to X handles using several techniques, including searching for X handles on outlets landing pages. Finally, we implement a series of cleaning procedures and manual verifications to ensure that the final database excludes non-hard-news outlets and outlets who focus on coverage outside the United States.

The resulting dataset includes 1,170 U.S.-based hard-news outlets, with information on each outlet's X handle, political slant, and an indicator distinguishing local from national outlets. In the experiment, we use the full set of outlets to generate the infographics and only the subset of national outlets (88) to produce the six suggested outlets that each participant receives.

Compared to existing datasets, the main advantage of our database is that it covers a

[10]The correlation between s_j^{raw} and s_j is very high at 0.96.

much more comprehensive set of outlets.[11] Figure A-5 displays the distribution of slant scores in our final dataset of news outlets.

2.4 Outcome Variables and Empirical Strategy

We designed the experiment to study effects on three complementary sets of outcomes. The first set of outcomes concerns changes in participants' beliefs about their own news diet and the news diets of their peers, as measured through the endline survey. The second set captures changes in participants' choices of news outlets to follow on the platform. This includes both short-term adjustments occurring during the experiment and longer-term changes in the news outlets participants follow, which we continue to track after the experiment concludes. The third set focuses on changes in engagement with news outlets, based on behavioral data collected on X after the experiment.

We estimate the treatment effects of social image concerns and peer information by comparing treatment and control groups within each randomization using the following intention-to-treat (ITT) regression:

$$Y_i = \alpha + \beta D_i + \gamma X_i + \varepsilon_i \tag{1}$$

where Y_i represents an outcome of interest for participant i, D_i is a dummy variable equal to 1 if the participant receives the relevant treatment, X_i denotes a set of pre-treatment covariates, and ε_i is an error term.[12] In the primary analysis, we report results without including pre-treatment covariates for transparency and ease of interpretation. Results that adjust for pre-treatment covariates, including using the double machine learning approach from Chernozhukov et al. [2018] described in our pre-registration, are reported in the appendix.[13] All statistical inference uses heteroskedasticity-robust standard errors.

In addition to variation induced by the random assignment of treatment and control

[11] Robertson et al. [2018] shows that their slant measure is highly correlated with those used in previous studies. Similarly, we find that the correlation between our adjusted slant measure and that of Bakshy et al. [2015]—a widely used database covering nearly 400 news outlets—is 0.93.

[12] Because of our cross-randomized design, the coefficient of interest β in equation (1) can be interpreted as the ITT effect averaged across both types of participants: those who received the other treatment and those who did not [Muralidharan et al., 2023]. We show in Appendix A.4 that our results are very similar when estimating a fully saturated model including both treatments and their interaction.

[13] In addition, our pre-registration specified an instrumental variables specification as a secondary analysis, in which treatment assignment serves as an instrument for participants' willingness to disclose their news diets and for changes in beliefs about the ideological slant of their peers' news choices. Estimates from this secondary specification are reported in Appendix A.5.

groups, our experiment introduces an independent source of variation that allows us to identify the effect of peer information on participants' news diets. As described in Section 2.1, the *peer signal* includes an estimate of the average slant of the news diets of a participant's peers—computed from a random sample of up to five of a participant's peers. Because it is based on a random sample of peers, this measure provides an unbiased but noisy estimate of the true average across the participant's peer group. Consequently, participants with identical followers could have been shown different estimates of the slant of their peers' news diets due solely to random sampling orthogonal to pre-treatment characteristics and potential outcomes. We exploit this feature to study how random variation in the reported peer slant affect posterior beliefs and subsequent news choices, which we estimate using the following specification:

$$Y_i = \alpha + \gamma \mathbb{I}[\sigma_i^{\text{signal}} \geq \sigma_i^{\text{truth}}] + \delta X_i + u_i \tag{2}$$

where $\mathbb{I}[c]$ is an indicator function that equals one if condition c holds. The main coefficient of interest, γ, captures the causal effect of providing participants with information indicating that their peers' news diets are more right-leaning than the *true* slant of their peers' news diets, which we term as a *signal with right noise*.[14] We estimate equation (2) separately for the peer information treatment and control groups. As only the former group observes the signal, the estimate for the control group serves as a placebo test. Finally, X_i is a vector of control variables. We present results without controls in the main text and show their robustness to the inclusion of controls in the appendix.

3 Descriptive Statistics

This section describes our sample and reports key summary statistics. Of the 4,546 participants who completed the baseline survey and were randomized into treatments, 3,755 were exposed to the experimental conditions. We define treatment exposure as viewing the relevant infographics and receiving the incentive to share a post. Among participants exposed to the experimental conditions, 2,745 responded to the endline survey, providing data to

[14]Due to X's API restrictions, we cannot observe the exact slant of the peer news diet (σ_i^{truth}). Instead, we construct an estimate based on a larger number of randomly selected peers—approximately twenty rather than the five used for the signal—which we refer to as the "true" slant to simplify exposition, albeit slightly abusing the term. Importantly, this approach does not require observing the true slant and remains valid as long as the estimate is unbiased and more precise than the provided signal, which in our case follows from the law of large numbers.

construct outcomes such as posterior beliefs. Because we observe behavioral outcomes (for example, changes in news-following behavior and engagement) for all 3,755 participants who are exposed to treatment, we use this larger group in the main analysis, but also use the subsample of participants who completed the endline survey in a robustness exercise and when outcomes rely on survey measures. In all cases, the results are highly consistent across samples.

The attrition rate in our main sample is 17%. Figure A-6A shows minimal differential attrition across the treatment groups. There is also no differential attrition in the probability of completing the endline survey (Figure A-6B). Table A-1 demonstrates the treatment groups are balanced on pre-treatment covariates, consistent with successful randomization.

Table A-2 contains a summary of the experiment participants. Panel A shows that our sample is predominantly White, more educated, more heavily male, and older than the U.S. adult population. As discussed in Appendix A.6, these characteristics are more common among politically engaged individuals, who are overrepresented in our sample. We view this group as substantively interesting, as politically engaged individuals are often exposed to the social influence that we study. Nevertheless, Appendix A.6 shows that our main findings are unchanged after accounting for these observable dimensions, whether by weighting the sample to match the observable traits of the U.S. adult population or by separately comparing politically engaged and less engaged individuals.

Panel B presents descriptive statistics on the engagement of our sample on X. Following Pew Research, which reports comparable statistics for a representative panel of X users, we present these measures separately for the median participant in the top 10% and bottom 90% of the non-like engagement distribution, where non-like engagement is defined as sum of posts, reposts, quotes and replies. The table shows that participants in our sample have a significant presence on the platform and that engagement is highly skewed: the 95th percentile sends 41 times more posts than those at the 45th percentile. The median participant in our sample posts around four times every five days, has 148 followers, and follows 500 accounts.

Figure A-7 displays the distributions of the number of news outlets individuals follow and the average slant of these outlets conditional on following at least one news outlet. We report these distributions across three samples: the study participants, the participants' peers, and a representative sample of X users (see Appendix A.7 for details on the construction of this sample). The median participant in our sample follows four news outlets and is much more politically engaged than both their peers and the representative sample of X

users. Around 20% of participants do not follow any outlets, which is substantially lower than the 38% or 55% for the participant's peers and representative sample, respectively.[15]

To assess the reliability of the summaries of participants' news diets that we construct during the experiment, we compare these summaries to their self-reported pre-experiment values. Figure A-8 plots the average slant of participant news diets based on their self-reported ideology and the self-reported slant of their news diet. The news diet summaries we construct during the experiment are highly correlated with the self-reported measures, increasing our confidence that the summaries capture important features of participants' news diets. Moreover, we find that participants' self-reported beliefs about the ideology and news diets of both their followers and who they follow on X is highly correlated with the average slant of outlets followed by their followers. This analysis suggests that the measure we show participants reflects, on average, participants' own beliefs about their peers on the platform.

4 Social Image Concerns

In this section, we study how social image concerns influence individuals' news diets. We do this by estimating the treatment effect of the disclosure condition using Equation (1), comparing subsequent adjustments in news diets between participants incentivized to publicize their news diets (and thus likely facing social image concerns) and those who are not. The control group is incentivized to post a placebo message, implying that treatment effects are not driven by the action of posting itself (or the incentive).

4.1 Do Incentives to Disclose Impact Disclosure?

Figure 1 indicates that our treatment had the intended effects on compliance. The two panels illustrate the fraction of participants that posted the placebo message and news diet summary for the disclosure control and treatment groups. The control group was 9.3 percentage points (standard error of 1.1) more likely to post the placebo message. In addition, 10.5% of participants in the treatment group posted the news diet infographic,

[15]Participants who do not follow any news outlets are classified as having a neutral news diet and are therefore assigned a slant score of zero. When presented with the news diet summary (e.g. Figure A-2), participants who do not follow any news outlets are depicted at the center of the political scale and at the bottom of the distribution for the number of followed outlets. We discuss differential effects for this subgroup when presenting the main results.

compared to only 2.3% in the control group. Thus the treatment group was 8.2 percentage points (standard error of 0.8) more likely than the control group to post the summary of their news diet. Both differences are statistically significant and show that participants complied with the incentives.

In addition, Figure A-9 demonstrates that compliance with the incentive is increasing in the monetary amount of the incentive. The difference in levels between sharing the news diet summary and the placebo message suggest that participants are less willing to share their own news diet, consistent with social image concerns creating an additional cost to sharing personal information.

4.2 Do Social Image Concerns Affect News Diets?

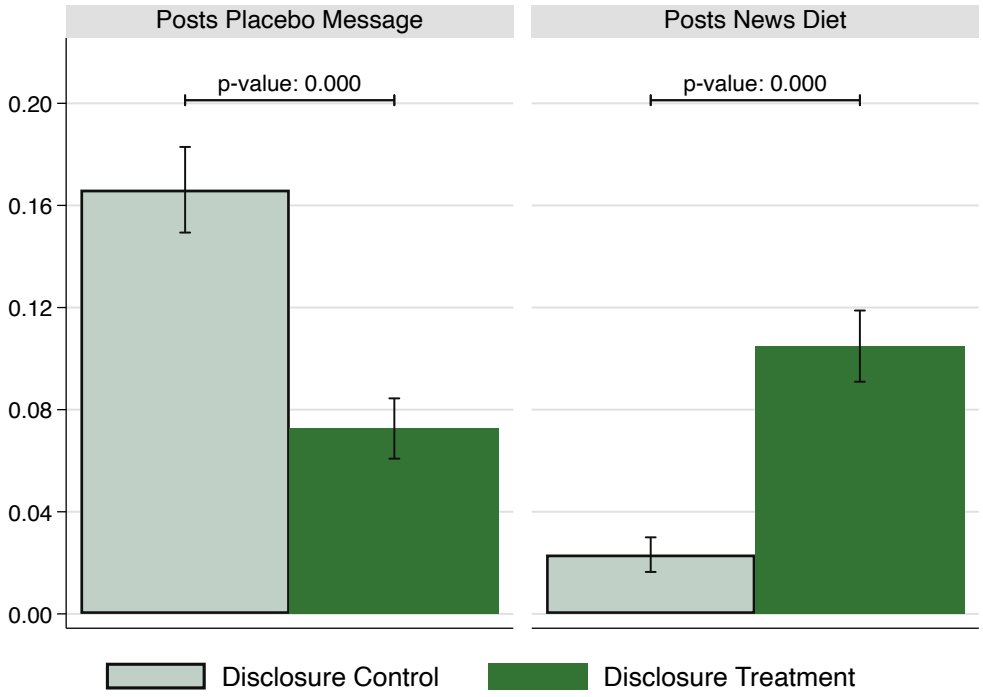
We now analyze how social image concerns influence the news outlets individuals choose to follow. We do so by examining changes that participants make to their news diets *during* the experiment in response to the disclosure condition.[16]

The three panels in Figure 2 show the probability of participants making at least one change to their news diet (Any Change), the absolute change in the total number of outlets followed ($|\Delta$ Followed Outlets$|$), and the absolute value of the change in the average slant of the outlets followed relative to those followed before the randomization ($|\Delta$ Slant$|$).

Participants in the control group update their news diets, potentially due to incorrect beliefs about their own news diets or because the outlet suggestions lower the search cost of finding new outlets. More importantly, we find that participants in the treatment group are significantly more likely to change their news diets. Relative to the control mean, the treatment increased the probability of making any change to the news diet by 21.4% (standard error of 4.6%), the absolute value of the change in the number of outlets followed by 28.6% (standard error of 6.0%), and the absolute value of the change in the slant of the outlets followed by 33.7% (standard error of 10.5%). In summary, the incentive to publicize one's news diet summary significantly increases the likelihood a participant adjusts their news diet consistent with the hypothesis that social image concerns influence the news outlets individuals follow.

[16] We observe the full set of outlets that participants follow immediately before and after the intervention, and therefore capture both changes driven by the six suggested outlets and any other adjustments (following or unfollowing other outlets). In practice, however, most observed changes reflect participants following the suggested outlets.

Figure 1: Effect of Disclosure Condition on Compliance

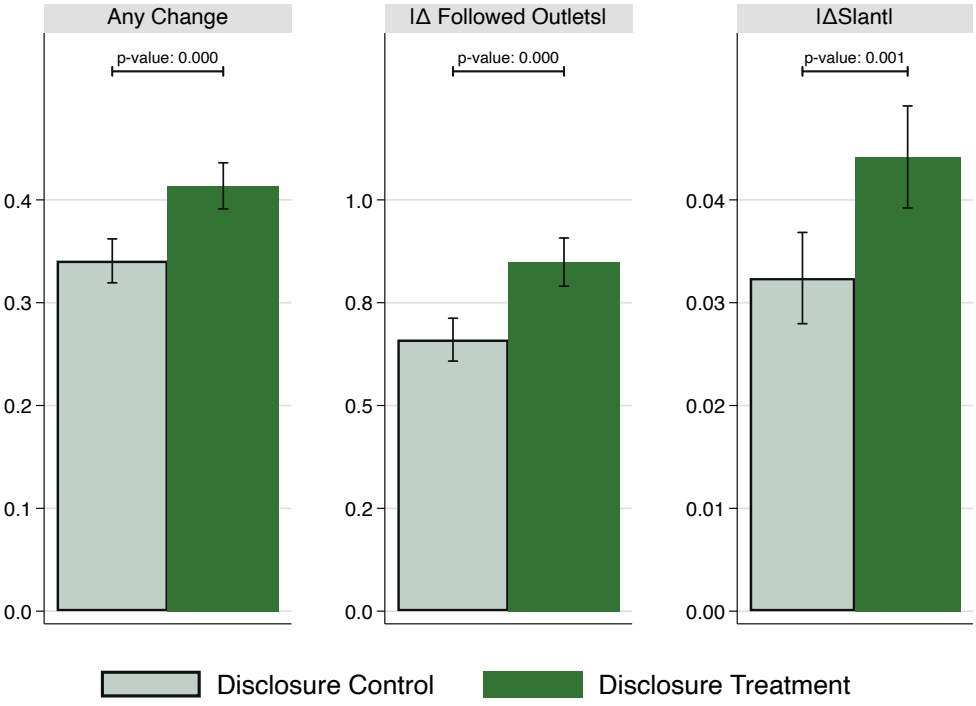


Notes: This figure reports the fraction of participants who post either the placebo message containing only a referral link with no personal information (left panel) or their own news diet summary (right panel) across the treatment and control groups in the disclosure condition. The difference between the treatment and control groups in the disclosure condition is that only participants in the former group are asked to reveal to their peers which news outlets they follow. The whiskers indicate the 95% confidence intervals. We also report the p-value for the null hypothesis of no difference between the treatment and control groups.

4.3 What Factors Drive Image Concerns?

The previous section demonstrates that social image concerns play an important role in shaping the news outlets participants choose to follow. In this section, we seek to better understand how participants would like to be perceived by their peers. In particular, we examine two potential mechanisms. First, individuals may be concerned about revealing the ideological leanings of the news outlets they follow. Some may wish to appear balanced and unbiased, forming opinions based on a broad spectrum of information. Others

Figure 2: Effect of Disclosure Condition on News Diets



Notes: This figure reports the mean of the following three outcomes across participants assigned to the treatment and control groups in the disclosure condition: an indicator for whether the participant makes any change to their news diets (left), the absolute change in the number of news outlets followed (center), and the change in the absolute slant of the outlets followed (right). In all cases, we construct these outcomes by comparing the set of outlets participants follow after the intervention—after they are given the opportunity to make changes—with those followed at baseline. The difference between the treatment and control groups in the disclosure condition is that only participants in the former group are asked to reveal to their peers which news outlets they follow. The whiskers indicate the 95% confidence intervals. We also report the p-value for the null hypothesis of no difference between the treatment and control groups.

might prefer to display more extreme positions to reinforce their ideological identity and demonstrate strong convictions. Second, individuals may be concerned about their relative standing within their social networks. A participant might prefer to align with the views of their peers, for instance, if they fear negative reactions when deviating from a news diet consistent with their ideological position. Alternatively, an individual might engage in a strategy of "digressing to impress". For example, a participant could believe that following

cross-cutting content signals to their peers an awareness of diverse perspectives, thereby enhancing their perceived intelligence and credibility in social interactions.

We test both of these hypotheses by investigating whether the disclosure treatment induces participants to choose news outlets that, on average, move the slant of their news diet toward (or away from) the ideological center and their peers. We focus on the group of news consumers, which we define as participants who follow at least one news outlet at baseline (approximately 80% of our sample), since pre-treatment slant and movement toward the center are not well-defined for non-news consumers.[17] We find that, compared to the control mean, participants in the treatment group are 7.7 and 6.3 percentage points (standard errors of 1.5 and 1.6, respectively) more likely to adjust their news diet toward the center and toward their peers, respectively (Figure A-11).

More interestingly, we estimate the heterogeneous treatment effect of the disclosure treatment conditioning on the relative slant of the participant's news diet, the news diets of their peers, and the center. Figure 3A displays the treatment effect of the disclosure treatment among participants whose peers have a news diet that is on average closer to the center than their own. These represent participants with left (right) leaning news diets whose peers have news diets with an average slant to their right (left). In this case, movements toward the center and toward peers are identical, and the disclosure treatment has a strong effect: the disclosure treatment induces an additional 8.0% of participants to adjust their news diet both toward the center and their peers.

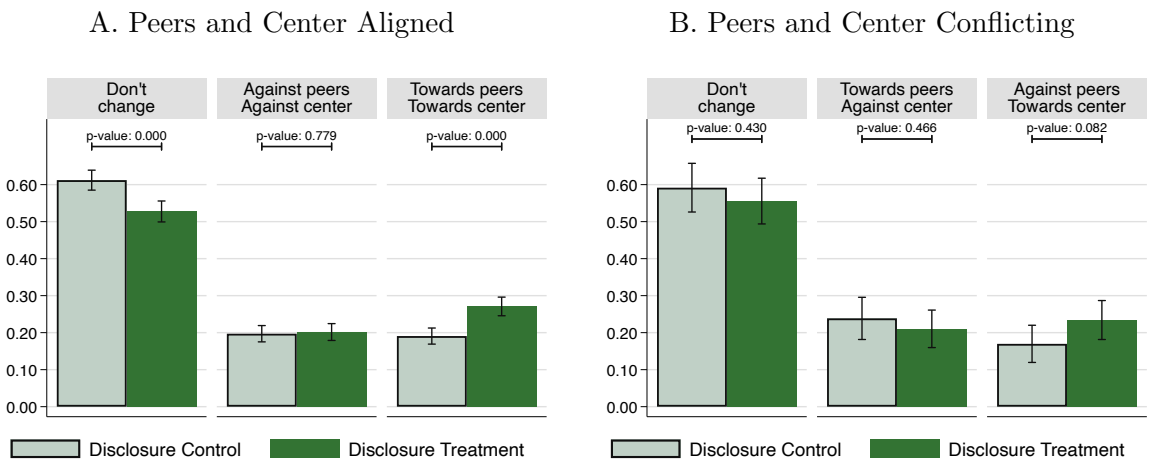
Figure 3B displays the treatment effect among participants whose peers and the center are in opposite directions: participants with left (right) leaning news diets whose peers have news diets to their left (right). In this case, the disclosure treatment induces an additional 6.4% of participants to move toward the center, and 2.8% fewer participants to move toward their peers. This suggests that social image concerns induce participants to moderate their news diets rather than conforming more with the news diets of their peers.

Together, these findings suggest that both the distance of an individuals' news diet from the ideological center and the positions of their peers' news diets are important in explaining how they choose news outlets in the presence of others. Social image concerns lead participants, on average, to select news diets with a slant closer to both their peers and the ideological center when a user's peers are closer to the center. By contrast, when a participant's peers are *away* from the center, participants tend to move their news diet

[17]Figure A-10 reports the results for non-news consumers. We again find that these participants move their news diets towards their peers when their peers follow any news outlets.

toward the center rather than toward their peers. This suggests that when the center and a participant's peers are in conflict, the desire to move toward the center, and thus be perceived as more moderate, tends to dominate. Finally, participants with little political engagement–those who do not follow any news outlets at baseline–tend to follow news outlets that are closer to their peers. Therefore, it appears that participants predominantly wish to be perceived as having a neutral news diet, though they also care about following news outlets with a slant similar to that of their peers to a lesser extent.

Figure 3: Directional Effect of Disclosure Condition — Ideological Center vs. Peers



Notes: This figure reports the fraction of participants who adjust the slant of their news diets toward the ideological center, toward their peers, and who make no changes, across the treatment and control groups in the disclosure condition. This analysis is restricted to news consumers—participants who follow at least one news outlet (80% of our sample). Panel A displays the results for participants whose peers and the center are aligned: participants with left (right) leaning news diets whose peers have news diets to their right (left). Panel B presents the results for participants whose peers and the center are in opposite directions: participants with left (right) leaning news diets whose peers have news diets to their left (right). The difference between the treatment and control groups in the disclosure condition is that only participants in the former group are asked to reveal to their peers which news outlets they follow. The whiskers indicate the 95% confidence intervals. We also report the p-value for the null hypothesis of no difference between the treatment and control groups.

4.4 Persistence of the Disclosure Treatment Effects

The treatment effects discussed above demonstrate that incentivizing participants to share their news diet summary induces them to change the news outlets they follow during the experiment. Here, we investigate whether these changes are short-lived or if they persist after the experiment by re-estimating equation (1) for outcomes that are observed periodically after the experiment. We estimate the effects of the disclosure treatment on

the probability of participant making at least one change to their news diet relative to baseline, the absolute value of the change in the total number of outlets followed, and the absolute value of the change in the average slant of the outlets followed relative to the outlets participants were following before the randomization.

Figure 4 shows that there is little decline in the treatment effects during the three months after the experiment. For all outcomes, we cannot reject that the treatment effects measured post-experiment are the equal to those estimated during the experiment.

4.5 Impact on Active Engagement with News Outlets

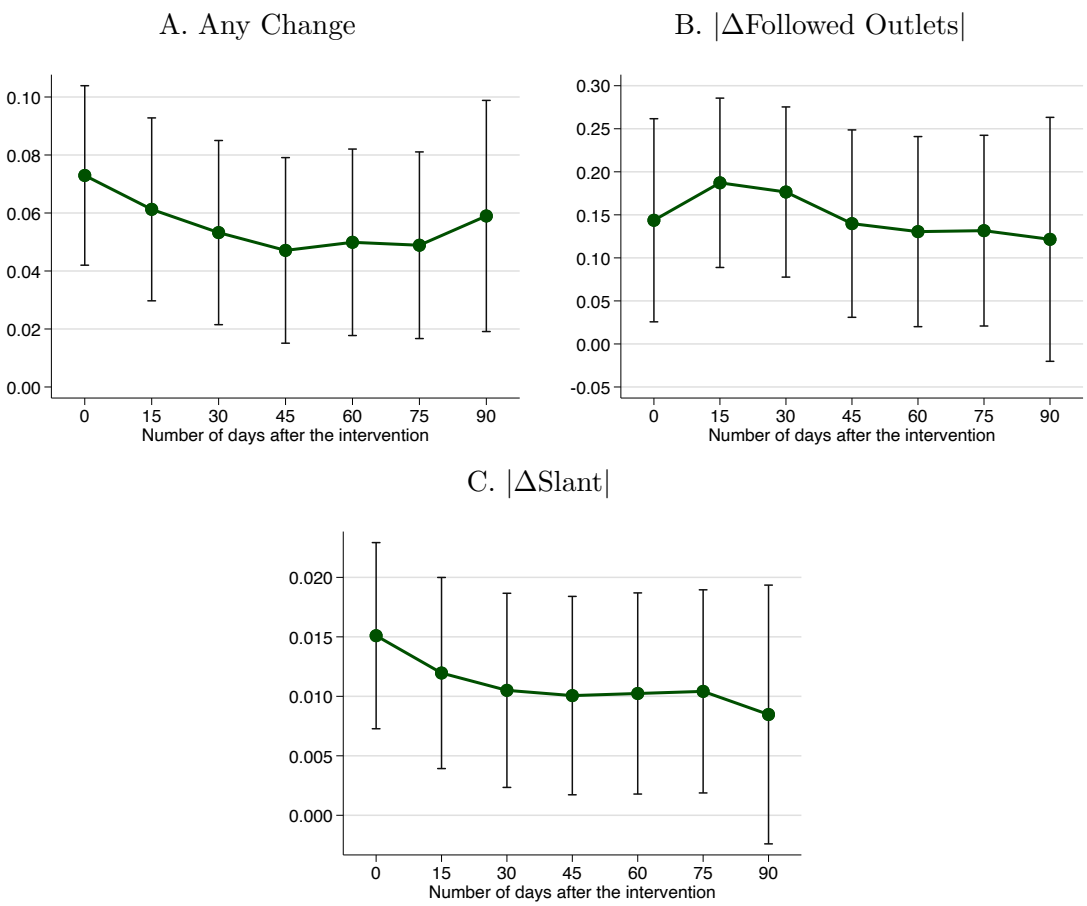
In this section, we analyze the extent to which participants engage with news outlets after the experiment by liking, posting, reposting, quoting or replying to content that was created by or mentions a news outlet. We define *news engagement* as interactions where the individual engages with content either created by or linking to a news outlet in our dataset and classify all other engagement as *non-news engagement*.[18]

Figure 5 plots the effect of the disclosure treatment on news engagement (Figure 5A) and non-news engagement (Figure 5B). The leftmost figure in each panel reports the effect on an engagement index constructed as the average of the standardized values of five components—likes, replies, reposts, posts, and quotes–where each component is standardized by subtracting the control-group mean and dividing by the control-group standard deviation. The remaining figures report the effects on each component separately. We find that the disclosure treatment leads to a significant increase in overall engagement with news outlets, while having no statistically significant effect on engagement with non-news outlets. We further find that the increase in overall engagement with news outlets induced by the disclosure treatment is explained by higher engagement with outlets participants followed at baseline and with outlets suggested to participants during the experiment (Figure A-12)

The findings presented in this section suggest the disclosure treatment has important effects on long-term engagement. Participants in the disclosure treatment group are not simply following or unfollowing outlets to curate a news diet for the experiment and then reverting back to their original news diet. The changes they make persist for months

[18]We classify a post as news engagement if it mentions (i.e. tags) a news outlet. We classify a reply or repost as news engagement if it is in response to a post created by a news outlet, or if the post to which it responds mentions a news outlet. We classify a quote as news engagement if it quotes a post created by a news outlet, or if either the quoting post or the quoted post mentions a news outlet. Finally, we classify a like as news engagement if the liked item–whether a post, reply, repost or quote–is itself classified as news engagement under the above rules.

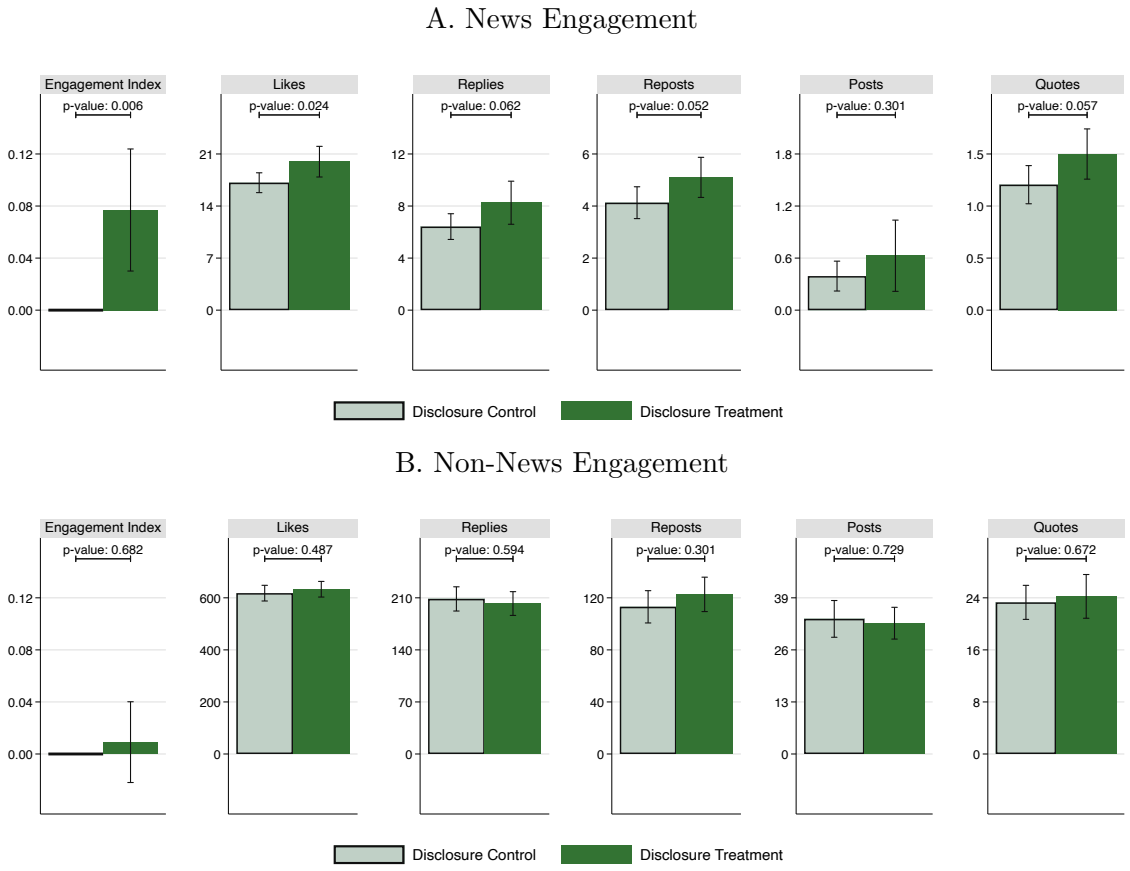
Figure 4: Effect of Disclosure Condition Over Time



Notes: This figure reports treatment effects of the disclosure condition for the 90 days after the intervention. Panels A, B, and C report effects on an indicator for whether the participant makes any change to the news outlets they follow, the absolute change in the number of outlets followed, and the change in the absolute slant of the outlets followed, respectively. In all cases, we construct these outcomes by comparing the set of outlets participants follow at the specified date (see x-axis) with those followed at baseline. The difference between the treatment and control groups in the disclosure condition is that only participants in the former group are asked to reveal to their peers which news outlets they follow. The whiskers indicate the 95% confidence intervals.

after the experiment, and they engage more with the outlets they follow. Together, these findings suggest that social image concerns are important determinants of the news outlets individuals choose to follow and engage with, and these changes persist over time.

Figure 5: Effect of Disclosure Condition on Active Engagement



Notes: This figure reports the mean number of reposts, likes, and posts across participants assigned to the treatment and control groups in the disclosure condition. Panels A and B report these metrics separately for news and non-news engagement, where an interaction (repost, like, or post) is classified as news engagement if it is made in response to a news outlet or directly mentions one. The difference between the treatment and control groups in the disclosure condition is that only participants in the former group are asked to reveal to their peers which news outlets they follow. The whiskers indicate the 95% confidence intervals. We also report the p-value for the null hypothesis of no difference between the treatment and control groups.

4.6 Robustness

We now assess the robustness of the disclosure treatment effect estimates. First, we study heterogeneity in the treatment effect of the disclosure treatment depending on whether or

not a participant was in the peer information condition. Appendix A.4 reports the main estimates from a fully saturated regression in which we simultaneously include both the disclosure and peer information treatment conditions. The interaction term between the two treatment conditions—peer information and disclosure—is quantitatively small and statistically insignificant. Therefore, the approach we take of estimating the treatment effect of each treatment in isolation gives quantitatively similar results.

We also show that our main findings remain unchanged when we account for observable differences in socio-demographic characteristics between our sample and the U.S. population, and thus are unlikely to be driven by sample selection (Appendix A.6).

Above we estimate the ITT effect of the disclosure condition on news diets. One could also estimate the treatment effect of disclosure on news diets by instrumenting the disclosure behavior with the disclosure treatment. As we discuss in Appendix A.5 this analysis relies on the stronger assumption that treatment assignments affects news diets only through the actual disclosure decision. Due to this exclusion restriction, our preferred estimates are the ITT results which require no such assumption. The instrumental variable analysis implies that increasing compliance rates by 1 percentage point leads to a 0.89 percentage point increase in the probability treated participants would adjust their news diets. This implies that roughly nine in ten individuals adjust their news diets when sharing on the platform in response to social image concerns.

Tables A-3 and A-4 examine the robustness of our main results, reported in Figures 2–5, to the inclusion of covariates. In both tables, we follow the same approach. Column (1) reports our baseline estimates without additional controls. Column (2) includes the set of self-reported demographic characteristics collected at baseline: year of birth and indicators for gender, education status, ethnicity, and state of residence. Column (3) adds the full set of self-reported measures of political engagement and political knowledge. Column (4) includes self-reported ideology; news ideology for the participant, their followers, and the users they follow; and thermometer feeling questions on political parties and candidates. Column (5) adds the slant of the participant's account and that of their peers, a credibility index of the news consumed, the number of followers and users followed, and dummy variables for the device used to complete the survey. Column (6) includes all aforementioned controls. Finally, Column (7) reports estimates from the double machine learning approach of Chernozhukov et al. [2018] (algorithm DML1), selecting control variables from the full set of available variables. Across all specifications, we find that our estimated coefficients remain remarkably stable with the inclusion of controls.

5 Peer Information

In this section, we study how the information that individuals receive about the news diets of their peers influences their own news diet. We begin by examining the extent to which the peer signal—an estimate of the average slant of the news diets of a participant's peers—shifts participants' *beliefs* about the news diets of their social networks. Next, we analyze how these beliefs influence the *news diets* of participants themselves. We conclude by exploring alternative explanations to our preferred interpretation and conducting additional robustness checks.

5.1 Does Peer Information Influence Beliefs?

We first examine whether participants revise their beliefs about the ideological slant of the news diets of their peers when presented with an estimate of the average slant of their peers' news diets. This is identified from variation between participants in the peer information treatment who receive these peer signals and those in the control group who do not. We measure prior and posterior beliefs on a seven-point scale ranging from "extremely liberal" to "extremely conservative."

We find that the peer information treatment substantially increases the probability a participant adjusts their beliefs about the average slant of the news diets of their peers relative to the control group (32% more likely to update beliefs). The peer information treatment also increases the probability a participant makes a large change in their beliefs about the news diets of their peers (Figure A-13).

While the analysis above demonstrates that participants respond to the peer signal, it does not reveal whether the peer signal induces participants to update their beliefs in the direction of the signal. To test this, we transform the key outcome—the change in a participant's beliefs[19]—by multiplying it by -1 when the signal lies to the left of the prior.[20] A positive (negative) treatment effect on this "signal-oriented" outcome indicates

[19]Change in beliefs is always defined as the posterior belief measured at endline minus prior belief measured at baseline.

[20]Formally, the transformation is defined as follows:

$$\widetilde{y}_i = \begin{cases} y_i & \text{if } \sigma_i^{\text{signal}} \geq \sigma_i^{\text{prior}} \\ -y_i & \text{if } \sigma_i^{\text{signal}} < \sigma_i^{\text{prior}} \end{cases} \tag{3}$$

where y_i is the outcome of interest (e.g., the change in a participants beliefs about their peers' news diets) and $\widetilde{y}_i$ its signal-oriented counterpart. Because we originally measure the prior (σ_i^{prior}) and signal (σ_i^{signal})

that the peer information treatment induces a participant to update their beliefs about their peers in the same (opposite) direction as the *surprise*.[21]

Panel A of Table 1 shows the treatment effects of the peer information treatment on the signal-oriented change in participant beliefs about the news diets of their peers. Column (1) indicates that treated participants update their beliefs in the direction of the surprise by a larger amount than the control group. This effect is statistically significant and suggests on average participants in the treatment group update their beliefs twice as much in the direction of the surprise relative to those in the control group.[22]

Column (2) examines heterogeneity in the treatment effect by the magnitude of the surprise (normalized by its standard deviation). When the prior and signal coincide (zero surprise), the treatment effect is negligible and not statistically distinguishable from zero. As the surprise increases in magnitude, the treatment effect increases: a one–standard deviation increase in the absolute surprise raises the treatment effect by 0.21 points on the seven-point scale, a 72% increase relative to the average treatment effect.

Columns (3) and (4) show the results of this analysis are similar to that in Columns (1) and (2) if we focus on the direction of belief updates rather than their magnitude.

The above results use variation from the random assignment of the peer information treatment. We now turn to the second source of variation—signal precision—captured by equation (2). As discussed in Section 2.4, the peer signal is an unbiased but noisy measure of the average slant of the news diet of a participant's peers. This random noise provides exogenous variation that we use to estimate the causal effect of receiving information suggesting that one's peers are more right- or left-leaning than they actually are on participants' beliefs about the news diet of their peers.

Panel B of Table 1 presents the corresponding results for participants assigned to the peer information treatment group, who receive the signal. The coefficient in Column (1) captures the effect of (randomly) receiving a signal with right noise on changes in participants' beliefs about the average slant of the news diets of their peers. Participants exposed to a signal with right noise—where the signal randomly lies to the right of the truth—shift

on different scales, we rescale the signal to match the prior's scale. Appendix A.8 details the procedure and shows robustness to alternative normalizations. We also note that peer signals were estimated for all participants during the experiment including those where the signal was not revealed to participants, ensuring this transformation is well-defined for the entire sample.

[21] We use the term *surprise* throughout to refer to the difference between the signal and the prior.

[22] Participants in the control group also update their beliefs in the direction of the surprise despite being exposed to it, which could be explained by participants recalibrating their beliefs in response to the information contained in their own news diet.

Table 1: Effects of Peer Signal on Beliefs About Peers' News Diet Slant

	(1)	(2)	(3)	(4)
	Change in Beliefs About Peer's News Diets			
	Slant Change		sign(Slant Change)	
A. Treatment Effects on Signal-Oriented Outcomes, Full Sample (N=2,695)				
Peer Treatment	0.288	0.027	0.212	0.066
	(0.043)	(0.071)	(0.029)	(0.048)
|Signal − Prior|		0.161		0.058
		(0.052)		(0.024)
Peer Treatment × |Signal − Prior|		0.208		0.116
		(0.057)		(0.027)
Constant	0.267	0.062	0.163	0.089
	(0.038)	(0.065)	(0.026)	(0.043)
B. Effects of a Peer Signal with Right Noise, Peer Treatment Sample (N=2,150)				
Signal with Right Noise	0.453	0.070	0.273	0.041
	(0.046)	(0.065)	(0.030)	(0.043)
|Signal − Truth|		-0.220		-0.138
		(0.035)		(0.020)
Signal with Right Noise × |Signal − Truth|		0.365		0.222
		(0.047)		(0.027)
Constant	-0.099	0.116	-0.057	0.077
	(0.032)	(0.045)	(0.021)	(0.029)
C. Effects of a Peer Signal with Right Noise, Peer Control Sample (N=539)				
Signal with Right Noise	-0.014	-0.081	0.023	-0.064
	(0.079)	(0.124)	(0.053)	(0.079)
|Signal − Truth|		-0.022		-0.026
		(0.056)		(0.039)
Signal with Right Noise × |Signal − Truth|		0.060		0.080
		(0.087)		(0.055)
Constant	-0.008	0.017	-0.049	-0.021
	(0.059)	(0.090)	(0.036)	(0.053)

Notes: This table presents the main effects of the peer signal on changes in beliefs about the slant of peers' news diets. Columns (1) and (2) report effects on the continuous change, and Columns (3) and (4) report effects on the discrete direction of change. Panel A reports the treatment effects of randomly receiving a signal about the slant of the news diets of the participant's peers, following Equation (1). The outcome variables for Panel A are signal-oriented, as defined in Equation (3). These estimates indicate the extent to which the peer signal leads participants to update their beliefs about the slant of their peers' news diets in the direction of the surprise. Panels B and C report the estimated impact of randomly receiving a signal with right-noise, using Equation (2), for participants in the peer information treatment and control groups, respectively. The outcome variables for Panel B and C are not signal-oriented. These estimates indicate the extent to which a signal with right-noise leads participants to shift their beliefs about the news diets of their peers to the right. The even-numbered columns present heterogeneous treatment effects by the magnitude of the surprise, measured in standard-deviation units. Standard errors reported in parentheses are heteroskedasticity robust.

their posterior beliefs to the right. The effect is large and statistically significant. Receiving a peer signal with right noise (as opposed to left noise) leads to a shift in participants' beliefs about the slant of their peers' news diets to the right by 0.45 on the seven point scale. This further confirms that participants find the signal informative and update their beliefs in the direction of the signal using an independent source of variation.

We also test whether the magnitude of the noise affects responsiveness to the noise. Following Panel A, we extend equation (2) to include the absolute value of the noise (in standard deviations) and an interaction of the magnitude of noise with an indicator if the signal contained right noise. Column (2) shows participants' beliefs update more in the direction of the noise when the noise component of the signal is larger. Columns (3) and (4) demonstrate the results are consistent when we focus on the discrete direction of changes.

Finally, Panel C repeats the analysis for participants assigned to the peer information control group, who serve as a placebo since they do not observe the signal and thus are not exposed to the noise.[23] Across all specifications, the coefficients are statistically indistinguishable from zero and are much smaller than those in Panel B as we expect.

Overall, the evidence in this section suggests that the peer signal influences participants beliefs about the news diets of their peers.

5.2 Does Peer Information Affect News Diets?

We now investigate whether changes in a participant's beliefs about the slant of the news diets of their peers impact their own news diet. We assess this by estimating the treatment effect of receiving the peer signal—which we show above shifts participant beliefs about the news diets of their peers—on changes to a participant's own news diet.

Table 2 presents estimates parallel to the results discussed in Section 5.1, but uses the change in the *observed* slant of the participant's news diet as the outcome rather than beliefs about the slant of the news diets of their peers.

The coefficient in Column (1) of Panel A is precisely estimated and not statistically significant, indicating that the peer signal does not affect the participant's own news diet. The estimated effects are also small, corresponding to less than 3% of a standard deviation of the outcome in the control group. This contrasts sharply with the large impact of the peer signal on participants' beliefs about the news diets of their peers, which amounts to

[23]We pre-registered that 80% of individuals would be randomized into the peer information treatment and 20% into the peer information control to increase the power of the analysis in Panel (B). Therefore, the analysis in panel (C) is based on a smaller sample (N=539).

33% of a standard deviation of directional change in beliefs in the control group.[24]

Column (2) shows no evidence of a differential treatment effect on the slant of an individual's news diet between those who receive signals further from their prior belief. Similarly, Columns (3) and (4) result in similar conclusions when focusing on discrete rather than continuous changes.

Panel B reports the results exploiting variation in the sampling noise of the peer signal for the treatment group. The small and non-significant coefficient on the Signal with Right Noise variable in Column (1) implies that individuals who receive a signal with right-noise do not adjust their own news diets in response to this information. This occurs despite the fact that they do update their beliefs about the news diets of their peers (Table 1). The estimated effects are also small relative to the corresponding effects on beliefs about the news diets of their peers presented in Table 1, amounting to 1% as opposed to 21% of a standard deviation of the respective outcome in the control group.

We likewise find no evidence that these results vary with the magnitude of the noise (Column (2)) or when focusing on discrete rather than continuous changes (Columns (3) and (4)). Finally, Panel C shows no effects in our placebo exercise, which replicates the analysis in Panel B for the peer-information control group who did not receive the signal.

Across the various approaches we take to identify the effect of peer information on news diets we find remarkably consistent results. We find that participants significantly adjust their beliefs about the news diets of their peers after receiving information about the slant of their peers' news diets. However, despite this notable shift in beliefs, we find no evidence of meaningful changes in participants' own news diets in response. This suggests the direct peer influence channel does not substantially influence news diets. In the remainder of the section, we discuss alternative interpretations and additional robustness checks.

5.3 Alternative Interpretations

We interpret the results thus far as indicative of peer information not influencing individuals' own news diets. We now discuss the limitations of our analysis and provide evidence that is inconsistent with the most likely alternative explanations.

[24]Figure A-14 reports results for key outcomes in absolute value. This captures whether there is any change in participant news diets, but not the direction of change. Consistent with Table 2, we find no evidence that the peer treatment significantly affects participants' news diets. Treatment effects are small, negative, and not statistically distinguishable from zero. One outcome is marginally significant—the absolute value of slant change—with a p-value = 0.07. All treatment effects are much smaller in magnitude than the effect of our disclosure treatment on the same set of outcomes, as reported in Figure 2.

Table 2: Effects of Peer Signals on Participants' Own News Diet Slant

	(1)	(2)	(3)	(4)
	Change in Participants' Own News Diets			
	Slant Change		sign(Slant Change)	
A. Treatment Effects on Signal-Oriented Outcomes, Full Sample (N=3,683)				
Peer Treatment	-0.003	-0.007	-0.013	0.002
	(0.005)	(0.010)	(0.026)	(0.042)
|Signal − Prior|		-0.002		0.003
		(0.005)		(0.023)
Peer Treatment × |Signal − Prior|		0.003		-0.011
		(0.005)		(0.025)
Constant	0.001	0.005	0.006	0.002
	(0.005)	(0.010)	(0.023)	(0.038)
B. Effects of a Peer Signal with Right Noise, Peer Treatment Sample (N=2,951)				
Signal with Right Noise	-0.001	-0.005	-0.018	-0.015
	(0.004)	(0.006)	(0.022)	(0.033)
|Signal − Truth|		-0.001		-0.007
		(0.003)		(0.018)
Signal with Right Noise × |Signal − Truth|		0.003		-0.002
		(0.004)		(0.023)
Constant	-0.003	-0.002	-0.009	-0.002
	(0.003)	(0.005)	(0.016)	(0.022)
C. Effects of a Peer Signal with Right Noise, Peer Control Sample (N=722)				
Signal with Right Noise	0.004	0.008	0.042	0.110
	(0.010)	(0.013)	(0.047)	(0.065)
|Signal − Truth|		0.004		0.048
		(0.003)		(0.030)
Signal with Right Noise × |Signal − Truth|		-0.004		-0.063
		(0.006)		(0.044)
Constant	-0.007	-0.012	-0.039	-0.091
	(0.007)	(0.009)	(0.032)	(0.043)

Notes: This table presents the main effects of the peer signal on changes in the slant of the participants' news diets. Columns (1) and (2) report effects on the continuous change, and Columns (3) and (4) report effects on the discrete direction of change. Panel A reports the treatment effects of randomly receiving a signal about the slant of the news diets of the participant's peers, following Equation (1). The outcome variables for Panel A are signal-oriented, as defined in Equation (3). These estimates indicate the extent to which the peer signal leads participants to update the slant of their news diets in the direction of the surprise. Panels B and C report the estimated impact of randomly receiving a signal with right-noise, using Equation (2), for participants in the peer information treatment and control groups, respectively. The outcome variables for Panel B and C are not signal-oriented. These estimates indicate the extent to which a signal with right-noise leads participants to shift the slant of their news diets to the right. The even-numbered columns present heterogeneous treatment effects by the magnitude of the surprise, measured in standard-deviation units. Standard errors reported in parentheses are heteroskedasticity robust.

5.3.1 Long-Term Treatment Effects

The main analysis reports effects on changes in participants' news diets during the experiment, comparing changes between the start and the end of the experiment. One alternative explanation is that, while participants update beliefs about the news diets of their peers upon receiving the peer signal, they may not immediately adjust their own news diets, instead gradually adapting over time. This is plausible given the absence of a timing incentive. This is distinct from the disclosure randomization, where participants are prompted to share information with peers during the experiment. This would imply the effect of the peer signal would arise over time as the updated beliefs about the news diets of one's peers are incorporated into subsequent news choices.

To explore this, Table A-5 replicates the estimates from Table 2 but uses participant news diets three months after the intervention to construct the outcome, Figure A-15 plots treatment effects over time on nondirectional outcomes (in absolute value), and Figure A-16 plots effects on news and non-news engagement (reposts, likes, and posts) over the three months after the intervention. Across these analyses, we find no evidence that the change in beliefs in response to the peer signal translate into changes in participant news diets over time.

5.3.2 Content of the Peer Signal

Another plausible interpretation is that, while our peer signal—the average ideological slant of the news diets of a participant's followers—may have limited influence on participants' own news diets, other peer-related information could be more influential. Given the many possible formats in which peer information could be conveyed, we cannot rule out an effect of peer information under alternative designs. Nevertheless, we do see our peer signal influence the beliefs of participants about the news diets of their peers making this signal both informative and a natural benchmark for comparison.

Followers One important feature of our signal is that it is based on peers who follow the participants. We chose to rely on this group rather than the set of accounts the participant follows, as our pilots showed that users often follow many accounts with whom they rarely interact (e.g., celebrities, brands, institutions), making them a noisy proxy for peers. Participants have fewer followers than accounts they follow (the median user follows 500 accounts but has only 148 followers), and many followers are also part of the set of

accounts the participant follows (for the median participant, 44% of their followers are accounts they also follow, yet only 13% of the accounts they follow are also followers). Relative to more complex peer definitions (e.g., the set of followers that the participant also follows, or the accounts the participant frequently interacts with), focusing on followers has the advantage of being a term easily understandable, aiding belief elicitation and clear communication.[25]

Given the considerable overlap between followers and accounts followed, there is little reason to expect that providing a signal based on the latter, or their intersection, would have led to meaningfully different results. We examine this more systematically in Table A-6, exploiting variation in the share of mutual connections between followers and accounts followed (*mutual connection share*) to assess whether peer signals that contain more information about the news diets of the accounts the participant follows correspond to differential treatment effects. The underlying idea is that individuals with a relatively higher mutual connection share should be more likely to update their priors about the accounts they follow in response to our followers-based peer signal and, ultimately, adjust their own news diets if peer information on accounts followed is indeed more influential than peer information on followers.[26]

Panel A shows that the treatment effects of the followers-based peer signal indeed leads participants to update beliefs about the slant of the news diets of users they follow in the direction of the surprise. Yet, this effect is concentrated among participants where the group of followers and accounts followed significantly overlap. More importantly, Panel B shows these updates do not translate into changes in participants' own news choices, reinforcing the robustness of our null effect.

Aggregating Peer News Diets Another potential explanation is that participants may be sensitive to some other aggregation method of peers that places a higher weight on more influential peers, while the peer signal we provide equally weights all peers. Defining influence is challenging, however, we attempt to rule out this possibility in Table A-7, which reports the results of two additional exercises.

[25]Furthermore, some of these definitions of peers would have been infeasible to implement in practice due to API limitations. Ex ante, another benefit of presenting information based on followers rather than accounts followed is that it aligns more directly with our disclosure condition: information shared by participants is likely to appear in the feed of their followers, but not in the feed of the accounts they follow.

[26]Consistent with this idea, we find in our baseline survey that participants with a higher share of mutual connections report a smaller ideological distance between the accounts they follow and the accounts that follow them.

In Panel A, we report the treatment effects of the peer information condition, allowing for a heterogeneous effect for participants where the signal is based on peers with whom the participant regularly engages. This exploits exogenous variation in whether the five randomly selected peers used to construct the signal happen to be among those with whom the participant most frequently exchanges posts, reposts, and likes on the platform. In Panel B, we split the sample by whether participants have a below- or above-median number of followers. A smaller follower base likely increases the probability that the random sample of peers used to construct the signal contains influential or important peers. Stronger effects among participants with small networks would therefore provide indirect evidence that they prioritize information about influential peers.

In both exercises, we are unable to reject that the treatment effect is the same for both subgroups. Importantly, when restricting to the groups for whom the peer signal is arguably more informative about influential peers or peers they engage with, we find no impact of the peer signal on participant news diets—despite a significant effect on participant beliefs about the average slant of the news diets of their peers. This suggests that our null result is unlikely to be driven by participants responding only to influential peers.

Ideology A final possible alternative explanation is that participants are sensitive to their beliefs about the ideology of their peers rather than the slant of their peers' news diets. In Table A-8 we estimate the effect of the peer signal on participant beliefs about the ideology of their peers. In addition to updating their beliefs about the slant of their peers' news diets, this table demonstrates that participants also update their beliefs about the ideology of their peers. We therefore conclude that individuals' beliefs about the ideology of their peers similarly do not impact participant's own news diets.

5.3.3 Experimenter Demand Effects

Our conclusion relies on the result that the peer signal shapes the *actual* beliefs of participants about the news diets of their peers. A concern is that the shifts in beliefs that we observe may reflect a desire to satisfy the experimenter rather than genuine changes in a participant's beliefs. If so, the absence of change in participants' own news diets in response to the signal would not necessarily indicate a lack of peer information effects, but could rather point to a limited capacity of the signal—and, by extension, our design—to capture them. While we cannot fully rule out this hypothesis, we provide several complementary pieces of evidence that are inconsistent with experimenter demand driving the results.

First, our design sought to minimize experimenter demand effects: posterior beliefs were not incentivized, and participants were reminded that honest responses constituted the most valuable contribution to the study. Second, participants appeared highly engaged with the information we provided, suggesting they regarded it as credible. For example, 64% voluntarily shared their email addresses for potential follow-up and to receive further information about their news diets. In addition, 71% completed a set of 16 *optional* questions at the end of the survey, despite the absence of any monetary incentive.

Third, we do observe direct evidence that participants adjusted their own news diets in response to another piece of information we provide: the slant of their own news diets. As Figure A-17 shows, participants who were informed that their own news diet leaned left (right) of their self-reported ideology are observed to follow more right- (left-) leaning outlets during the experiment, thereby aligning their news diet more closely with their ideology. This pattern indicates that participants regarded our information as credible and incorporated it into their news choices. What we do not observe is a comparable adjustment in response to information about their peers.

Fourth, we find no evidence that the ideology of peers influence participant news diets when directly asking participants about the relative importance of various factors influencing the decision to follow or unfollow outlets (Figure A-18). Only 5% reported that the ideological distance between outlets and their friends matters when choosing their news diets. In contrast, 35% pointed to the distance between outlets and their own ideology and 73% emphasized the trustworthiness of outlets. In a follow-up question on the single most important factor, only 2% cited their peers. In contrast, we find that a substantial share of individuals report being sensitive to the information they share consistent with the large impact of social image concerns we estimate in Section 4.

5.4 Robustness

As discussed in Section 3, our main sample includes participants who were exposed to the treatment conditions but did not necessarily complete the endline survey. This explains the difference in the number of observations between Tables 1 and 2. Table A-9 replicates Table 2 for the subsample of participants who responded the endline survey and confirms that our conclusions remain unchanged.

Appendix A.4 shows that our results on the peer treatment are very similar when estimating an extended regression that simultaneously estimates the effects for the disclosure

condition. In addition, Appendix A.6 shows that the findings in this section are unchanged when adjusting our sample to match observable characteristics of the U.S. population.

Appendix A.5 discusses an instrumental variable analysis, where treatment assignment is used to instrument changes in participants' reported beliefs about the slant of their peers' news diets. The 2SLS estimates are small in magnitude and statistically insignificant, consistent with our earlier findings.

Finally, Tables A-10 and A-11 assess the robustness of the main results in this section (Tables 1 and 2) to the inclusion of additional covariates—ranging from specific groups of baseline variables to the full baseline set—as well as to estimation using the double machine learning approach of Chernozhukov et al. [2018]. Across all specifications, the estimated coefficients remain remarkably stable.

6 A Natural Experiment on Social Image Concerns

In this section, we analyze a policy change made by the social media platform X that provides a unique opportunity to study how social image concerns shape engagement with biased news content. This natural experiment complements our field experiment and allows us to examine the role of social image concerns in everyday platform use.

6.1 Policy Change in the Visibility of Endorsements

On June 12, 2024, the X platform introduced a policy change that made the action of *liking a post* private to everyone except for the user liking the post and the author of the post. Thus, as before the intervention, any user can still see the total count of *Likes* on a post. However, unlike before the intervention, a user can no longer see which other users (including users they follow and who follow them) liked a post made by anyone other than themselves. This change was widely advertised by X engineers through several posts and garnered considerable attention from dozens of media outlets, as well as millions of people.[27]

[27]These posts for instance received substantial attention on the platform, with at least one reaching more than 230 million views `https://x.com/XEng/status/1800634371906380067`. For reference, this number is twice the total active number of U.S. users on the platform and around 38% of worldwide users (see `https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/` and `https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/`.) Interest outside the platform was also notable, as documented in several Reddit posts and online blogs. According to Google Trends, searches for the terms "X Likes" or "Twitter Likes" in the search engine spiked more than sixfold during the month of the intervention, reaching its highest level since 2004.

By altering the public visibility of user endorsements, this intervention provides an opportunity to test the importance of social image concerns in shaping news engagement in a natural setting. In fact, public statements by platform managers suggest that affecting visibility and social image concerns is a primary motivation for why X implemented this change in the first place. In a post justifying the change, Haofei Wang, X's Director of Engineering, stated, "Public likes are incentivizing the wrong behavior. For example, many people feel discouraged from liking content that might be 'edgy' in fear of retaliation from trolls, or to protect their public image."[28]

As explained below, we leverage the plausibly exogenous change in the visibility of likes to identify the causal effect of social image concerns on the probability of liking a post. To do so, we collected the universe of over nine million posts created in 2024 by all news outlets in our database.[29] For each post, we observe the exact timestamp the post was created as well as engagement metrics: the number of views, likes, reposts, replies, quotes, and bookmarks. The timestamp allows us to identify whether a post was created before or after the policy change was implemented. Table A-12 presents summary statistics for the data. On average, posts in our sample received 8,822 views, 25 likes, 7 reposts, 5 replies, 1 bookmark, and 1 quote.

6.2 Identifying the Impact of the Policy Change

For every post j, we observe the number of users who view the post (V_j) and the number of users who like, repost, bookmark, reply to, and quote a post (denoted L_j, R_j, B_j, C_j, and Q_j, respectively). This allows us to calculate the probability a user takes each action conditional on being exposed to the post. The probability of liking a post conditional on exposure is given by $p_j^L = L_j/V_j$. We can similarly define the probability of reposting ($p_j^R = R_j/V_j$), bookmarking ($p_j^B = B_j/V_j$), replying ($p_j^C = C_j/V_j$), and quoting ($p_j^Q = Q_j/V_j$) a post.

Here we introduce notation and assumptions that allow us to identify the effect of social image concerns on the probability of liking a post. Let $p_j^L(A)$ denote the potential outcome for the probability that a user likes post j under treatment status $A \in \{0, 1\}$, where A is a dummy variable equal to one after the policy change. We assume that the potential

[28] See `https://x.com/wanghaofei/status/1793096366132195529`

[29] We excluded 37,577 posts (approximately 0.4%) that were not original content from these news outlets but rather reposts from other users, as X metrics for reposts reflect engagement with the original post rather than the repost itself.

outcome before the policy change is a function of the probability of non-like actions:

$$p_j^L(0) = g\left(p_j^{-L}, X_j\right) + \varepsilon_j \tag{4}$$

where $g\left(p_j^{-L}, X_j\right) = E\left[p_j^L | p_j^{-L}, X_j\right]$, p_j^{-L} is the vector of non-like actions for post j, X_j is a vector of covariates for post j, and ε_j is a mean-zero error term. Importantly, we assume that this conditional mean is structural and, absent the policy change, the relationship between the probability of non-like actions and the probability of liking a post would have been the same. We further assume an additive treatment effect such that

$$p_j^L(1) = p_j^L(0) + \tau(X_j) \tag{5}$$

where $\tau(X_j)$ is the treatment effect. We allow for treatment effects to depend on post characteristics X_j. Critically, the policy change only affects the visibility of likes and not the visibility of reposts, bookmarks, replies, or quotes. Finally, we map potential outcomes to the observed data as $p_j^L = (1 - A_j)p_j^L(0) + A_j p_j^L(1)$ where A_j is an indicator equal to one if post j occurred after the policy change.

6.2.1 Discussion of Identification Strategy

This identification strategy is a generalized version of the synthetic control method [Abadie et al., 2015]. In principle, $g(\cdot)$ can be a non-linear function, though we will impose a linear functional form which fits the data well as we show in Section 6.3.1. This model is a version of the model analyzed in Doudchenko and Imbens [2016], who generalize the synthetic control method to allow for negative weights, an intercept term, and weights that do not sum to one. In addition, we allow for weights to depend on post features X_j.

We allow for these deviations from the more standard synthetic control method for several reasons. First, as shown in Table A-12, the probability of liking a post is much higher than the probability of taking any non-like action. Therefore, any convex combination of the probability of non-like actions would have an extremely poor fit. In addition, given our dataset consists of over nine million observations, we are able to flexibly fit the pre-treatment outcomes without overfitting and while still withholding several pre-treatment months. This allows us to compare the out-of-sample fit to test the assumption that the estimated relationship between $p^L(0) = g\left(p_j^{-L}, X_j\right)$ is structural.

6.3 Estimation and Results

6.3.1 Do Social Image Concerns Really Matter?

We take this model to the data by splitting our sample into three periods. The first three months of 2024 constitute a training period where we estimate the function $g(\cdot)$ before the policy was implemented. The period April 1, 2024 through June 11, 2024 is the test period where we evaluate the assumption that $g(\cdot)$ represents a good counterfactual for the probability of liking a post without the policy change. Finally, the period after June 12, 2024 contains the treated period after the policy change. Comparing the actual probability of liking a post to the counterfactual probability allows us to identify the treatment effect of the policy change on the probability of liking a post.

We estimate $g(\cdot)$ using a weighted least squares regression of p_j^L on p_j^{-L}, the absolute slant of the news outlet creating the post j discretized into quartiles, and the interaction of the slant quartile indicators with p_j^{-L}. Figure 6A plots the raw time series of p_j^L and $g\left(p_j^{-L}, X_j\right)$. The vertical lines separate the train, test, and treatment periods. The test period provides an out-of-sample test of the assumption that $g\left(p_j^{-L}, X_j\right)$ is structural. The model fits the data well in the test period, consistent with our identification assumption. Figure 6B plots the difference in the actual and counterfactual like probabilities. We cannot reject the joint test that all pre-treatment differences are zero ($p = 0.53$).

We see a clear increase in p_j^L relative to the counterfactual ($g\left(p_j^{-L}, X_j\right)$) shortly after the treatment indicating that the policy change did indeed increase the total number of likes. We also estimate the average treatment effect of the policy change during the six months after its introduction, by estimating the following regression:

$$100 \times \left(p_j^L - g\left(p_j^{-L}, X_j\right)\right) = \tau_0 + \tau Post_j + \varepsilon_j. \tag{6}$$

where τ is the average treatment effect and $Post_j$ is a dummy variable equal to one if post j is created after the intervention. Column (1) of Table 3 presents the results and shows that the policy change that privatized likes increases the probability of liking a post conditional on exposure by 0.03 percentage points. This effect is not only statistically significant (standard error of 0.007), but also economically large, representing a 12.2% increase relative to the pre-treatment mean.[30]

[30]There was a presidential election in November 2024, 5 months after the policy change. Our estimated treatment effect of the policy change is substantially larger during this month. To rule out concerns that our treatment effect estimates are driven by this election month, we estimate the average treatment effect

Overall, these findings provide clear evidence that social image concerns shape user engagement with news outlets during regular social media use, complementing the broader results from Section 4.

6.3.2 Do Social Image Concerns Moderate News Engagement?

A key finding in our field experiment is that social image concerns moderate individual news diets (Section 4.3). We now assess the generalizability of this result in the context of the policy change that made likes private. We do so by estimating the heterogeneous treatment effect of the policy by the political slant of the news outlet creating the post. If social image concerns cause individuals to moderate their news diets, then making likes private—which alleviates the social image concerns—should disproportionately increase the likes received by more politically slanted outlets.

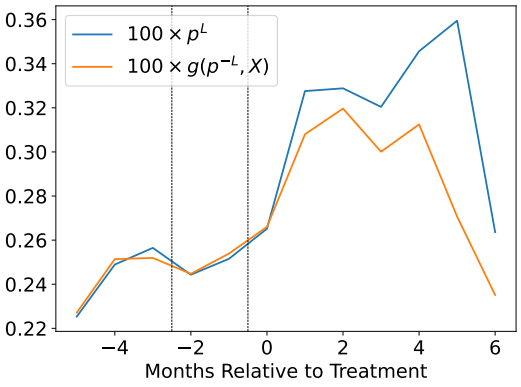
Figure 6C plots the dynamic treatment effects of the policy change conditional on the quartile of the absolute slant of the news outlet publishing the post. The first quartile corresponds to the most politically neutral outlets and the fourth quartile corresponds to the most politically slanted outlets. We find that the treatment effect of concealing likes is significantly larger for posts from the most politically slanted news outlets relative to neutral outlets. Furthermore, Column (2) of Table 3 report heterogeneous treatment effect estimates based on Equation (6) with additional fixed effects for the quartile of absolute slant and their interaction with the $Post_j$ dummy. These estimates imply that the effect of the policy making likes private on the probability of liking a post is five times larger for posts created by the most politically slanted outlets relative to posts from the most neutral ones. Column (4) reports the same heterogeneous treatment effect estimates using only data before the 2024 elections and we see the same qualitative patterns.

Overall, the results in this section align closely with the findings from our field experiment. Namely, we find in both cases that social image concerns lead individuals to moderate their news diets. This strengthens the case that these results generalize beyond the controlled experimental setting.

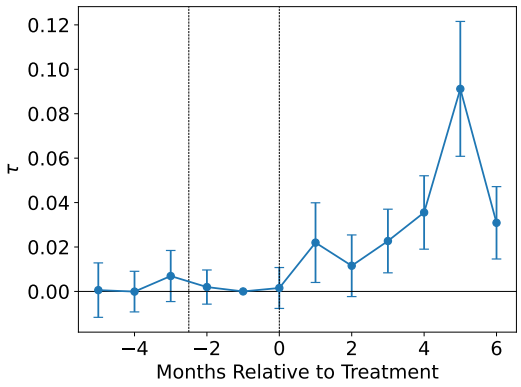
in Column (3) using only data through October 2024. While the magnitude of the treatment effect is lower, the qualitative results remain.

Figure 6: Effect of Making Likes Private on Engagement with News Outlets

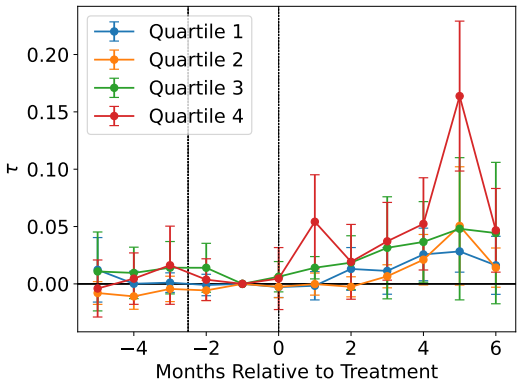
A. Time Series of p_j^L and $g\left(p_j^{-L}, X_j\right)$



B. Event Study



C. Event Study by Outlet Slant



Notes: This figure shows the main estimates of the effect of the policy change that made *likes* private on the number of likes received by all posts published by U.S. news outlets. Panel A plots the time series of the average $100 \times p_j^L$ and the average fitted values of $100 \times g\left(p_j^{-L}, X_j\right)$, both weighted by the number of views. Panel B plots an event study comparing the difference in the average $100 \times p_j^L$ and the average fitted values of $100 \times g\left(p_j^{-L}, X_j\right)$, again weighted by the number of views. Panel C plots the same event study as panel B separately for quartiles of news outlet slant. In both Panel B and Panel C, the month before the policy was changed is normalized to zero and the vertical bars represent 95% confidence intervals based on bootstrapped standard errors that are clustered at the news outlet level. All estimates are reported in percentage points.

Table 3: Treatment Effect of Concealing Likes on the Probability of Liking a Post

	(1)	(2)	(3)	(4)
	100 × Probability of Liking a Post			
	All Data January-December (2024)		Data Before Elections January-October (2024)	
Constant	-0.001	-0.002	-0.001	-0.002
	(0.008)	(0.011)	(0.008)	(0.011)
2nd Quartile		0.004		0.004
		(0.013)		(0.013)
3rd Quartile		0.000		0.000
		(0.019)		(0.019)
4th Quartile		0.001		0.001
		(0.023)		(0.023)
Post	0.030	0.010	0.017	0.006
	(0.007)	(0.006)	(0.007)	(0.005)
Post × 2nd Quartile		0.010		0.005
		(0.012)		(0.008)
Post × 3rd Quartile		0.008		0.005
		(0.010)		(0.008)
Post × 4th Quartile		0.044		0.024
		(0.015)		(0.017)
Pre-Treatment Average ($100 \times p_j^L$)	0.245	0.245	0.245	0.245
Observations	9,073,600	9,073,600	7,563,526	7,563,526
R^2	0.002	0.003	0.001	0.001

Notes: This table presents estimates of the effect of the policy change that made *likes* private on the probability a user likes a news outlet's post conditional on viewing the post in percentage points. Coefficients are based on Equation (6), estimated using weighted least squares in which observations are weighted by the total number of views. The unit of analysis is a post, and the sample consists of the universe of posts created by news outlets in 2024 on the X platform. Even columns include fixed effects for the quartile of the absolute slant of the news outlet and interact the *Post* dummy with these quartile fixed effects. Standard errors calculated from a bootstrap clustered at the outlet level.

7 Conclusion

There is a widespread belief that interactions with like-minded peers tend to limit exposure to differing viewpoints and can reinforce existing beliefs, thus polarizing societies. However, there is limited empirical evidence of how these interactions lead to changes in the demand for biased information. Even less is known about the mechanisms through which this happens. In this paper, we study the role of two important mechanisms – social image concerns and peer information – in explaining preferences for biased news. Our field experiment on X induced variation in both an individual's perceptions of the political leanings of the news diets of their peers and the visibility of their own news diets to their social media followers.

We find little support for the peer information channel: changes in beliefs about the political slant of the news consumed by peers have minimal impact on individuals' own news diets. By contrast, individuals significantly adjust their news diets when they believe their behavior is observable to peers, indicating that social-image concerns constitute an important mechanism. We further exploit a plausibly exogenous policy change on the X platform that altered the visibility of endorsements and find consistent evidence that engagement with news content responds to these social-image incentives.

Our study has important policy implications. It provides evidence that interactions with peers can increase the demand for more moderate content. As we demonstrate, individuals care about how their peers perceive the content with which they engage. Furthermore, individuals significantly value demonstrating to their peers that they consume more moderate news relative to their private bliss point. Therefore, by amplifying the visibility of user interactions, social media can help moderate the content that users choose to consume. Importantly, our results indicate that these effects are not only statistically significant but also substantial in magnitude. Thus, they must be carefully considered in the formulation of policies designed to mitigate polarization. To the extent that encouraging users to consume balanced news diets is a desirable goal, our results suggest clear policies that would encourage this behavior. Given the significant influence of the social image concern channel and the moderating impact of the incentive to publicize news diets, fostering enhanced transparency on platforms can moderate the news consumption patterns of users. Features that balance user preferences for privacy and foster transparency of news choices on social media can have desirable effects.[31] On the other hand, as we show in Section 6,

[31]The blindspotter tool by ground.news is one such example https://ground.news/blindspotter/twitter.

policies that mitigate social-image concerns induce individuals to interact more with biased content.

Despite evidence that social influence mitigates demand for polarized news content through the two key channels discussed in this paper, further research is needed to understand its aggregate effects on polarization. First, social media is not only believed to affect polarization through the creation of echo chambers. To maximize engagement, ranking algorithms may limit individual exposure to counter-attitudinal content. The degree to which social media algorithms expose users to segregated content, and how this exposure contributes to polarization, form a central pillar of debate in recent research Levy [2021]. Further evidence is needed to deepen our understanding of this phenomenon [González-Bailón et al., 2023, Guess et al., 2023]. Second, interactions with others may affect polarization through channels not directly studied here. For example, our study does not consider the impact of direct conversations with peers on news diets. The importance of these alternative mechanisms is also a promising direction for future research. Finally, while our results speak to polarization in news consumption, an open question is how these mechanisms influence other dimensions of polarizing behavior, including affective polarization, ideological extremism, and hostility toward political outgroups.

References

A. Abadie, A. Diamond, and J. Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510, 2015.

Z. Ahmed, S. Amizadeh, M. Bilenko, R. Carr, W.-S. Chin, Y. Dekel, X. Dupre, V. Eksarevskiy, S. Filipi, T. Finley, et al. Machine learning at microsoft with ml. net. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2448–2458, 2019.

H. Allcott, L. Braghieri, S. Eichmeyer, and M. Gentzkow. The welfare effects of social media. *American Economic Review*, 110(3):629–76, 2020.

H. Allcott, M. Gentzkow, and L. Song. Digital addiction. *American Economic Review*, 112 (7):2424–2463, 2022.

S. Aral and M. Zhao. Social media sharing and online news consumption. *Available at SSRN 3328864*, 2019.

S. Athey, M. Mobius, and J. Pal. The impact of aggregators on internet news consumption. Technical report, National Bureau of Economic Research, 2021.

E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132, June 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaa1160. URL `https://science-sciencemag-org.libproxy.mit.edu/content/348/6239/1130`. Publisher: American Association for the Advancement of Science Section: Report.

L. Boxell, M. Gentzkow, and J. M. Shapiro. Cross-country trends in affective polarization. *Review of Economics and Statistics*, pages 1–60, 2022.

L. Braghieri, S. Eichmeyer, R. Levy, M. Mobius, J. Steinhardt, and R. Zhong. Article level slant and polarization of news consumption on social media. *Available at SSRN*, 4932600, 2024.

D. Broockman and J. Kalla. The impacts of selective partisan media exposure: a field experiment with fox news viewers. *OSF Preprints*, 2022.

J. Bruner. Tweets loud and quiet. *O'Reilly Radar. Retrieved October*, 6:2015, 2013.

L. Bursztyn and R. Jensen. How does peer pressure affect educational investments? *The quarterly journal of economics*, 130(3):1329–1367, 2015.

L. Bursztyn and R. Jensen. Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure. *Annual Review of Economics*, 9:131–153, 2017.

L. Bursztyn, T. Fujiwara, and A. Pallais. 'acting wife': Marriage market incentives and labor market investments. *American Economic Review*, 107(11):3288–3319, 2017.

L. Bursztyn, G. Egorov, and S. Fiorin. From extreme to mainstream: The erosion of social norms. *American economic review*, 110(11):3522–3548, 2020a.

L. Bursztyn, A. L. González, and D. Yanagizawa-Drott. Misperceived social norms: Women working outside the home in saudi arabia. *American economic review*, 110(10):2997–3029, 2020b.

A. Casas, E. Menchen-Trevino, and M. Wojcieszak. Exposure to extremely partisan news from the other political side shows scarce boomerang effects. *Political Behavior*, pages 1–40, 2022.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

F. Chopra, I. Haaland, and C. Roth. Do people demand fact-checked news? evidence from us democrats. *Journal of Public Economics*, 205:104549, 2022.

F. Chopra, I. Haaland, and C. Roth. The demand for news: Accuracy concerns versus belief confirmation motives. *NHH Dept. of Economics Discussion Paper*, (01), 2023.

S. DellaVigna, J. A. List, U. Malmendier, and G. Rao. Voting to tell others. *The Review of Economic Studies*, 84(1):143–181, 2016.

N. Doudchenko and G. W. Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research, 2016.

R. Enikolopov, A. Makarin, and M. Petrova. Social media and protest participation: Evidence from russia. *Econometrica*, 88(4):1479–1514, 2020.

J. Fardouly and L. R. Vartanian. Social media and body image concerns: Current research and future directions. *Current opinion in psychology*, 9:1–5, 2016.

R. G. Fryer Jr and P. Torelli. An empirical analysis of 'acting white'. *Journal of Public Economics*, 94(5-6):380–396, 2010.

P. Funk. Social incentives and voter turnout: evidence from the swiss mail ballot system. *Journal of the European economic association*, 8(5):1077–1103, 2010.

M. Gentzkow and J. M. Shapiro. Media bias and reputation. *Journal of political Economy*, 114(2):280–316, 2006.

M. Gentzkow and J. M. Shapiro. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71, 2010.

M. Gentzkow and J. M. Shapiro. Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839, 2011.

A. S. Gerber, D. P. Green, and C. W. Larimer. Social pressure and voter turnout: Evidence from a large-scale field experiment. *American political Science review*, 102(1):33–48, 2008.

S. González-Bailón, D. Lazer, P. Barberá, M. Zhang, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Freelon, M. Gentzkow, A. M. Guess, et al. Asymmetric ideological segregation in exposure to political news on facebook. *Science*, 381(6656):392–398, 2023.

A. Guess and A. Coppock. Does counter-attitudinal information cause backlash? results from three large survey experiments. *British Journal of Political Science*, 50(4):1497–1515, 2020.

A. M. Guess, N. Malhotra, J. Pan, P. Barberá, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Dimmery, D. Freelon, M. Gentzkow, et al. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, 381(6656):398–404, 2023.

Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE international conference on data mining*, pages 263–272. IEEE, 2008.

R. Levy. Social media, news consumption, and polarization: Evidence from a field experiment. *American economic review*, 111(3):831–70, 2021.

A. Lindbeck. Incentives and social norms in household behavior. *The American Economic Review*, 87(2):370–377, 1997.

S. Messing and S. J. Westwood. Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication research*, 41(8):1042–1063, 2014.

S. Mullainathan and A. Shleifer. The market for news. *American economic review*, 95(4): 1031–1053, 2005.

K. Muralidharan, M. Romero, and K. Wüthrich. Factorial designs, model selection, and (incorrect) inference in randomized experiments. *The Review of Economics and Statistics*, pages 1–44, 2023.

R. E. Robertson, S. Jiang, K. Joseph, L. Friedland, D. Lazer, and C. Wilson. Auditing partisan audience bias within google search. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–22, 2018.

J. E. Settle. *Frenemies: How social media polarizes America*. Cambridge University Press, 2018.

W. Suen. The self-perpetuation of biased beliefs. *The Economic Journal*, 114(495):377–396, 2004.

C. R. Sunstein. *# Republic: Divided democracy in the age of social media*. Princeton University Press, 2018.

A Supplementary Material

A.1 Recruitment and Data Collection

We used X ads to recruit adult English speakers located in the U.S. between March and June 2023. The ads are shown in Figure A-19. Each ad featured a map indicating the winning party of each county in the 2020 U.S. presidential election, along with one of four different recruitment texts.

A total of 951,470 unique users saw the ad, of whom 12,940 clicked on it. This 1.36% click-through rate is slightly higher than the average rate for X ads across all industries (0.86%, see `https://www.brafton.com/blog/social-media/social-advertising-benchmarks/`) and comparable to other studies in the literature reporting similar metrics on Facebook (Allcott et al. [2020] and Allcott et al. [2022] report a click-through rate of 1.7% and 0.8%, respectively). The average acquisition cost per link clicked was $0.14. The click-through rate and acquisition cost per link were similar across all ads, ranging from 1.26% to 1.47% and from $0.13 to $0.19, respectively.

A.2 Constructing News Outlets Suggestions

Each participant receives six suggested news outlets and access to a button that allows them to explore the full set of outlets in our database for additional options. We randomize whether participants receive the direct suggestions based on one of two algorithms. The algorithms select three outlets each from the list of national news outlets that have a slant to the left and right of the slant of the participant's news diet. The first presents the most *popular* outlets that the user does not already follow. The second system provides personalized suggestions generated by a collaborative-filtering-based recommender system using five latent factors. The algorithm we use is described in full in Hu et al. [2008].

The three panels in Figure A-20 plot (i) the probability of making at least one change to their news diet (*Any Change*), (ii) the absolute change in the total number of outlets followed ($|\Delta$Followed Outlets$|$), and (iii) the absolute change in the average slant of the outlets followed relative to those followed before the randomization ($|\Delta$Slant$|$) conditional on the content-selection algorithm assigned to the participant. The treatment effects are very consistent across the two algorithms, indicating that the method used to construct the suggested outlets is not driving the results.

A.3 Construction of News Outlets Dataset

The construction of the database proceeds in several steps:

1. **Potential set of news outlets.** We create a unified list of potential news outlets by harmonizing the active domains compiled by Athey et al. [2021] and Braghieri et al. [2024]. We focus on the data scraped in the first three months of 2021, which includes 4,412 outlets.

2. **Filtering by activity.** We exclude 1,363 outlets that published fewer than ten articles per month on average during this period. This ensures that we retain only outlets actively publishing articles with a sufficient online presence, resulting in 3,049 remaining outlets.

3. **Mapping outlets to X handles.** We attempt to link each outlet's website domain to its corresponding X account as follows:

   (a) For each outlet, we scrape its landing page and extract any X handle linked directly from it (e.g., through a header or footer social media icon).

   (b) If no valid handle is found on the landing page, we query the X API using the outlet's domain as a keyword and select the account with the largest number of followers among the top 20 results.

   (c) We manually verify all selected handles.

   This process yields a valid X handle for 2,651 outlets (87%).

4. **Deduplication by domain.** A small number of outlets share the same canonical domain (e.g., local editions of the same website). In these cases, we keep only the outlet with the largest number of published articles, resulting in 2,636 unique domains.

5. **Assigning political slant scores.** We assign each outlet a *slant score* using Robertson et al. [2018]. This measure captures the extent to which a domain is preferred by voters of one party relative to the other. It ranges from -1 to 1, taking a value of 0 when a domain is shared equally by Republicans and Democrats, and approaching 1 (-1) when it is shared exclusively by Republicans (Democrats). The original measure is constructed by:

(a) Matching voter registration records for roughly half a million U.S. voters to their linked X accounts;

(b) Collecting over 100 million posts from these accounts over time;

(c) Identifying posts that contain URLs; and

(d) Calculating, for each domain, the relative frequency with which Democrats versus Republicans share links to it. This produces a continuous slant score where positive values indicate a Republican-leaning audience and negative values a Democratic-leaning audience.

We are able to match slant scores for 1,918 domains (73%).

6. **Removing duplicate handles.** Some domains share the same X handle (e.g., affiliated or subsidiary publications). For these, we retain the domain with the largest number of articles, reducing the dataset to 1,908 outlets.

7. **Filtering for political ("hard") news.** To focus on politically relevant content, we estimate the share of "hard" or political news articles as opposed to "soft" content (e.g., sports, entertainment, lifestyle) for each outlet. This comes from an AutoML classifier predicting hard news labels at the article level using the article text [Ahmed et al., 2019]. This algorithm classifies each article as hard or soft news and we aggregate this to the outlet level by calculating the share of a outlet's articles that are hard news. We manually verify the algorithm correctly classifies 84% of articles in a randomly selected subset of 200 articles. We exclude 292 outlets at the lowest end of the distribution that can be clearly classified as publishing primarily soft-news, leaving 1,616 outlets.

8. **Filtering by top-level domain.** We further exclude 239 outlets that do not have a ".com" or ".net" top-level domain, ensuring that the remaining outlets are primarily U.S.-based and use standard commercial or network domains, yielding 1,377 outlets.

9. **Manual verification and exclusion of remaining soft or non-U.S. outlets.** We manually inspect all remaining outlets and remove 207 that:

   - Are primarily non-U.S.-based (e.g., international editions of foreign outlets);
   - Are not genuine news outlets (e.g., corporate or platform websites such as `facebook.com` or `google.com`); or

- Predominantly publish soft news that was not removed in Step 6.

After this final cleaning, 1,170 outlets remain in the final sample.

10. **Classification of outlet size.** We manually classify outlets by size and prominence, flagging *large news outlets* that are national in scope thus excluding very local or small-scale outlets (e.g., blogs). Out of the 1,170 outlets, 88 are classified as large. While the full set of 1,170 outlets is used to generate infographics for participants in our experiment, only the subset of large outlets is used to produce the six suggested outlets that each participant receives during the experiment.

The resulting dataset includes 1,170 U.S.-based hard-news outlets, with information on each outlet's canonical domain, verified X handle, political slant, an indicator of outlet size, and additional characteristics (including measures of total and hard-news article production, credibility and engagement measures in social media such as number of followers).

A.4 Interaction between Treatments

In the main text, we report the results of a regression that estimates effects separately for the peer information and disclosure conditions following equation (1). The coefficient of interest, β, represents the ITT estimate of receiving peer information (an incentive to disclose the news diet summary) across both types of participants: those in the disclosure (peer information) treatment and those in the disclosure (peer information) control.

Figure A-21 reports the main estimates of a fully saturated regression in which we simultaneously estimate the outcomes among individuals under the four possible treatment combinations: (1) no peer information and no incentive to disclose the news diet summary, (2) peer information but no incentive to disclose, (3) no peer information but an incentive to disclose, and (4) peer information *and* an incentive to disclose. Panel A reports estimates of the probability a participant updates their beliefs about the slant of their peers' news diet by treatment. We find that, compared to participants who do not receive peer information or an incentive to disclose (those in group 1), only the groups that receive peer information (groups 2 and 4) differentially update their beliefs. Both of these effects are statistically significant and of a similar magnitude: moving from group 1 to group 2 (4) leads to a 12.7 (13.0) percentage points increase in the probability of a participant updating their posterior beliefs about their peers. Instead, we find that receiving an incentive to disclose has a quantitatively small and statistically insignificant effect on beliefs across both types

of participants: those who do not receive the peer information treatment (group 3 vs. group 1) and those who do (group 4 vs. group 2). Panel B reports the effects on the probability of making changes to the news diet summary (any change). We find that, compared to participants who do not receive peer information or an incentive to disclose (those in group 1), only the groups that receive an incentive to disclose their news diet summary (groups 3 and 4) are more likely to make changes: moving from group 1 to group 3 (4) leads to a 6.2 (4.3) percentage points increase in the probability of making changes to the news diet. Both of these effects are marginally significant (p-values are 0.086 and 0.124, respectively). We find that the peer condition instead reduces the probability of making changes to the news diet, although these effects are insignificant and smaller in magnitude.

Overall, the interaction term between the two treatment conditions (peer information and disclosure) is quantitively small and statistically insignificant, which explains why the estimates using this approach are very similar to those in the main text.[32]

A.5 Instrumental Variable Estimates

In the main text, we report causal effects based on intention-to-treat (ITT) estimates derived from reduced-form regressions, which capture the impact of assignment to different informational treatments. An alternative way to compare the quantitative effects of our experiments is to estimate local average treatment effects (LATE) using two-stage least squares (2SLS), where the endogenous regressor is whether participants disclosed their news diets (in the disclosure condition) or reported believed changes in the ideological slant of their peers' news choices (in the peer information condition).[33]

Table A-16 reports these 2SLS estimates for our main results. Panel A focuses on the disclosure condition, with the three columns presenting effects on our three key outcomes measuring changes in the news outlets followed. For instance, the 2SLS coefficient reported in Column (1) implies that full compliance (moving from not posting to posting the news

[32]The magnitude associated with the interaction term is -0.012 (standard error of 0.047) in Panel A and 0.014 (standard error of 0.040) in Panel B.

[33]We pre-registered the ITT estimates as our baseline in part because we could not guarantee the absence of exclusion restriction violations in the 2SLS specification. In the disclosure condition, for instance, participants may have modified their news diets with the intention of sharing them, but after observing the resulting summary, decided not to follow through. The data provide some suggestive evidence of this behavior: although only marginally, participants who did not share their news diet summary despite being incentivized to do so were slightly more likely to change their news diets than those in the control group. Our ITT specification instead captures the overall change induced by the incentive, reflecting the effect on both participants who ultimately shared and those who intended to share but did not follow through.

diet summary) would increase the probability of making any change to news outlets by 89 percentage points.

Panel B focuses on the peer information condition. The two columns report effects on observed changes in participants' own news diets, with Column (1) and Column (2) showing effects on the intensive and extensive margins, respectively. The 2SLS estimates are small in magnitude and statistically insignificant, consistent with our earlier findings.

A.6 Sample Selection and External Validity

Here we compare our sample to the U.S. adult population. Our sample contains a larger share of white, highly educated, older, and male individuals than the U.S. adult population (Panel A of Table A-2). In addition, the majority of users in our sample report following politics "very closely."

To assess the generalizability of our results, first, we estimate our main results using sample weights constructed to ensure that the analysis reflects the demographic composition of the U.S. adult population relying on the observables demographic characteristics reported in Table A-2.

Second, our sample overrepresents politically engaged U.S. adults. This likely reflects both the higher political engagement of X users relative to the all U.S. adults and the possibility that our ads disproportionately attracted individuals with political interests. 60% of participants report following politics "very closely," 37% "somewhat closely," and 3% "not very closely." In comparison, World Values Survey (WVS) data from 2018 indicates that only 21% of U.S. adults were "very interested" in politics, 43% were "somewhat interested," and 35% reported being "not very interested" or "not interested at all."[34] In light of this evidence, we explore the robustness of our main results by focusing on the 40% of participants with low political interest (those who follow politics somewhat or not very closely), who are underrepresented in our sample relative to the U.S. adult population. For comparison, we also include the results for participants who follow politics very closely.

Figure A-22 reports the robustness of the social image concern results, replicating Figure 2, after re-weighting the sample and conditional on how closely the participant follows politics. Table A-15 reports the robustness of the peer information results, focusing on

[34]Interestingly, the demographic composition of politically engaged individuals in the WVS mirrors the differences between our sample and the U.S. adult population. Male, white, college-educated, and above-median-age individuals are, on average, 14, 6, 16, and 23 percentage points more likely to be interested in politics than female, non-white, non-graduate, and below-median-age individuals, respectively.

Panel A of Tables 1 and 2.

We find that our main results remain broadly robust across these exercises.[35] Specifically, both the social image concern and peer information effects are qualitatively unchanged when we reweight the sample to better approximate the U.S. adult population and when we focus on participants with lower political interest. While these exercises cannot fully rule out the influence of unobservable differences, they help shed light on the representativeness of our findings relative to the U.S. adult population and provide additional confidence in the stability of our main results. Regardless, we view our final sample as substantively interesting in its own right, as it captures effects among politically interested participants who are more likely to engage with political news regularly, either directly or through their peers.

A.7 Construction of Random Sample of X Users

We create a random sample of X users following Bruner [2013]. X assigns each account a unique integer ID. Until 2015, these IDs were approximately sequential, which made it feasible to obtain a random sample of X accounts by randomly drawing IDs from the known range of possible values. After 2015, however, X changed its ID assignment algorithm, which made the ID space substantially more sparse and rendered this approach computationally infeasible. We rely on the dataset provided by Bruner [2013], which contains a random sample of 400,000 X accounts collected in 2013 following the aforementioned approach. From this dataset, we draw a random subsample of these accounts—limited by the constraints of the X API—for which we collect a set of recent metrics (e.g., outlets followed, engagement). This updated information serves as a reference point for comparison with our sample. Our final dataset therefore consists of 5,097 randomly selected X accounts originally sampled in 2013, with account-level information updated to 2022.

[35] We note some small quantitative differences in our results, but the overall pattern remains consistent. For instance, the effect of the disclosure treatment on the probability of making any change is slightly smaller when using the weighted specification, yet it remains statistically significant at the 5% level. Participants who follow politics very closely tend to make slightly more changes to the news outlets they follow in both the treatment and control groups; however, the heterogeneous treatment effect is not statistically significant.

A.8 Alternative Normalizations of the Prior and Peer Signal

We elicit participants' prior and posterior beliefs about the ideological slant their own news diets and the news diets of their peers using a seven-point scale ranging from "Extremely liberal" to "Extremely conservative." The options are: "Extremely liberal," "Slightly liberal," "Moderate," "Slightly conservative," "Conservative," and "Extremely conservative."

We measure the actual ideological slant of participants' news diets and their peers' news diets as the average ideological score of the outlets each individual follows. As explained in Section 2.3, outlet scores are coded on a scale from -1 to 1. Because the peer signal is constructed as an average of these outlet scores, it is by construction bounded within the same interval ($[-1, 1]$).

As explained in Footnote 20, to calculate our main results with the signal-oriented outcomes, we calculate the *surprise* as the difference between the peer signal and the peer prior. Because the prior and the signal are measured on different scales, we first normalize them to ensure comparability.

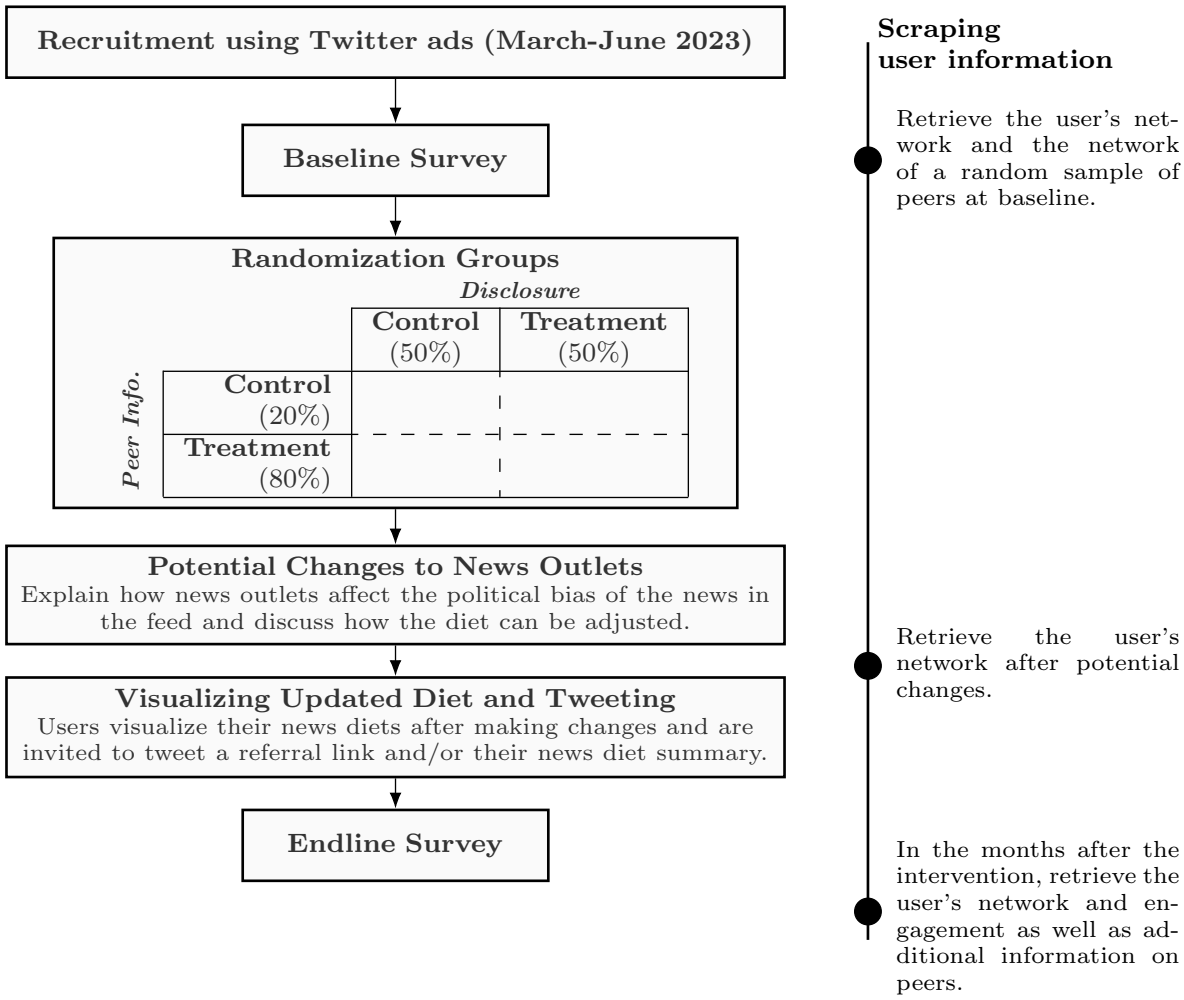
In what follows, we examine the robustness of our results to alternative normalization methods. The exact formulas used to calculate these normalizations are presented in Table A-13. In the main results, we follow Normalization 1, which transforms the signal by defining cutoffs that divide the interval $[-1, 1]$ into seven equally sized intervals and assigning to each observation the value of the interval in which the signal falls. This places both the signal and the prior on a common seven-point scale. Normalization 2 uses the means of the signal within each of the seven prior categories as anchors to set the cutoffs, while Normalization 3 is similar but uses medians instead of means. Normalization 4 applies a uniform transformation that maps the signal from $[-1, 1]$ to $[1, 7]$. Finally, Normalization 5 standardizes both signal and prior by subtracting their means and dividing by their standard deviations.

Table A-14 presents treatment effects of the peer treatment on participants' beliefs about the news diets of their peers in Columns (1)-(4) and on changes in participant news diets in Columns (5)-(8) using the alternative normalizations described above. These results are provided for comparability with Panel A in Tables 1 and 2, which follow Normalization 1.

Regardless of the method used to align the scales of the signal and the prior, we find that exposure to a peer signal consistently shifts participants' beliefs about the news diets of their peers. However, we find no evidence that these peer signals translate into real changes in the participants' own news diets. Taken together, these results reinforce our interpretation that peer effects in this context influence beliefs but not actual news consumption behavior.

A.9 Additional Figures and Tables

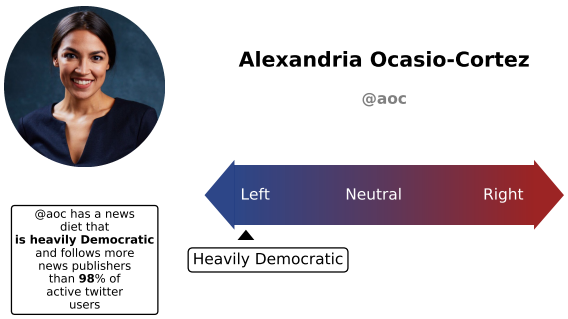
Figure A-1: Experimental Design



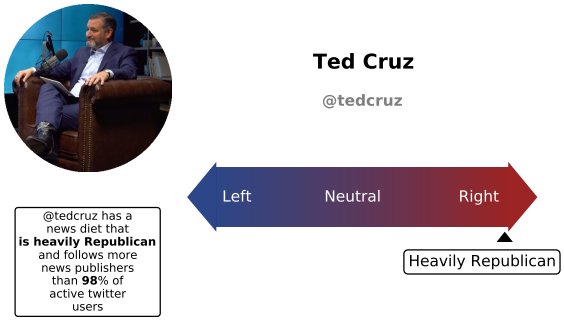
Notes: This figure illustrates the experimental design. The left side shows the process that participants follow to complete the experiment and the randomization assignments. The right side highlights the periods in which we scrape information about the participants, their peers (e.g. accounts followed), and their engagement. A total of 951,470 unique X users were shown the recruitment ads. 12,940 clicked on the ads, 5,190 consented to participate in the study and 3,755 finished the study.

Figure A-2: Examples of News Diet Summaries

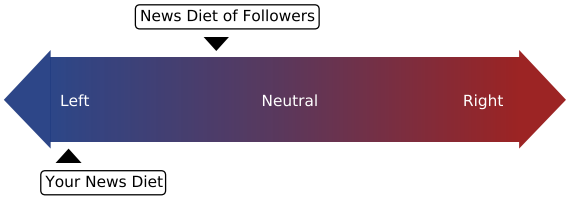
A. @AOC's News Diet Summary



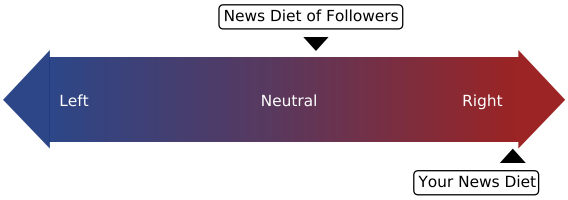
B. @TedCruz's News Diet Summary



C. @AOC Followers' News Diet Summary

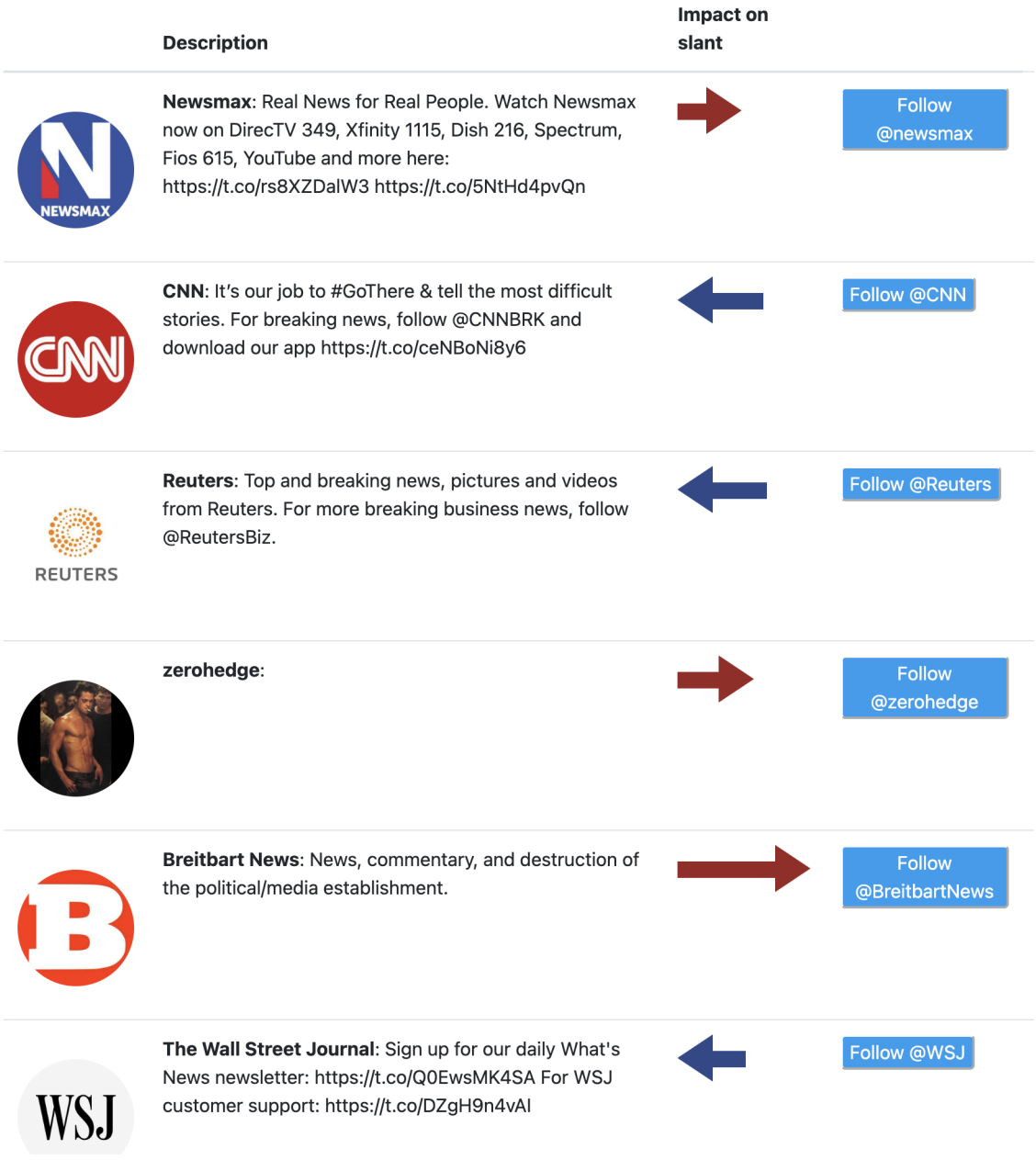


D. @TedCruz Followers' News Diet Summary



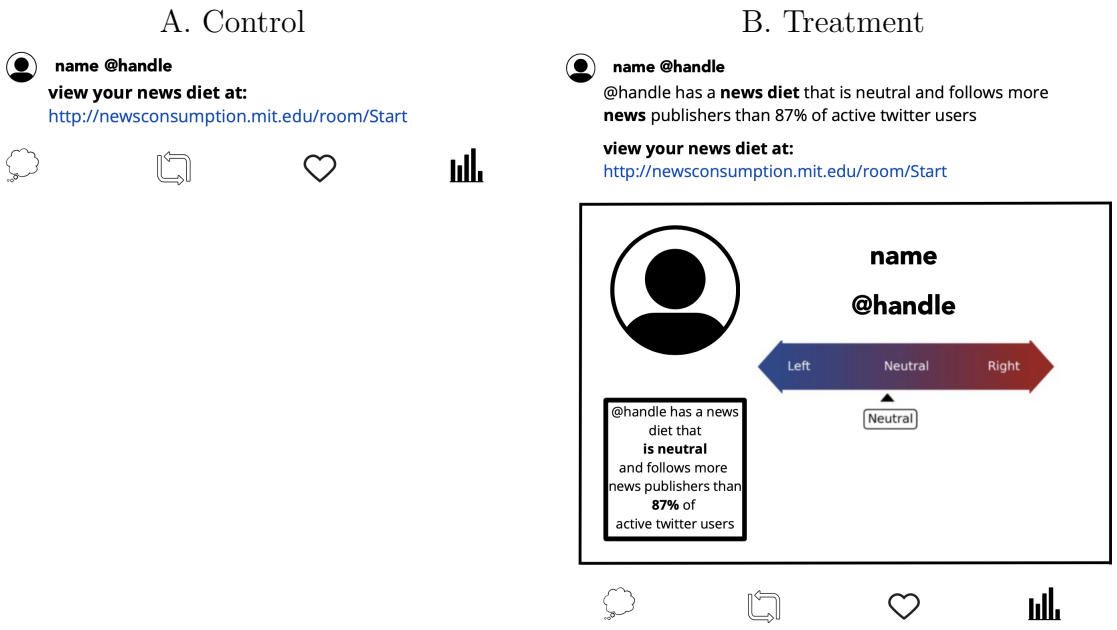
Notes: The figure illustrates the infographics used to convey information about participants' news diets using an example involving two well-known politicians. All study participants viewed an infographic summarizing their own news diet, similar to the one in Panels A or B, but populated with information from their own account. This infographic reported how much they engage with news outlets relative to the average X user, as well as the slant of those outlets. Participants assigned to the peer information treatment also viewed a second infographic showing the average slant of the news diets of their peers (followers), similar to the one in Panels C or D. All information displayed in these infographics is collected in real time using the X API.

Figure A-3: Example of News Outlet Suggestions



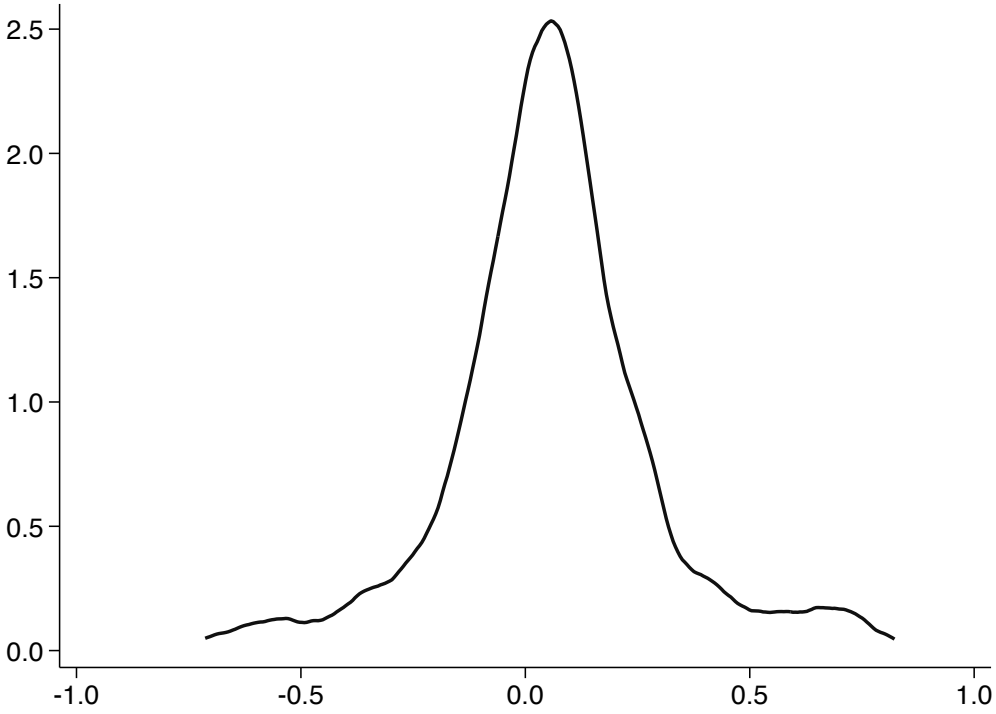
Notes: The figure illustrates the design of the six outlets suggested to the participants during the experiment. All participants viewed a list of six news outlets along with an arrow indicating the impact that following each outlet would have on the average slant of the news outlets they follow on X. The outlets were selected to be balanced, with three options that would shift the slant of a participant's news diet to the right and three that would shift it to the left.

Figure A-4: Example Posts for Compliance in the Disclosure Condition



Notes: This figure illustrates the posts that participants were incentivized to share in the disclosure conditions. These posts were automatically drafted by our system using each participant's information, but participants ultimately decided whether to post the message to their account.

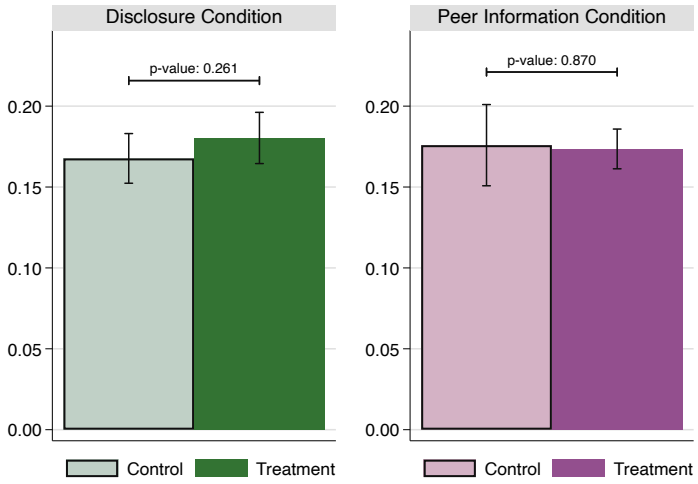
Figure A-5: Distribution of the Slant of the X News Outlets



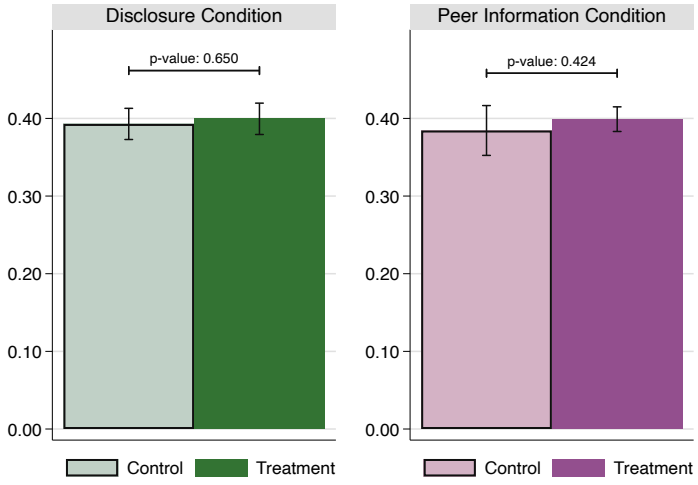
Notes: This figure plots the distribution of the slant for the final dataset of 1,170 news outlets on X.

Figure A-6: Differential Attrition Rate for Disclosure and Peer Information Conditions

A. Probability of Not Reaching Sharing Page (Main Sample)



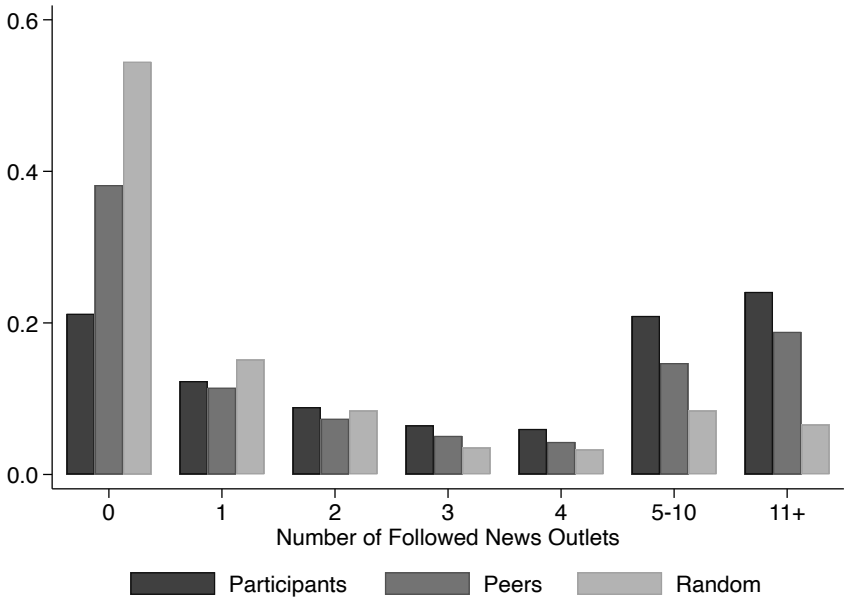
B. Probability of Not Completing Endline Survey



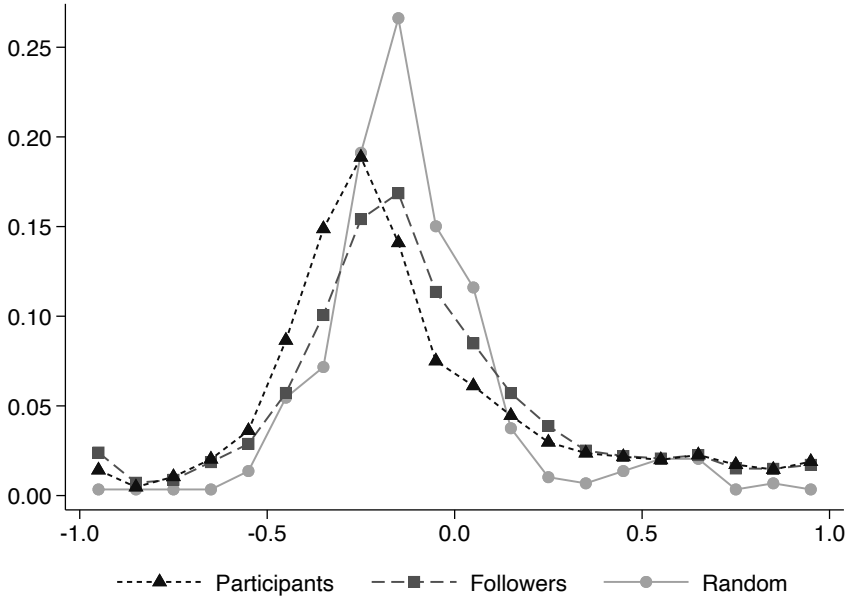
Notes: This figure reports the fraction of participants that did not complete the experiment across participants assigned to the treatment and control groups in the disclosure condition (left) and the peer information condition (right). Panel A plots the probability a participant does not reach the sharing page where the participants had the opportunity to share either the placebo message or the message containing the participant's news diet. Panel B plots the probability a participant does not complete the endline survey. The difference between the treatment and control groups in the disclosure condition is that only participants in the former group are asked to reveal to their peers which news outlets they follow. The difference between the treatment and control groups in the peer information condition is that only the former receives a signal about the slant of the news diets of their peers. The whiskers indicate the 95% confidence intervals. We also report the p-value for the null hypothesis of no difference between the treatment and control groups.

Figure A-7: Summary of Participant News Diets

A. Total Number of News Outlets Followed



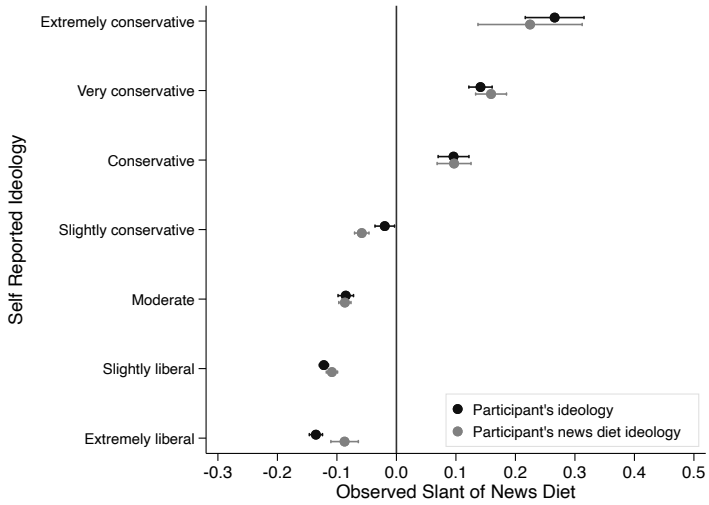
B. Slant of News Diets Followed Conditional on Following Outlets



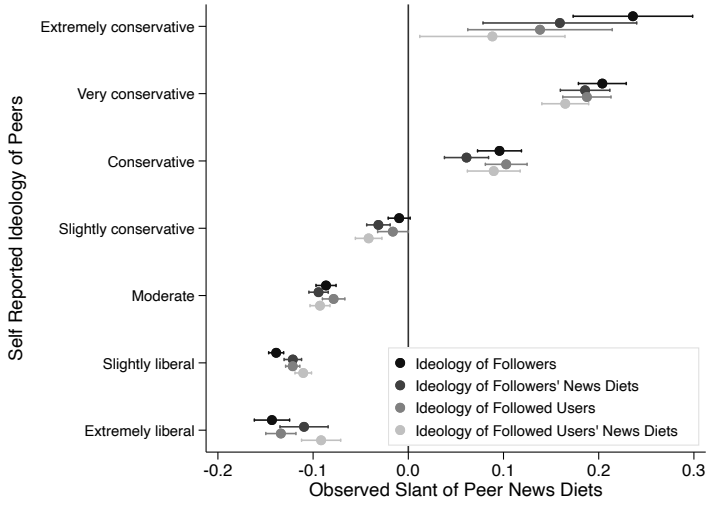
Notes: This figure presents the distribution of the total number of news outlets followed (Panel A) and the average slant of these outlets conditional on following at least one outlet (Panel B) for three samples: the participants, the participant's followers, and a random sample of active X users (see Appendix A.7).

Figure A-8: Self Reported and Observed Measures of Ideology

A. Own News Diet Conditional on Self-Reported Ideology

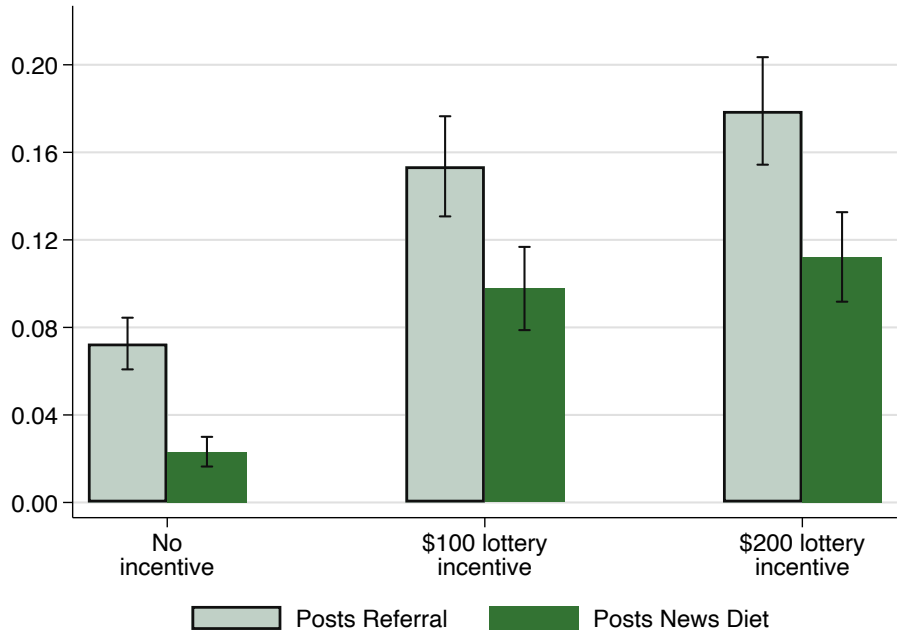


B. Peer News Diet Conditional on Beliefs About Peer Ideology



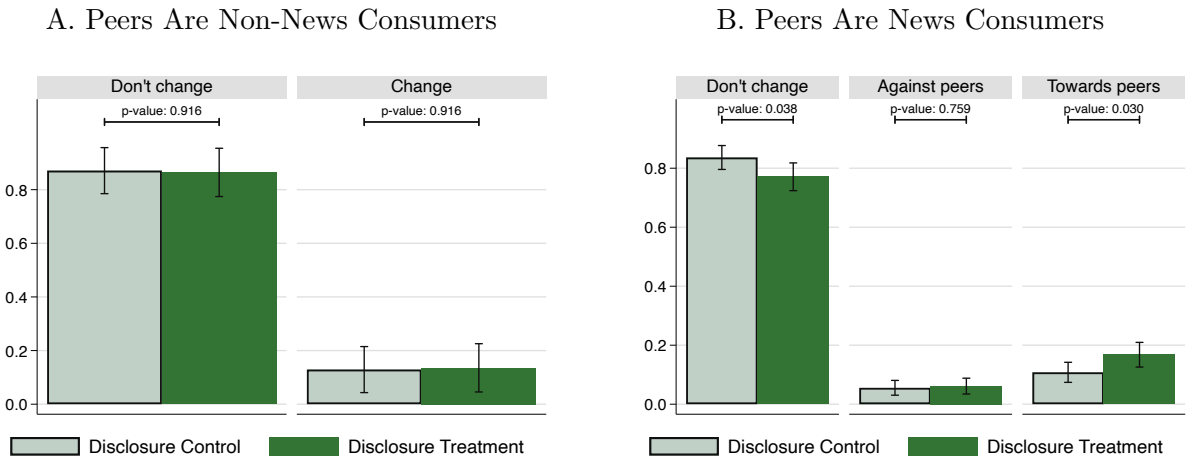
Notes: This figure illustrates the relationship between self-reported ideology elicited in the baseline questionnaire and the observed slant of the news diets measured at baseline. Panel A focuses on these measures for the participants, while Panel B focuses on the participants' peers—contrasting participants' beliefs about their peers (followers or followed users) with the actual slant of those peers' news diets. Each circle in Panel A plots the average slant of news diets of users conditional on self-reported ideology either about themselves or their news diets. Each circle in Panel B plots the average slant of the news diets of participants' peers conditional on a participant's beliefs about the ideology of their peers. The whiskers indicate the 95% confidence intervals.

Figure A-9: Differential Compliance Rates in the Disclosure Condition by Incentive Amount



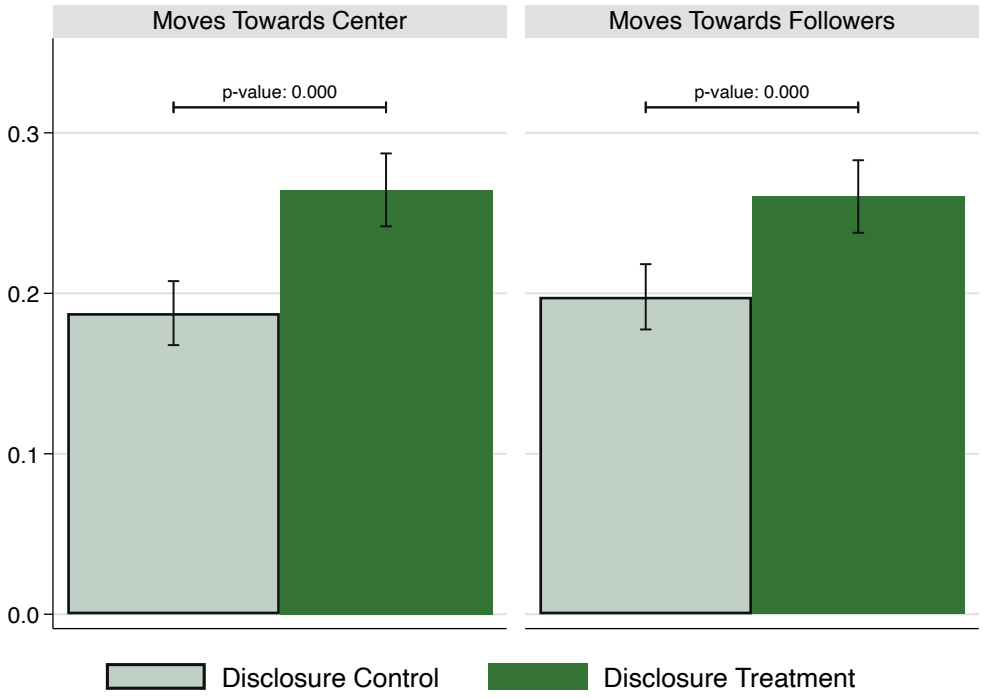
Notes: This figure reports the fraction of participants that complied with the disclosure treatment incentive conditional on the incentive they received. Each bar captures the share of participants who posted either the placebo message with the referral link (light bars) or the message containing their news diet summary (dark bars) conditional on the incentive they received. The 'No incentive' bars correspond to individuals who received an incentive to post the other message: the light bar is the share of participants who posted the placebo message among those in the disclosure treatment group and dark bar is the share of participants in the disclosure control group posting the message containing their news diet. The $100 and $200 lottery incentive bars plot the share individuals who post the placebo message (message with news diet) among participants in the disclosure control (treatment) groups. The dollar amounts corresponds to the bonus received if a participant complied with the treatment and won the lottery. The whiskers indicate the 95% confidence intervals.

Figure A-10: Directional Impact of Disclosure Condition – Non-News Consumers



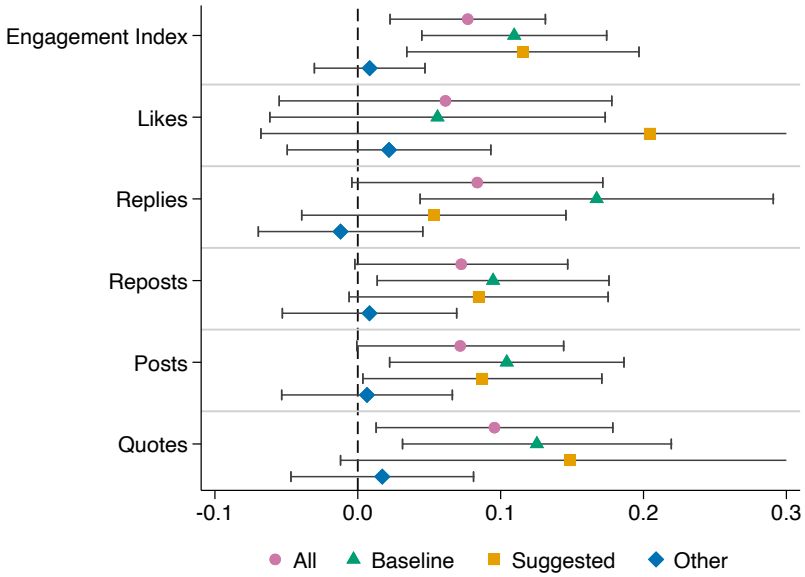
Notes: This figure reports how non-news consumers (participants who do not follow any news outlet at baseline) adjust the slant of their news diets in the disclosure condition. Panel A reports the results for the subgroup of participants whose peers are also non-news-consumers, while Panel B displays the results for participants whose peers are news-consumers. Panel A plots the share of participants who make any change to their news diet conditional on treatment status. Panel B plots the share of participants who make a change in the news diet towards the news diets of their peers or against the news diets of their peers by treatment status. The difference between the treatment and control groups in the disclosure condition is that only participants in the former group are asked to reveal to their peers which news outlets they follow. The whiskers indicate the 95% confidence intervals. We also report the p-value for the null hypothesis of no difference between the treatment and control groups.

Figure A-11: Directional Impact of Disclosure Condition — Center vs. Peers



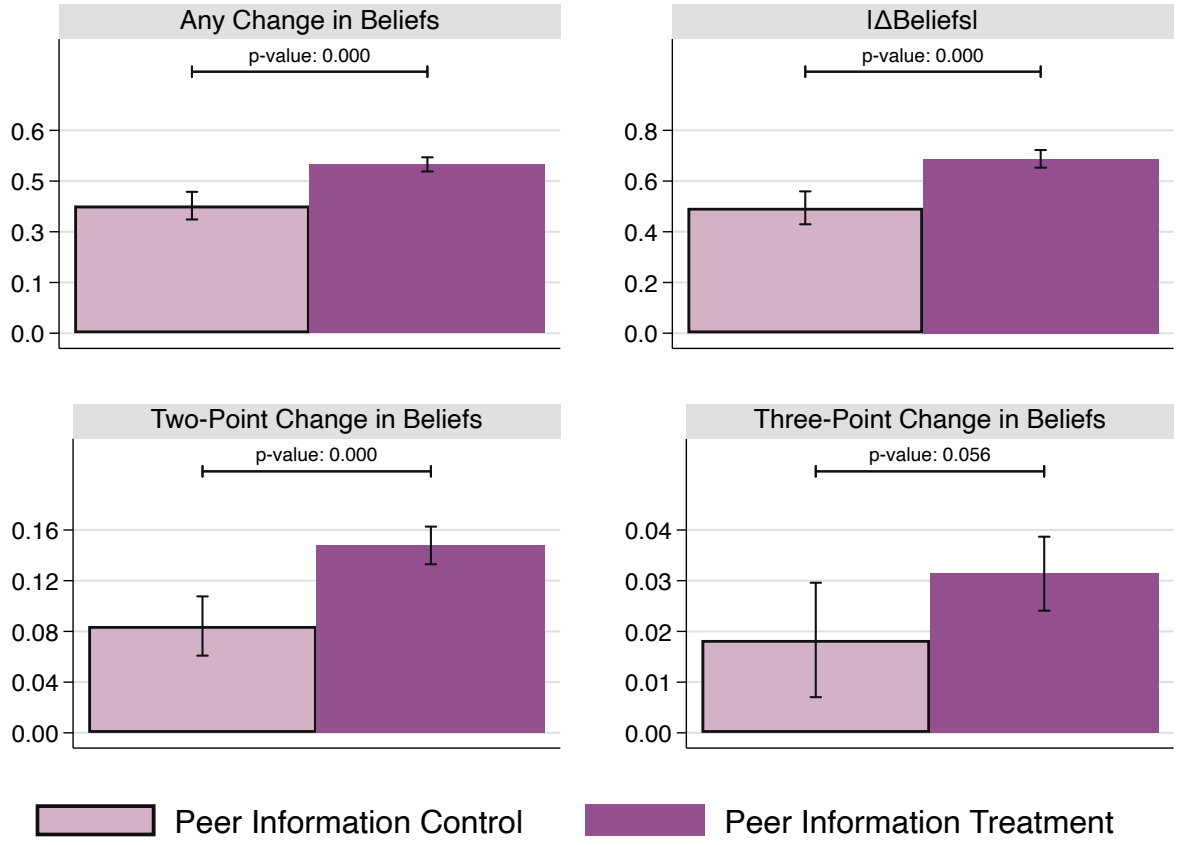
Notes: This figure reports the fraction of news-consumers (participants who do follow any news outlet at baseline) that adjust the slant of their news diets toward the center (left) vs. toward their peers (right) across the treatment and control groups in the disclosure condition. The difference between the treatment and control groups in the disclosure condition is that only participants in the former group are asked to reveal to their peers which news outlets they follow. The whiskers indicate the 95% confidence intervals. We also report the p-value for the null hypothesis of no difference between the treatment and control groups.

Figure A-12: Effect of Disclosure Condition on Engagement with News Outlets



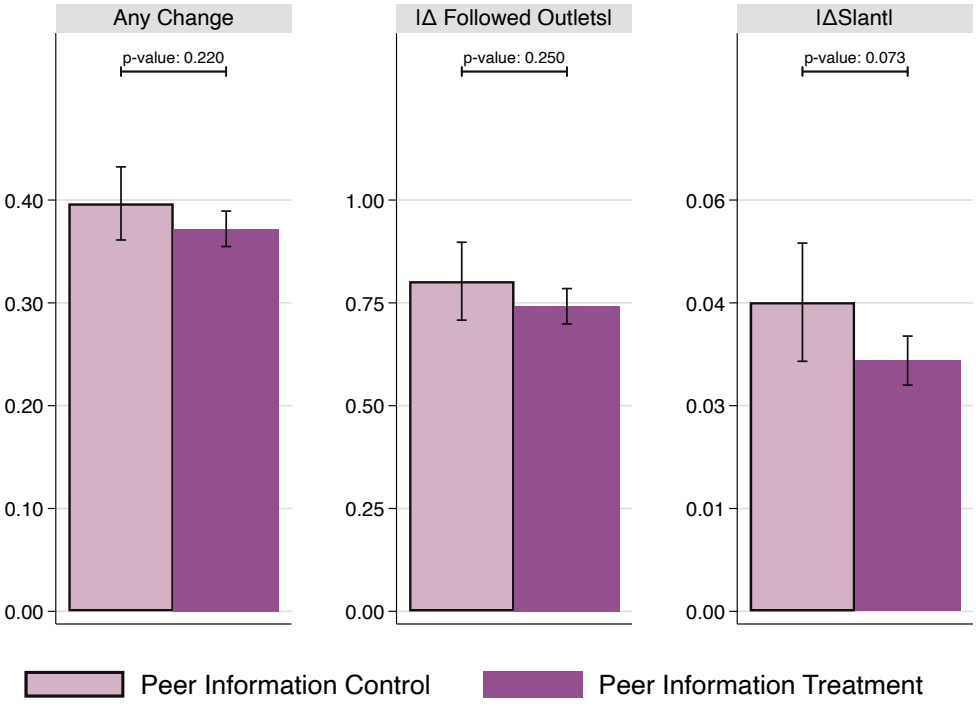
Notes: This figure reports estimates of the treatment effect of the disclosure condition on the index of engagement, the number of likes, replies, reposts, posts and quotes associated with the following subsets of news outlets: Any news outlet (All), the set of news outlets followed at baseline (Baseline), the six suggested outlets (Suggested), and the set of news outlets neither initially followed nor suggested (Other). The outcomes are standardized, so the plotted coefficients represent treatment effect estimates in standard deviation units of the outcome in the control group. The definition of association with a news outlet is provided in footnote 18. The difference between the treatment and control groups in the disclosure condition is that only participants in the former group are asked to reveal to their peers which news outlets they follow. The whiskers indicate the 95% confidence intervals.

Figure A-13: Effect of Peer Information Treatment on Beliefs



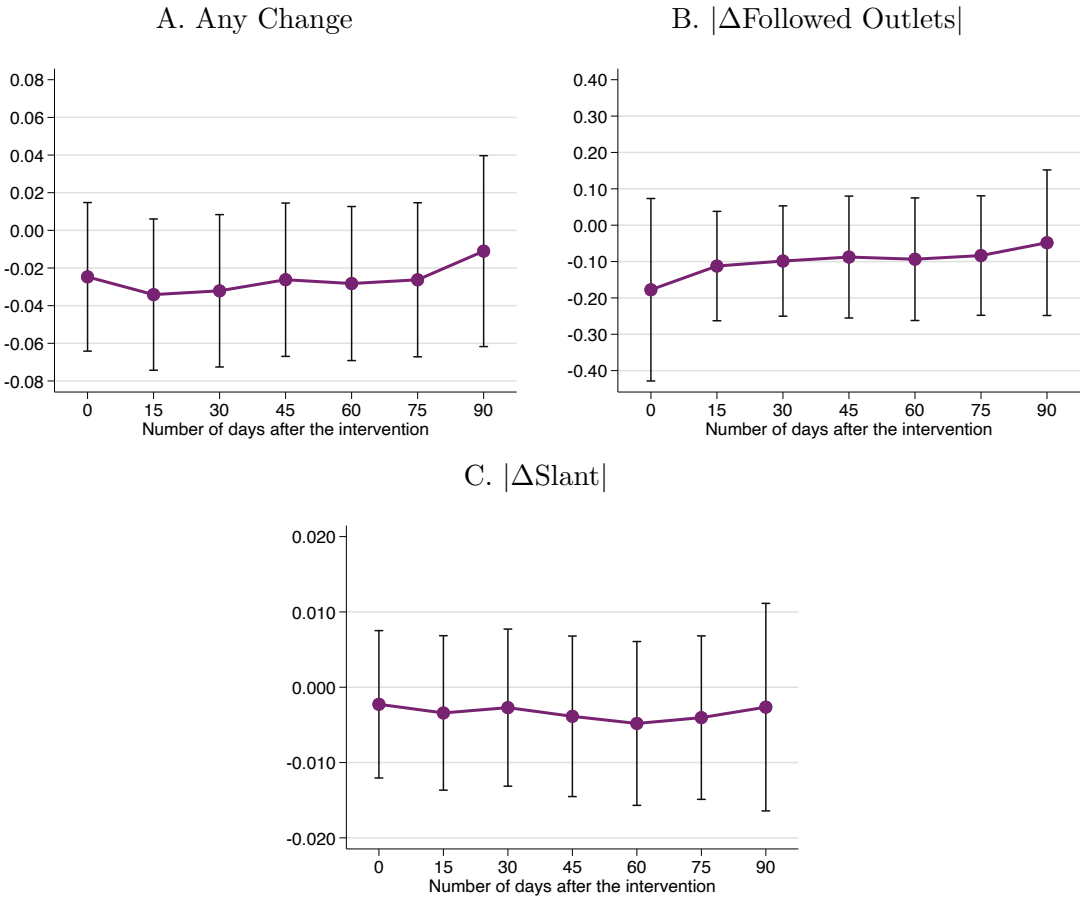
Notes: This figure reports the fraction of participants that update their beliefs about the average slant of their peers' news diets separately for participants in the peer information control and treatment. The top left panel plots the share of participants that update their beliefs about the slant of their peers' news diets at all. The top right panel plots the absolute change in the beliefs about the slant of their peers' news diets. The bottom left (right) panel plots the share of participants who update their beliefs about the slant of their peers' news diets by at least two (three) points on the seven point-scale in which beliefs are measured. In all cases, we compare these outcomes in the post-intervention period relative to the baseline. The difference between the treatment and control groups in the peer information condition is that only the former receives a signal about the slant of the news diets of their peers. The whiskers indicate the 95% confidence intervals. We also report the p-value for the null hypothesis of no difference between the treatment and control groups.

Figure A-14: Effect of Peer Information Condition on News Diets



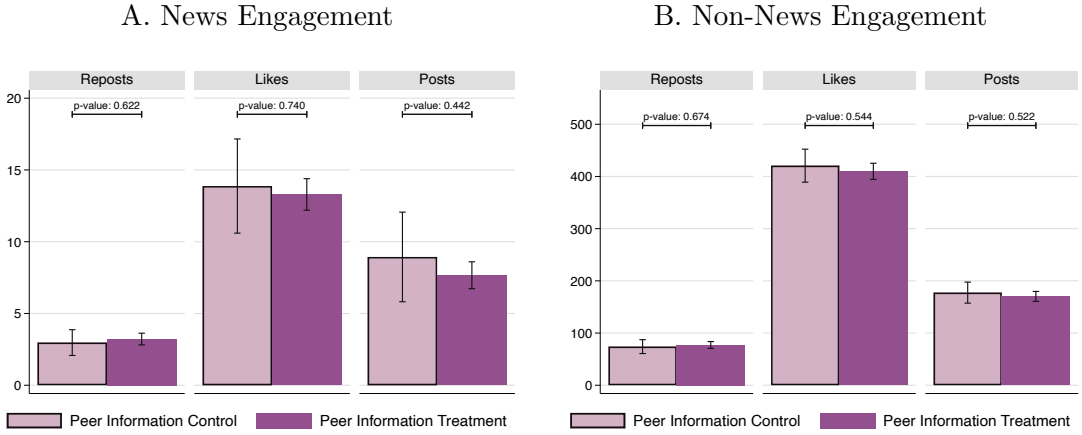
Notes: This figure reports the mean of three outcomes across participants assigned to the peer information treatment and control groups: an indicator variable if participants make any change to their news diets (left), the absolute change in the number of outlets followed (center), and the absolute change in the slant of the participants' news diets (right). In all cases, we compare these outcomes in the post-intervention period relative to the baseline. The difference between the treatment and control groups in the peer information condition is that only the former receives a signal about the slant of the news diets of their peers. The whiskers indicate the 95% confidence intervals. We also report the p-value for the null hypothesis of no difference between the treatment and control groups.

Figure A-15: Persistence of the Peer Information Treatment Effects



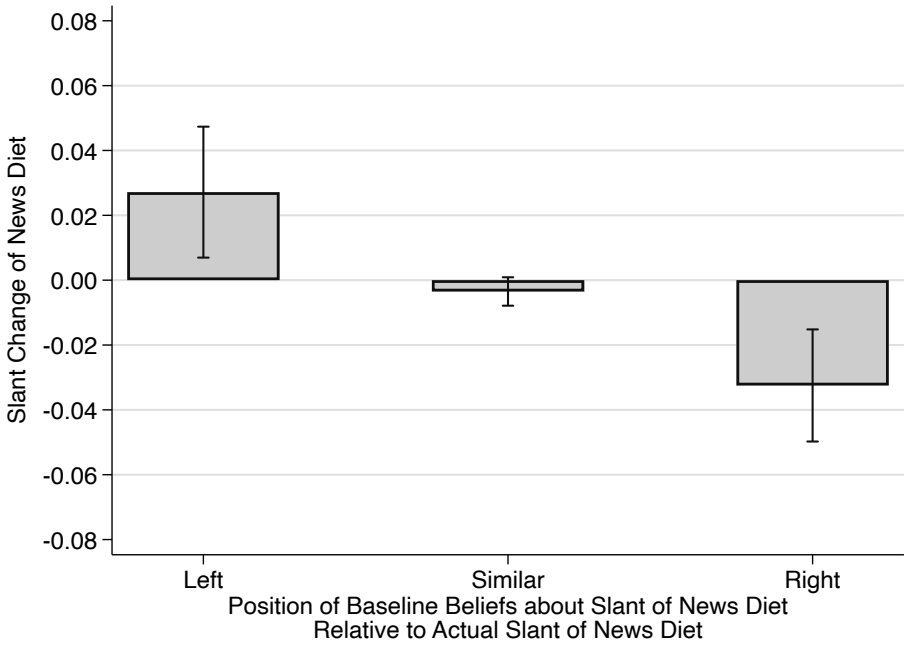
Notes: This figure reports treatment effects of the peer information condition for the 90 days after the intervention. Panels A, B, and C report effects on an indicator for whether the participant makes any change to the news outlets they follow, the absolute change in the number of outlets followed, and the change in the absolute slant of the outlets followed, respectively. In all cases, we construct these outcomes by comparing the set of outlets participants follow at the specified date (see x-axis) with those followed at baseline. The difference between the treatment and control groups in the peer information condition is that only the former receives a signal about the slant of the news diets of their peers. The whiskers indicate the 95% confidence intervals.

Figure A-16: Effect of Peer Information Condition on X Engagement



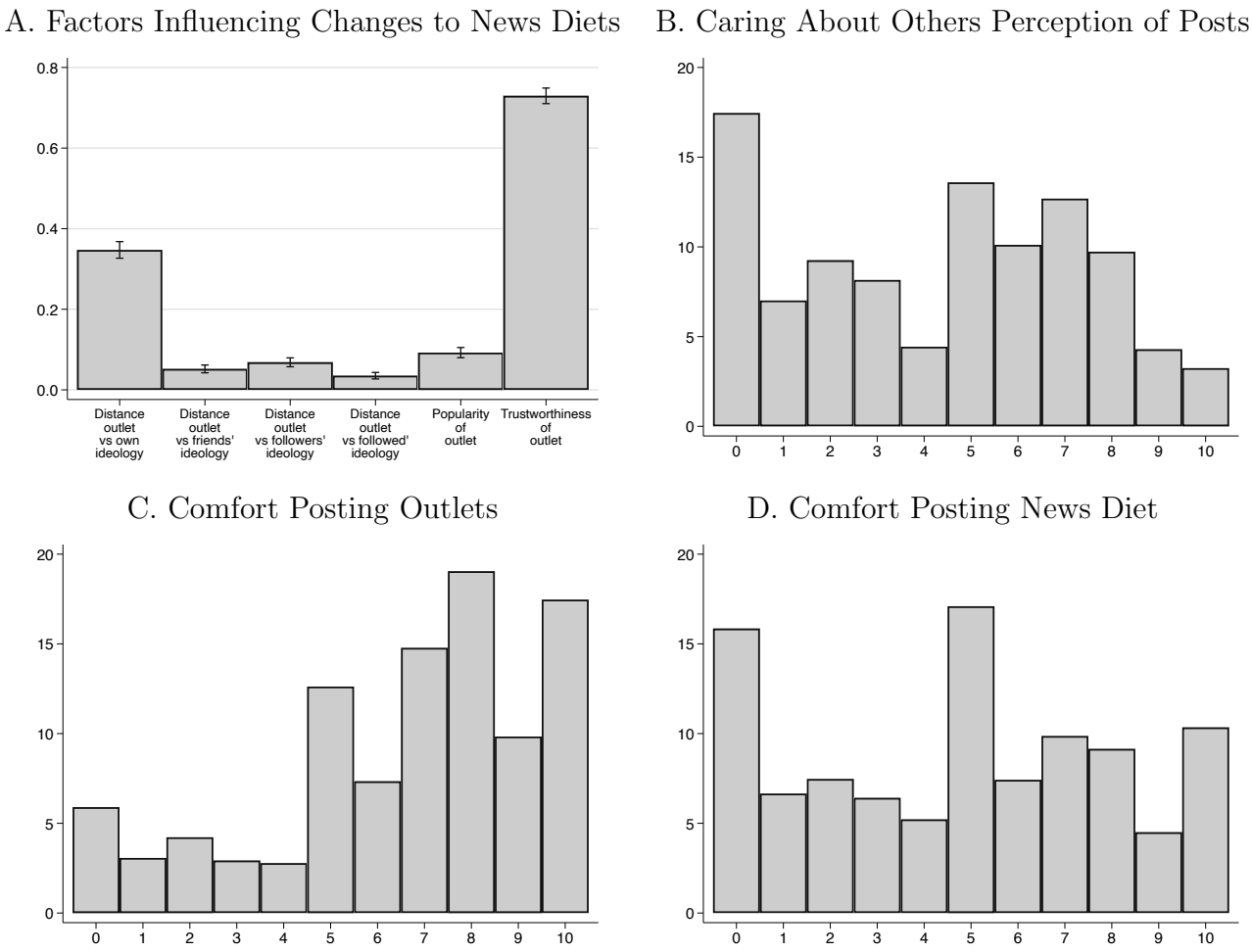
Notes: This figure displays the mean of the number of reposts, likes, and posts associated with news outlets (Panel A) and non-news outlets (Panel B) across participants assigned to the treatment and control groups in the peer information condition. Reposts, likes, and posts are defined as associated with news outlets if the post mentions a news outlet or responds to a post created by a news outlet. The difference between the treatment and control groups in the peer information condition is that only the former receives a signal about the slant of the news diets of their peers. The whiskers indicate the 95% confidence intervals. We also report the p-value for the null hypothesis of no difference between the treatment and control groups.

Figure A-17: News Diets Changes by Relative Position of Baseline News Diet and Beliefs



Notes: This figure plots, for participants in the disclosure control condition, the change in the slant of a participant's news diet throughout the experiment based on the relative location of their baseline news diet and their beliefs about the slant of their news diet. 'Left' ('Right') represents participants whose beliefs about the slant of their news diets are to the left (right) of their actual news diet at baseline. 'Similar' corresponds to participants with correct beliefs about the slant of their news diet. We rescale the slant of a participant's news diet at baseline following Footnote 20 to form the x-axis groups. The whiskers indicate the 95% confidence intervals.

Figure A-18: Distribution of Responses to Endline Questions



Notes: This figure plots the distribution of responses across four different endline survey questions. Panel A shows the proportion of respondents that answer each category in the x-axis in response to the question: Could you tell us if any of these factors play a role to influence your decision to follow or unfollow any news publishers during the study? (i) How close or far the publishers are from my ideology (ii) How close or far the publishers are from the ideology of my friends (iii) How close or far the publishers are from the ideology of my followers (iv) How close or far the publishers are from the ideology of the people that I follow (v) The popularity of the publishers (vi) The trustworthiness of the publishers. Panel B plots the distribution of responses to the question: On a scale from 0 to 10 where 0 "None" and 10 means "A lot", how much do you care about what others think of the content you tweet and retweet?. Panel C plots the distribution of responses to the question: On the same scale, how comfortable would you feel retweeting information about a publisher you follow?. Panel D plots the distribution fo responses to the question: On the same scale, how comfortable would you feel retweeting information about your news diet summary?

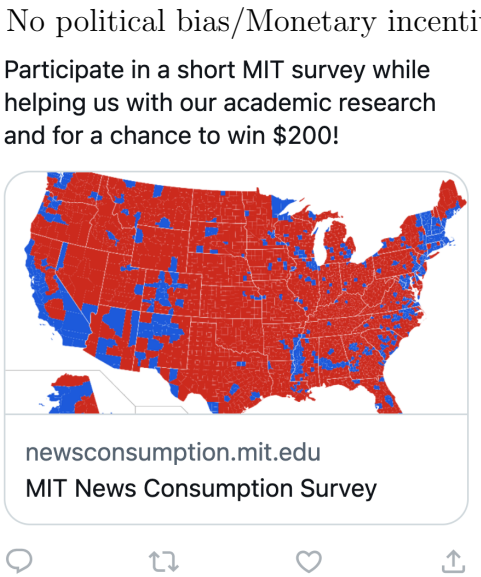
Figure A-19: Recruitment Ads

A. Political bias/No monetary incentive

B. Political bias/Monetary incentive

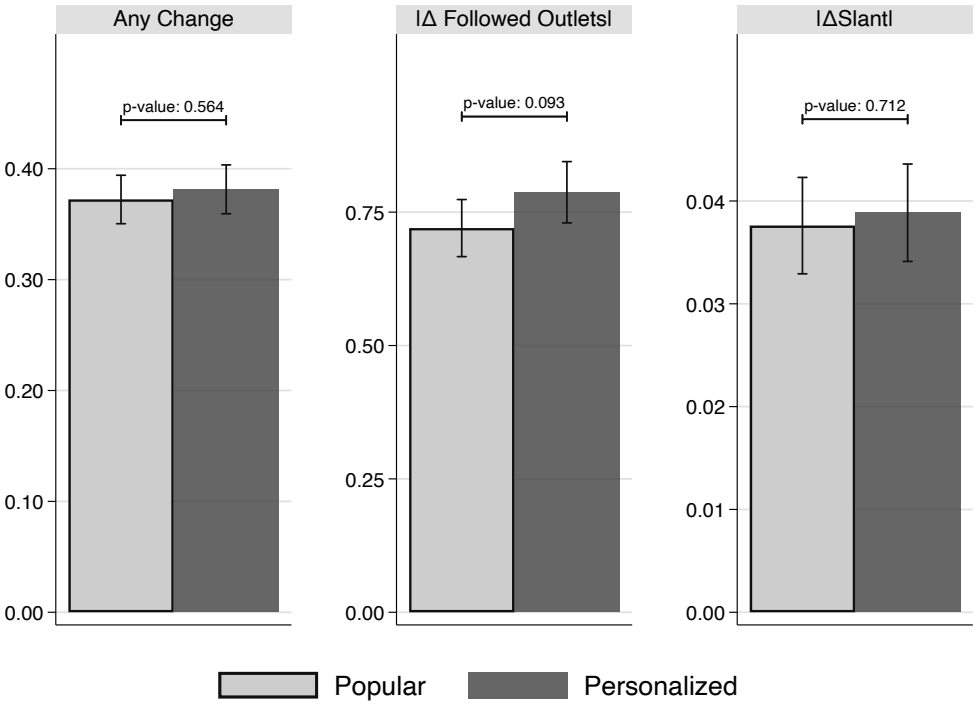
C. No political bias/No monetary incentive

D. No political bias/Monetary incentive



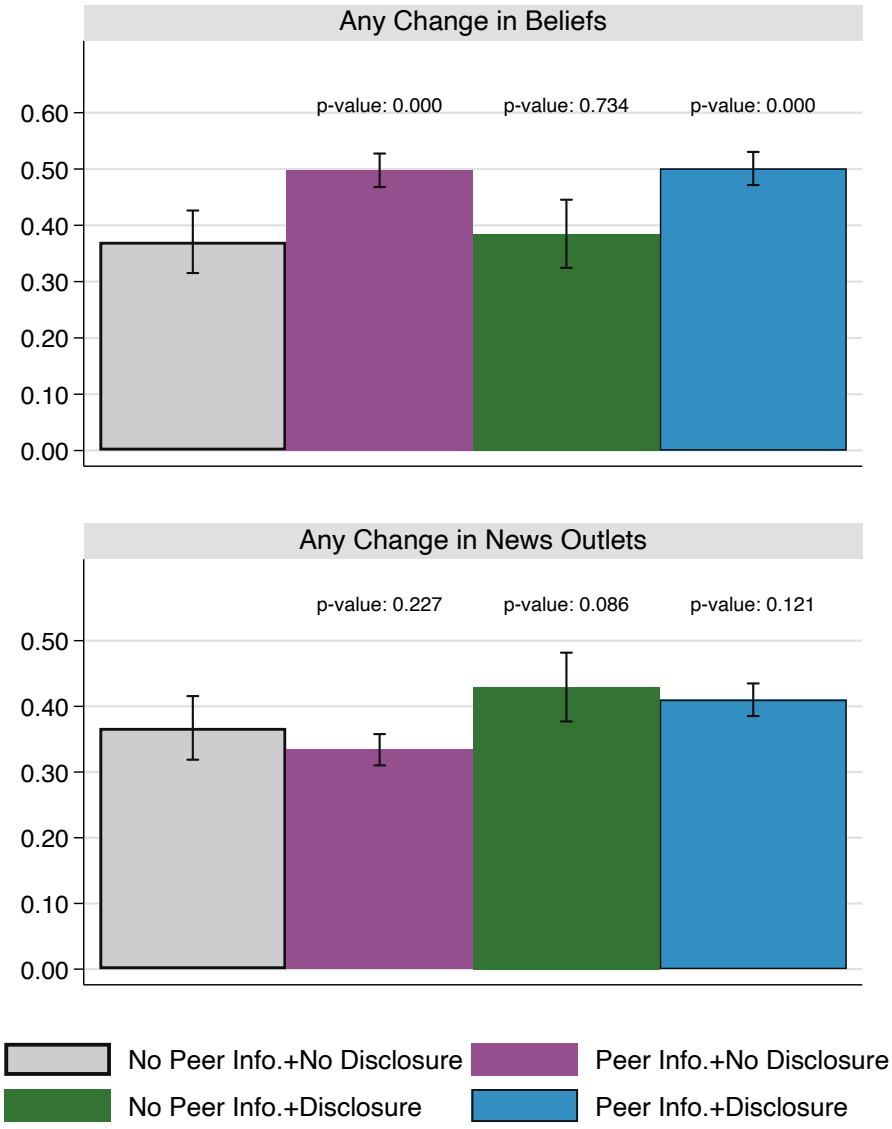
Notes: The figure shows the ads used for recruitment. Each map indicates whether the Republican Party (red) or Democratic Party (blue) received the majority of votes in the 2020 U.S. presidential election in each county, along with text that varies depending on whether a monetary incentive is mentioned and whether we highlight the opportunity for participants to assess their own political bias.

Figure A-20: Effect of Popular vs Personalized Suggestions on News Diets



Notes: This figure reports the mean of the outcome conditional on whether the participant received suggestions from the popular or personalized algorithm described in Section A.2. The outcomes are an indicator variable if the participant makes any change to their news diet (left), the absolute change in the number of news outlets followed (center), and the change in the absolute slant of their news diet (right). In all cases, we compare these outcomes in the post-intervention period relative to baseline. The whiskers indicate the 95% confidence intervals. We also report the p-value for the null hypothesis of no difference between the treatment and control groups.

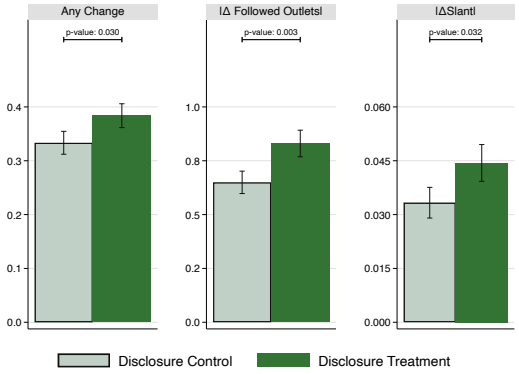
Figure A-21: Interactions between Disclosure and Peer Information Conditions



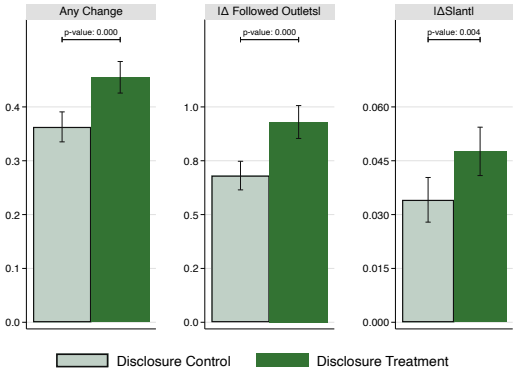
Notes: This figure reports the mean of the outcome across participants assigned to any factorial combination between the treatment and control groups in the peer information and disclosure conditions. The outcomes are: an indicator variable if participants update their beliefs regarding the slant of the news consumed by their peers (top), and an indicator variable if a participant makes any change to the news outlets followed (bottom). In all cases, we compare these outcomes in the post-intervention period vs. baseline. The difference between the treatment and control groups in the disclosure condition is that only participants in the former group are asked to reveal to their peers which news outlets they follow. The difference between the treatment and control groups in the peer information condition is that only the former receives a signal about the slant of the news diets of their peers. The whiskers indicate the 95% confidence intervals. We also report the p-value for the null hypothesis of no difference between each treatment group and the control group in both the disclosure and peer information conditions (shown in the leftmost bar).

Figure A-22: Effect of Disclosure Condition on News Diets – Sample Selection

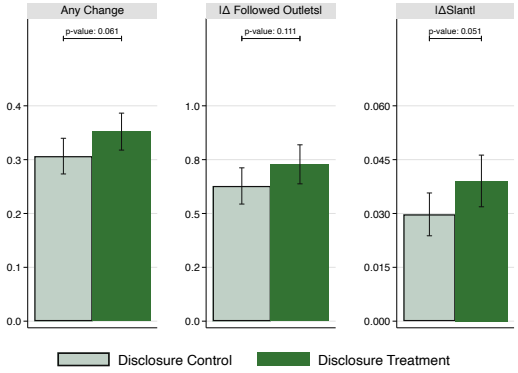
A. Weighting on Observables



B. Don't Follow Politics Closely



C. Follow Politics Closely



Notes: This figure assesses the robustness of the treatment effects estimated in Figure 2 to sample selection. Panel A weights the observations to match the observable characteristics in the U.S. adult population in terms of gender, education, ethnicity, and age. Panel B focuses on participants who report not following politics closely. Panel C focuses on participants who report following politics closely. Each panel estimates the mean of the following three outcomes across participants assigned to the treatment and control groups in the disclosure condition: an indicator variable if the participant makes any change to their news diet (left), the change in the number of news outlets followed (center), and the change in the absolute slant of their news diet (right). In all cases, we compare this outcome in the post-intervention period relative to the baseline. The difference between the treatment and control groups in the disclosure condition is that only participants in the former group are asked to reveal to their peers which news outlets they follow. The whiskers indicate the 95% confidence intervals. We also report the p-value for the null hypothesis of no difference between the treatment and control groups.

Table A-1: Balance Across Disclosure and Peer Information Conditions

	(1)	(2)	(3)	(4)	(5)	(6)
	Disclosure Condition			Peer Information Condition		
	Mean Control	Mean Treatment	Difference p-value	Mean Control	Mean Treatment	Difference p-value
A. Baseline Survey						
Male	0.69	0.67	0.09	0.68	0.68	0.83
Age	50.64	50.36	0.54	49.82	50.67	0.14
White	0.90	0.91	0.30	0.88	0.91	0.01
Graduate Degree	0.49	0.48	0.54	0.48	0.49	0.74
Ideology	3.05	2.99	0.25	2.98	3.03	0.52
Followers Ideology	3.07	3.04	0.43	3.08	3.05	0.61
Followings Ideology	3.21	3.16	0.29	3.15	3.19	0.59
News Diet	3.29	3.23	0.21	3.24	3.26	0.65
Followers News Diet	3.29	3.29	0.84	3.25	3.30	0.38
Followings News Diet	3.23	3.20	0.55	3.20	3.22	0.77
B. X Data						
Slant	-0.09	-0.09	0.63	-0.09	-0.09	0.89
Peer Slant	-0.05	-0.05	0.25	-0.05	-0.05	0.98
Number of Followers	1,230.01	1,044.12	0.48	1,096.04	1,148.12	0.87
Number of Followings	947.51	1,050.46	0.04	1,043.60	987.57	0.42
Number of Outlets	7.52	8.36	0.10	7.69	7.99	0.57

Notes: This table reports means of pre-treatment covariates by treatment status, with Columns (1) and (2) presenting results for the disclosure condition and Columns (4) and (5) for the peer information condition. Columns (3) and (6) report the p-values for the null hypothesis of no difference between the treatment and control groups. Panels A and B report covariates based on the baseline questionnaire and data scraped from X, respectively. The difference between the treatment and control groups in the disclosure condition is that only participants in the former group are asked to reveal to their peers which news outlets they follow. The difference between the treatment and control groups in the peer information condition is that only the former receives a signal about the slant of the news diets of their peers.

Table A-2: Descriptive Statistics for Demographic Characteristics and X Engagement

	(1)	(2)
A. Demographics Characteristics	Experiment Sample	U.S. Adult Population
Male	0.68	0.49
Age	50.50	47.6
White	0.91	0.79
Graduate degree	0.49	0.14
B. X Engagement Statistics	Experiment Sample	Representative Panel of X Users
Median Non-like actions per month for...		
Sample with Bottom 90% Engagement	20	2
Sample with Top 10% Engagement	816	138
Median Likes per month for...		
Sample with Bottom 90% Engagement	78	1
Sample with Top 10% Engagement	197	70
Median Followers for...		
Sample with Bottom 90% Engagement	124	19
Sample with Top 10% Engagement	672	387
Median Accounts Followed for...		
Sample with Bottom 90% Engagement	469	456
Sample with Top 10% Engagement	1,014	74

Notes: This table reports summary statistics on demographic characteristics (Panel A) and X engagement measures (Panel B). Column (1) presents these statistics for the experimental sample, while Column (2) presents them for a representative sample of U.S. adults (Panel A) or X users (Panel B). Panel B reports the median monthly number of non-like (defined as the sum of posts, reposts, quotes and replies) and likes actions, as well as the median number of followers and accounts followed. These statistics are reported separately for users in the top 10% or bottom 90% of the distribution of non-like actions. Data in Column (2) come from U.S. Census data (Panel A) and Pew Research (Panel B). See `https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/`. To estimate monthly rates of likes and non-like actions, we account for the fact that the X API limits observable actions to the most recent 200 per category. Engagement measures are therefore right-censored for highly active users. When fewer than 200 actions occur within 30 days, we use the observed count. When the 200-action cap binds, we estimate the monthly rate by dividing 200 by the fraction of a month required to generate those 200 actions. In our data, 16.5% of non-like actions and 4.5% of likes are censored.

Table A-3: Robustness to Controls in the Disclosure Condition I

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Outcome is Posts Placebo (N=3,755)							
Disclosure Treatment	-0.094	-0.093	-0.093	-0.092	-0.095	-0.093	-0.093
	(0.010)	(0.011)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
B. Outcome is Posts News Diet (N=3,755)							
Disclosure Treatment	0.082	0.081	0.082	0.081	0.080	0.079	0.082
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
C. Outcome is Any Change (N=3,755)							
Disclosure Treatment	0.073	0.071	0.073	0.077	0.070	0.072	0.074
	(0.016)	(0.016)	(0.016)	(0.016)	(0.015)	(0.015)	(0.015)
D. Outcome is |ΔFollowed Outlets| (N=3,754)							
Disclosure Treatment	0.189	0.187	0.191	0.198	0.181	0.190	0.189
	(0.040)	(0.040)	(0.040)	(0.040)	(0.039)	(0.040)	(0.039)
E. Outcome is |ΔSlant| (N=3,752)							
Disclosure Treatment	0.012	0.012	0.012	0.012	0.012	0.013	0.012
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
F. Outcome Is Don't Change, Sample of Peers and Center Aligned (N=2,458)							
Disclosure Treatment	-0.085	-0.086	-0.086	-0.087	-0.083	-0.088	-0.086
	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
G. Outcome Is Against Peers/Against Center, Sample of Peers and Center Aligned (N=2,458)							
Disclosure Treatment	0.005	0.002	0.005	0.006	0.004	0.003	0.003
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
H. Outcome Is Towards Peers/Towards Center, Sample of Peers and Center Aligned (N=2,458)							
Disclosure Treatment	0.080	0.085	0.081	0.081	0.079	0.085	0.080
	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)
I. Outcome Is Don't Change, Sample of Peers and Center Conflicting (N=470)							
Disclosure Treatment	-0.036	-0.018	-0.034	-0.044	-0.022	-0.014	-0.038
	(0.046)	(0.048)	(0.046)	(0.046)	(0.046)	(0.050)	(0.046)
J. Outcome Is Towards Peers/Against Center, Sample of Peers and Center Conflicting (N=470)							
Disclosure Treatment	-0.028	-0.042	-0.029	-0.038	-0.036	-0.048	-0.029
	(0.039)	(0.041)	(0.039)	(0.039)	(0.039)	(0.043)	(0.039)
K. Outcome Is Against Peers/Towards Center, Sample of Peers and Center Conflicting (N=470)							
Disclosure Treatment	0.064	0.061	0.063	0.082	0.058	0.062	0.063
	(0.037)	(0.040)	(0.037)	(0.037)	(0.038)	(0.042)	(0.037)
Controls:	None	Demographics	Political Interest	Ideology	Platform	All	All DML1

Notes: This table examines the robustness of the treatment effects of the disclosure condition, reported in Figures 1, 2 and 3, to the inclusion of controls. The outcome is specified in each panel header. Column 1 reports the baseline estimate with no controls. Column 2 includes year of birth and dummy variables for education status, ethnicity, and state of residence. Column 3 adds self-reported measures of political engagement, likelihood of voting, and political knowledge based on four true/false questions about relevant political events. Column 4 includes self-reported measures of ideology and news ideology for the participant, their followers, and the accounts they follow, as well as feeling-thermometer questions about political parties and candidates. Column 5 incorporates statistics from the participant's X account, including the number of followers, the number of accounts followed, an index of the credibility of the news outlets followed, and the device used to complete the survey (e.g., Windows, Mac). Column 6 includes all aforementioned controls. Column 7 reports the estimate from the double machine learning approach of Chernozhukov et al. [2018] (algorithm DML1), selecting control variables from the full set of available covariates. Standard errors reported in parentheses are heteroskedasticity robust.

Table A-4: Robustness to Controls in the Disclosure Condition II

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Outcome is Any Change Over Three Months After Intervention (N=2,402)							
Disclosure Treatment	0.059	0.068	0.061	0.063	0.053	0.062	0.062
	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.019)
B. Outcome is |ΔFollowed Outlets| Over Three Months After Intervention (N=2,402)							
Disclosure Treatment	0.122	0.137	0.123	0.128	0.090	0.097	0.107
	(0.072)	(0.074)	(0.072)	(0.071)	(0.070)	(0.070)	(0.071)
C. Outcome is |ΔSlant| Over Three Months After Intervention (N=2,402)							
Disclosure Treatment	0.008	0.008	0.008	0.008	0.009	0.009	0.009
	(0.006)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)
D. Outcome Is News Engagement Index (N=3,755)							
Disclosure Treatment	0.077	0.081	0.078	0.081	0.069	0.073	0.075
	(0.028)	(0.027)	(0.028)	(0.028)	(0.026)	(0.026)	(0.027)
E. Outcome Is Non-News Engagement Index (N=3,755)							
Disclosure Treatment	0.009	0.005	0.010	0.007	-0.002	-0.009	-0.010
	(0.022)	(0.022)	(0.022)	(0.022)	(0.021)	(0.021)	(0.021)
Controls:	None	Demographics	Political Interest	Ideology	Platform	All	All DML1

Notes: This table examines the robustness of the treatment effects of the disclosure condition reported in Figures 4 and 5 to the inclusion of controls. The outcome is specified in each panel header. Column 1 reports the baseline estimate with no controls. Column 2 includes year of birth and dummy variables for education status, ethnicity, and state of residence. Column 3 adds self-reported measures of political engagement, likelihood of voting, and political knowledge based on four true/false questions about relevant political events. Column 4 includes self-reported measures of ideology and news ideology for the participant, their followers, and the accounts they follow, as well as feeling-thermometer questions about political parties and candidates. Column 5 incorporates statistics from the participant's X account, including the number of followers, the number of accounts followed, an index of the credibility of the news outlets followed, and the device used to complete the survey (e.g., Windows, Mac). Column 6 includes all aforementioned controls. Column 7 reports the estimate from the double machine learning approach of Chernozhukov et al. [2018] (algorithm DML1), selecting control variables from the full set of available covariates. Standard errors reported in parentheses are heteroskedasticity robust.

Table A-5: Effect of Peer Signal on Participants' News Diets After the Intervention

	(1)	(2)	(3)	(4)
	Change in Participants' Own News Diets			
	Slant Change		sign(Slant Change)	
A. Treatment Effects on Signal-Oriented Outcomes, Full Sample (N=2,355)				
Peer Treatment	-0.002	-0.012	-0.005	0.033
	(0.008)	(0.012)	(0.034)	(0.056)
|Signal − Prior|		-0.010		0.009
		(0.006)		(0.030)
Peer Treatment × |Signal − Prior|		0.007		-0.030
		(0.007)		(0.033)
Constant	-0.003	0.010	0.013	0.002
	(0.007)	(0.011)	(0.031)	(0.051)
B. Effects of a Peer Signal with Right Noise, Peer Treatment Sample (N=1,895)				
Signal with Right Noise	-0.009	-0.010	-0.009	-0.018
	(0.007)	(0.010)	(0.029)	(0.044)
|Signal − Truth|		0.005		0.000
		(0.005)		(0.025)
Signal with Right Noise × |Signal − Truth|		0.000		0.008
		(0.007)		(0.034)
Constant	-0.001	-0.005	-0.019	-0.019
	(0.005)	(0.007)	(0.021)	(0.030)
C. Effects of a Peer Signal with Right Noise, Peer Control Sample (N=458)				
Signal with Right Noise	0.002	0.005	0.046	0.074
	(0.013)	(0.021)	(0.062)	(0.088)
|Signal − Truth|		0.003		0.027
		(0.007)		(0.036)
Signal with Right Noise × |Signal − Truth|		-0.003		-0.026
		(0.011)		(0.058)
Constant	-0.009	-0.012	-0.050	-0.079
	(0.009)	(0.015)	(0.045)	(0.060)

Notes: This table examines the longer-term effects of the peer signal on changes in the slant of the participants' news diets three months following the intervention. Columns (1) and (2) report effects on the continuous change, and Columns (3) and (4) report effects on the discrete direction of change. Panel A reports the treatment effects of randomly receiving a signal about the slant of the news diets of the participant's peers, following Equation (1). The outcome variables for Panel A are signal-oriented, as defined in Equation (3). These estimates indicate the extent to which the peer signal leads participants to update the slant of their news diets in the direction of the surprise. Panels B and C report the estimated impact of randomly receiving a signal with right-noise, using Equation (2), for participants in the peer information treatment and control groups, respectively. The outcome variables for Panel B and C are not signal-oriented. These estimates indicate the extent to which a signal with right-noise leads participants to shift the slant of their news diets to the right. The even-numbered columns present heterogeneous treatment effects by the magnitude of the surprise, measured in standard-deviation units. Standard errors reported in parentheses are heteroskedasticity robust.

Table A-6: Effect of Peer Signal on Beliefs About the Slant of the News Diet of Accounts Followed

	(1)	(2)	(3)	(4)
Treatment Effects on Signal-Oriented Outcomes				
	A. Change in Beliefs About News Diet of Accounts Followed (N=2,695)			
	Slant Change		Sign(Slant Change)	
Peer Treatment	0.114	0.013	0.106	0.051
	(0.046)	(0.065)	(0.031)	(0.042)
Mutual Connection		-0.060		-0.030
		(0.044)		(0.024)
Peer Treatment × Mutual Connection		0.104		0.056
		(0.049)		(0.028)
Constant	0.089	0.149	0.039	0.069
	(0.041)	(0.058)	(0.028)	(0.037)
	B. Change in Participants' Own News Diets (N=3,683)			
	Slant Change		Sign(Slant Change)	
Peer Treatment	-0.003	0.001	-0.013	0.004
	(0.005)	(0.007)	(0.026)	(0.035)
Mutual Connection		0.006		0.013
		(0.004)		(0.019)
Peer Treatment × Mutual Connection		-0.004		-0.017
		(0.005)		(0.023)
Constant	0.001	-0.004	0.006	-0.007
	(0.005)	(0.007)	(0.023)	(0.032)

Notes: This table reports the treatment effects of of the peer information treatment. Panel A examines the effect of the treatment on changes in participants' beliefs about the average slant of the news diets of the accounts they follow, and Panel B examines changes in the slant of the participants' own news diets. All outcome variables are signal-oriented, as defined in Equation (3). Columns (1) and (2) report effects on the continuous change, and Columns (3) and (4) report effects on the discrete direction of change. The even-numbered columns present heterogeneous treatment effects by the share of mutual connections (size of the intersection between a participant's followers and the accounts they follow, divided by the size of the union of these two sets), measured in standard-deviation units. Standard errors reported in parentheses are heteroskedasticity robust.

Table A-7: Heterogeneous Effect of Peer Signal by Measures of Signal Salience

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Effects on Signal-Oriented Outcomes							
	Change in Beliefs About Peers' News Diets				Change in Participants' Own News Diets			
	Slant Change		sign(Slant Change)		Slant Change		sign(Slant Change)	
A. Heterogeneous Treatment Effects by Engagement with Peers in Signal								
Peer Treatment	0.288	0.284	0.212	0.209	-0.003	-0.003	-0.013	-0.022
	(0.043)	(0.047)	(0.029)	(0.032)	(0.005)	(0.006)	(0.026)	(0.029)
Signal Engagement		0.053		0.023		0.003		-0.007
		(0.096)		(0.061)		(0.016)		(0.060)
Peer Treatment × Signal Engagement		0.022		0.021		0.001		0.049
		(0.109)		(0.069)		(0.017)		(0.066)
Constant	0.267	0.256	0.163	0.159	0.001	0.001	0.006	0.007
	(0.038)	(0.042)	(0.026)	(0.029)	(0.005)	(0.005)	(0.023)	(0.026)
B. Heterogeneous Treatment Effects by Number of Followers								
Peer Treatment	0.288	0.356	0.212	0.226	-0.003	0.001	-0.013	-0.004
	(0.043)	(0.055)	(0.029)	(0.040)	(0.005)	(0.006)	(0.026)	(0.039)
Low followers		0.065		-0.017		0.009		0.044
		(0.076)		(0.051)		(0.010)		(0.047)
Peer Treatment × Low followers		-0.140		-0.028		-0.009		-0.017
		(0.086)		(0.057)		(0.011)		(0.052)
Constant	0.267	0.236	0.163	0.171	0.001	-0.003	0.006	-0.016
	(0.038)	(0.047)	(0.026)	(0.036)	(0.005)	(0.006)	(0.023)	(0.035)

Notes: This table reports heterogeneous treatment effect of the peer information treatment for two different measures of salience of the peer signal. Panel A uses an indicator equal to one for participants with above-median engagement with peers who were randomly selected to be used in calculating the peer signal. Panel B uses an indicator equal to one for participants with a below-median number of followers. The first four columns examine the effects on changes in participants' beliefs about the average slant of the news diets of the accounts they follow (as in Panel A of Table 1), and the last four columns examine changes in the actual slant of the participants' news diets (as in Panel A of Table 2). All outcome variables are signal-oriented, as defined in Equation (3). Columns (1), (2), (5) and (6) report effects on the (continuous) change, whereas Columns (3), (4), (7) and (8) report effects on the discrete change (-1 for a leftward shift, 0 for no change, and 1 for a rightward shift). The even-numbered columns present heterogeneous treatment effects by the magnitude of the surprise, measured in standard-deviation units. Standard errors reported in parentheses are heteroskedasticity robust.

Table A-8: Effects of Peer Signal on Beliefs About Peers' Ideology

	(1)	(2)	(3)	(4)
	Change in Beliefs About Peers' News Diets			
	Slant Change		sign(Slant Change)	
A. Treatment Effects on Signal-Oriented Outcomes, Full Sample (N=2,695)				
Peer Treatment	0.188	-0.043	0.131	-0.030
	(0.035)	(0.056)	(0.027)	(0.043)
|Signal − Prior|		0.002		-0.005
		(0.031)		(0.023)
Peer Treatment × |Signal − Prior|		0.183		0.127
		(0.037)		(0.026)
Constant	-0.033	-0.036	-0.030	-0.023
	(0.030)	(0.048)	(0.023)	(0.037)
B. Effects of a Peer Signal with Right Noise, Peer Treatment Sample (N=2,150)				
Signal with Right Noise	0.367	0.123	0.266	0.099
	(0.037)	(0.056)	(0.027)	(0.041)
|Signal − Truth|		-0.145		-0.107
		(0.030)		(0.019)
Signal with Right Noise × |Signal − Truth|		0.233		0.161
		(0.043)		(0.028)
Constant	-0.067	0.074	-0.037	0.067
	(0.025)	(0.038)	(0.019)	(0.027)
C. Effects of a Peer Signal with Right Noise, Peer Control Sample (N=539)				
Signal with Right Noise	-0.007	-0.002	-0.007	-0.003
	(0.060)	(0.094)	(0.046)	(0.070)
|Signal − Truth|		-0.005		-0.006
		(0.030)		(0.022)
Signal with Right Noise × |Signal − Truth|		-0.004		-0.004
		(0.050)		(0.043)
Constant	-0.023	-0.017	-0.019	-0.012
	(0.044)	(0.068)	(0.031)	(0.047)

Notes: This table presents the effects of the peer signal on changes in beliefs about peers' ideology, in contrast to Table 1, which focuses on the slant or ideology of peers' news diets. Columns (1) and (2) report effects on the continuous change, and Columns (3) and (4) report effects on the discrete direction of change. Panel A reports the treatment effects of randomly receiving a signal about the slant of the news diets of the participant's peers, following Equation (1). The outcome variables for Panel A are signal-oriented, as defined in Equation (3). Panels B and C report the estimated impact of randomly receiving a signal with right-noise, using Equation (2), for participants in the peer information treatment and control groups, respectively. The outcome variables for Panel B and C are not signal-oriented. The even-numbered columns present heterogeneous treatment effects by the magnitude of the surprise, measured in standard-deviation units. Standard errors reported in parentheses are heteroskedasticity robust.

Table A-9: Effect of Peer Signals on Participants' Own News Diet — Sample with Non-Missing Beliefs

	(1)	(2)	(3)	(4)
	Change in Participants' Own News Diets			
	Slant Change		sign(Slant Change)	
A. Treatment Effects on Signal-Oriented Outcomes, Full Sample (N=2,695)				
Peer Treatment	-0.005	-0.008	-0.019	0.002
	(0.007)	(0.012)	(0.031)	(0.050)
|Signal − Prior|		-0.003		0.006
		(0.006)		(0.027)
Peer Treatment × |Signal − Prior|		0.002		-0.016
		(0.006)		(0.030)
Constant	0.003	0.007	0.007	0.000
	(0.006)	(0.011)	(0.028)	(0.045)
B. Effects of a Peer Signal with Right Noise, Peer Treatment Sample (N=2,150)				
Signal with Right Noise	0.001	-0.002	-0.013	0.003
	(0.005)	(0.008)	(0.028)	(0.040)
|Signal − Truth|		-0.002		-0.020
		(0.004)		(0.022)
Signal with Right Noise × |Signal − Truth|		0.003		-0.010
		(0.005)		(0.028)
Constant	-0.006	-0.004	-0.019	-0.000
	(0.004)	(0.006)	(0.019)	(0.028)
C. Effects of a Peer Signal with Right Noise, Peer Control Sample (N=539)				
Signal with Right Noise	-0.003	0.012	0.000	0.117
	(0.012)	(0.016)	(0.056)	(0.079)
|Signal − Truth|		0.004		0.058
		(0.004)		(0.036)
Signal with Right Noise × |Signal − Truth|		-0.013		-0.105
		(0.006)		(0.054)
Constant	-0.004	-0.009	-0.019	-0.083
	(0.009)	(0.012)	(0.039)	(0.053)

Notes: This table presents the effects of the peer signal on changes in the actual slant of the news that participants consume, as in Table 2, but restricted to the subsample of participants with non-missing beliefs (i.e., those who reported beliefs in the endline questionnaire). presents the main effects of the peer signal on changes in the slant of the participants' news diets. Columns (1) and (2) report effects on the continuous change, and Columns (3) and (4) report effects on the discrete direction of change. Panel A reports the treatment effects of randomly receiving a signal about the slant of the news diets of the participant's peers, following Equation (1). The outcome variables for Panel A are signal-oriented, as defined in Equation (3). These estimates indicate the extent to which the peer signal leads participants to update the slant of their news diets in the direction of the surprise. Panels B and C report the estimated impact of randomly receiving a signal with right-noise, using Equation (2), for participants in the peer information treatment and control groups, respectively. The outcome variables for Panel B and C are not signal-oriented. These estimates indicate the extent to which a signal with right-noise leads participants to shift the slant of their news diets to the right. The even-numbered columns present heterogeneous treatment effects by the magnitude of the surprise, measured in standard-deviation units. Standard errors reported in parentheses are heteroskedasticity robust.

Table A-10: Robustness to Controls in the Peer Information Condition I

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Treatment Effects on Signal-Oriented Outcomes, Full Sample (N=2,695)							
A1. Outcome is Signal-Oriented Change in Beliefs About Peers							
Peer Treatment	0.288	0.286	0.293	0.287	0.288	0.290	0.289
	(0.043)	(0.043)	(0.042)	(0.042)	(0.043)	(0.043)	(0.043)
A2. Outcome is Signal-Oriented sign(Slant Change) in Beliefs About Peers							
Peer Treatment	0.212	0.207	0.212	0.213	0.213	0.209	0.211
	(0.029)	(0.029)	(0.029)	(0.029)	(0.029)	(0.029)	(0.029)
B. Effects of a Peer Signal with Right Noise, Peer Treatment Sample (N=2,150)							
B1. Outcome is Change in Beliefs About Peers							
Signal with Right Noise	0.453	0.452	0.456	0.421	0.461	0.415	0.421
	(0.046)	(0.046)	(0.045)	(0.036)	(0.045)	(0.036)	(0.036)
B2. Outcome is sign(Slant Change) in Beliefs About Peers							
Signal with Right Noise	0.273	0.272	0.276	0.252	0.279	0.250	0.254
	(0.030)	(0.030)	(0.030)	(0.025)	(0.030)	(0.025)	(0.025)
C. Effects of a Peer Signal with Right Noise, Peer Control Sample (N=539)							
C1. Outcome is Change in Beliefs About Peers							
Signal with Right Noise	-0.014	-0.035	-0.010	-0.003	-0.011	-0.021	-0.014
	(0.079)	(0.086)	(0.078)	(0.063)	(0.078)	(0.070)	(0.064)
C2. Outcome is sign(Slant Change) in Beliefs About Peers							
Signal with Right Noise	0.023	0.021	0.021	0.023	0.026	0.022	0.037
	(0.053)	(0.059)	(0.053)	(0.046)	(0.053)	(0.051)	(0.047)
Controls:	None	Demographics	Political Interest	Ideology	Platform	All	All DML1

Notes: This table examines the robustness of the results reported in Table 1 to the inclusion of controls. The outcome is the change in beliefs about about the average slant of their peers' news diets. Column 1 reports the baseline estimate with no controls. Column 2 includes year of birth and dummy variables for education status, ethnicity, and state of residence. Column 3 adds self-reported measures of political engagement, likelihood of voting, and political knowledge based on four true/false questions about relevant political events. Column 4 includes self-reported measures of ideology and news ideology for the participant, their followers, and the accounts they follow, as well as feeling-thermometer questions about political parties and candidates. Column 5 incorporates statistics from the participant's X account, including the number of followers, the number of accounts followed, an index of the credibility of the news outlets followed, and the device used to complete the survey (e.g., Windows, Mac). Column 6 includes all aforementioned controls. Column 7 reports the estimate from the double machine learning approach of Chernozhukov et al. [2018] (algorithm DML1), selecting control variables from the full set of available covariates. Standard errors reported in parentheses are heteroskedasticity robust.

Table A-11: Robustness to Controls in the Peer Information Condition II

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Treatment Effects on Signal-Oriented Outcomes, Full Sample (N=3,683)							
A1. Outcome is Signal-Oriented Slant Change in Beliefs About Peers							
Peer Treatment	-0.003	-0.003	-0.003	-0.003	-0.004	-0.003	-0.004
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
A2. Outcome is Signal-Oriented sign(Slant Change) in Beliefs About Peers							
Peer Treatment	-0.013	-0.012	-0.012	-0.013	-0.014	-0.011	-0.013
	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)
B. Effects of a Peer Signal with Right Noise, Peer Treatment Sample (N=2,951)							
B1. Outcome is Slant Change in Beliefs About Peers							
Signal with Right Noise	-0.001	-0.001	-0.001	0.001	-0.002	0.000	0.000
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
B2. Outcome is sign(Slant Change) in Beliefs About Peers							
Signal with Right Noise	-0.018	-0.011	-0.018	-0.002	-0.020	-0.003	-0.004
	(0.022)	(0.023)	(0.022)	(0.022)	(0.023)	(0.023)	(0.022)
C. Effects of a Peer Signal with Right Noise, Peer Control Sample (N=722)							
C1. Dependent Variable is Slant Change in Beliefs About Peers							
Signal with Right Noise	0.004	0.002	0.003	0.003	0.002	0.000	0.003
	(0.010)	(0.011)	(0.010)	(0.010)	(0.010)	(0.011)	(0.010)
C2. Dependent Variable is sign(Slant Change) in Beliefs About Peers							
Signal with Right Noise	0.042	0.030	0.037	0.041	0.037	0.026	0.043
	(0.047)	(0.048)	(0.046)	(0.046)	(0.047)	(0.049)	(0.046)
Controls:	None	Demographics	Political Interest	Ideology	Platform	All	All DML1

Notes: This table examines the robustness to the inclusion of controls of the results reported in Table 2. Column 1 reports the baseline estimate with no controls. Column 2 includes year of birth and dummy variables for education status, ethnicity, and state of residence. Column 3 adds self-reported measures of political engagement, likelihood of voting, and political knowledge based on four true/false questions about relevant political events. Column 4 includes self-reported measures of ideology and news ideology for the participant, their followers, and the accounts they follow, as well as feeling-thermometer questions about political parties and candidates. Column 5 incorporates statistics from the participant's X account, including the number of followers, the number of accounts followed, an index of the credibility of the news outlets followed, and the device used to complete the survey (e.g., Windows, Mac). Column 6 includes all aforementioned controls. Column 7 reports the estimate from the double machine learning approach of Chernozhukov et al. [2018] (algorithm DML1), selecting control variables from the full set of available covariates. Standard errors reported in parentheses are heteroskedasticity robust.

Table A-12: Summary Statistics for the Natural Experiment on the Likes Policy Change

	(1)	(2)	(3)	(4)	(5)
	All	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Number of Posts	9,073,600	2,275,026	2,269,033	2,278,988	2,250,553
Views per Post	8,822.0	5,457.9	11,215.0	6,460.7	12,201.2
Likes per Post	25.2	9.8	18.8	12.1	60.6
Quotes per Post	1.0	0.6	1.2	0.7	1.7
Reposts per Post	6.6	2.5	4.8	3.4	15.7
Replies per Post	5.0	3.2	4.6	2.5	9.8
Bookmarks per Post	1.1	0.6	1.1	0.7	2.1

Notes: This table reports summary statistics for the natural experiment on the Likes Policy Change on the X platform. The first row contains the total number of posts, while the remaining rows contain the mean of each variable. The first column reports the values for all posts, while the remaining four columns report the values for each quartile of outlet slant.

Table A-13: Alternative Normalizations of the Peer Signal, Prior, and Surprise

Let $\widetilde{\sigma}$ and σ represent the non-normalized and normalized variables respectively, then:

Normalization 1 (Baseline): Equal-width discretization of the signal (divide range $[-1, 1]$ into seven bins with width $2/7$).

$$\sigma_i^{\text{prior}} = \widetilde{\sigma}_i^{\text{prior}};$$
$$\sigma_i^{\text{signal}} = k \quad \text{if } \widetilde{\sigma}_i^{\text{signal}} \in \left[-1 + \tfrac{2}{7}(k-1),\ -1 + \tfrac{2}{7}k\right],\ k = 1, \ldots, 7$$

Normalization 2 (Mean cutoffs): Use mean signal values within each prior category as anchors to define cutoffs:

$$\sigma_i^{\text{prior}} = \widetilde{\sigma}_i^{\text{prior}};$$
$$\sigma_i^{\text{signal}} = \arg\min_k |\mu_k - \tilde{\sigma}_i^{\text{signal}}|$$
$$\text{where } \mu_k = \mathbb{E}[\widetilde{\sigma}_i^{\text{signal}} \mid \widetilde{\sigma}_i^{\text{prior}} = k] \text{ (for } k = 1, \ldots, 7).$$

Normalization 3 (Median cutoffs): Same as Normalization 2, but use medians instead of means:

$$\sigma_i^{\text{prior}} = \widetilde{\sigma}_i^{\text{prior}};$$
$$\sigma_i^{\text{signal}} = \arg\min_k |m_k - \tilde{\sigma}_i^{\text{signal}}|$$
$$\text{where } m_k = \text{Median}[\widetilde{\sigma}_i^{\text{signal}} \mid \widetilde{\sigma}_i^{\text{prior}} = k] \text{ (for } k = 1, \ldots, 7).$$

Normalization 4 (Uniform transformation): Apply a linear transformation to the signal mapping $[-1, 1]$ to $[1, 7]$:

$$\sigma_i^{\text{prior}} = \widetilde{\sigma}_i^{\text{prior}};$$
$$\sigma_i^{\text{signal}} = 3\,\widetilde{\sigma}_i^{\text{signal}} + 4$$

Normalization 5 (Standardization): Standardize both variables by subtracting their means and dividing by their standard deviations:

$$\sigma_i^{\text{prior}} = \frac{\widetilde{\sigma}_i^{\text{prior}} - \mu_{\text{prior}}}{sd_{\text{prior}}};$$
$$\sigma_i^{\text{signal}} = \frac{\widetilde{\sigma}_i^{\text{signal}} - \mu_{\text{signal}}}{sd_{\text{signal}}}$$

where $\mu_{\text{prior}}, \mu_{\text{signal}}$ are the sample means and $sd_{\text{prior}}, sd_{\text{signal}}$ are the sample standard deviations.

Table A-14: Effect of Peer Signal Under Alternative Normalizations of Peer Surprise

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Effects on Signal-Oriented Outcomes							
	Change in Beliefs About Peers' News Diets				Change in Participants' Own News Diets			
	Slant Change		sign(Slant Change)		Slant Change		sign(Slant Change)	
A. Surprise Under Normalization 2								
Peer Treatment	0.212	-0.058	0.142	0.007	-0.001	-0.003	-0.010	-0.004
	(0.043)	(0.075)	(0.029)	(0.046)	(0.006)	(0.009)	(0.026)	(0.044)
|Signal − Prior|		0.255		0.139		-0.002		-0.006
		(0.049)		(0.023)		(0.005)		(0.024)
Peer Treatment × |Signal − Prior|		0.193		0.096		0.001		-0.004
		(0.055)		(0.026)		(0.005)		(0.026)
Constant	0.227	-0.129	0.147	-0.048	0.001	0.003	0.016	0.025
	(0.038)	(0.067)	(0.026)	(0.040)	(0.005)	(0.009)	(0.023)	(0.040)
B. Surprise Under Normalization 3								
Peer Treatment	0.194	-0.144	0.129	-0.046	-0.005	-0.012	-0.016	-0.026
	(0.044)	(0.081)	(0.029)	(0.050)	(0.006)	(0.010)	(0.026)	(0.047)
|Signal − Prior|		0.248		0.133		-0.007		-0.025
		(0.050)		(0.024)		(0.005)		(0.024)
Peer Treatment × |Signal − Prior|		0.230		0.120		0.005		0.006
		(0.056)		(0.027)		(0.005)		(0.026)
Constant	0.216	-0.153	0.139	-0.059	0.005	0.015	0.027	0.064
	(0.038)	(0.072)	(0.026)	(0.044)	(0.005)	(0.010)	(0.023)	(0.042)
C. Surprise Under Normalization 4								
Peer Treatment	0.343	0.171	0.262	0.198	-0.004	-0.005	-0.014	0.020
	(0.044)	(0.081)	(0.029)	(0.053)	(0.005)	(0.011)	(0.026)	(0.046)
|Signal − Prior|		0.286		0.165		0.000		0.031
		(0.052)		(0.023)		(0.005)		(0.023)
Peer Treatment × |Signal − Prior|		0.121		0.045		0.001		-0.024
		(0.057)		(0.026)		(0.006)		(0.025)
Constant	0.178	-0.238	0.082	-0.158	0.001	0.000	-0.008	-0.054
	(0.039)	(0.074)	(0.026)	(0.047)	(0.005)	(0.010)	(0.023)	(0.041)
D. Surprise Under Normalization 5								
Peer Treatment	0.318	0.183	0.233	0.211	-0.005	-0.013	-0.024	-0.042
	(0.043)	(0.070)	(0.028)	(0.045)	(0.005)	(0.010)	(0.026)	(0.043)
|Signal − Prior|		0.324		0.176		-0.004		-0.003
		(0.052)		(0.024)		(0.005)		(0.024)
Peer Treatment × |Signal − Prior|		0.100		0.014		0.006		0.014
		(0.058)		(0.026)		(0.005)		(0.027)
Constant	0.263	-0.139	0.171	-0.048	0.004	0.010	0.022	0.026
	(0.038)	(0.063)	(0.025)	(0.041)	(0.005)	(0.009)	(0.023)	(0.039)

Notes: This table reports the treatment effects of the peer information treatment under alternative normalization of the surprise (See Section A.8). The first four columns examines the effects on changes in participants' beliefs about the slant of the news diets of their peers (as in Panel A of Table 1), and the last four columns examine changes in the participants' own news diets (as in Panel A of Table 2). All outcome variables are signal-oriented, as defined in Equation (3). Columns (1), (2), (5) and (6) report effects on the (continuous) change, whereas Columns (3), (4), (7) and (8) report effects on the discrete change (-1 for a leftward shift, 0 for no change, and 1 for a rightward shift). The even-numbered columns present heterogeneous treatment effects by the magnitude of the surprise, measured in standard-deviation units. Standard errors reported in parentheses are heteroskedasticity robust.

Table A-15: Effect of Peer Information Treatment – Sample Selection

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Effects on Signal-Oriented Outcomes							
	Change in Beliefs About Peers' News Diets				Change in Participants' Own News Diets			
	Slant Change		sign(Slant Change)		Slant Change		sign(Slant Change)	
A. Weighting on Observables								
Peer Treatment	0.182	0.206	0.215	0.126	-0.007	-0.011	-0.033	-0.008
	(0.136)	(0.202)	(0.047)	(0.083)	(0.008)	(0.015)	(0.033)	(0.052)
|Signal − Prior|		0.406		0.116		-0.004		0.005
		(0.204)		(0.044)		(0.008)		(0.028)
Peer Treatment × |Signal − Prior|		-0.035		0.060		0.003		-0.018
		(0.207)		(0.047)		(0.008)		(0.032)
Constant	0.363	-0.164	0.141	-0.009	-0.000	0.006	-0.004	-0.010
	(0.132)	(0.197)	(0.043)	(0.077)	(0.007)	(0.014)	(0.029)	(0.047)
B. Don't Follow Politics Closely								
Peer Treatment	0.280	0.038	0.238	0.049	0.006	0.008	0.014	0.089
	(0.070)	(0.115)	(0.043)	(0.069)	(0.008)	(0.015)	(0.038)	(0.059)
|Signal − Prior|		0.268		0.068		0.001		0.054
		(0.099)		(0.036)		(0.008)		(0.033)
Peer Treatment × |Signal − Prior|		0.172		0.144		-0.002		-0.061
		(0.105)		(0.040)		(0.008)		(0.037)
Constant	0.311	-0.006	0.155	0.075	-0.005	-0.006	-0.003	-0.069
	(0.062)	(0.105)	(0.037)	(0.060)	(0.008)	(0.014)	(0.034)	(0.053)
C. Follow Politics Closely								
Peer Treatment	0.303	0.008	0.195	0.073	-0.010	-0.019	-0.030	-0.068
	(0.052)	(0.085)	(0.039)	(0.067)	(0.007)	(0.014)	(0.035)	(0.059)
|Signal − Prior|		0.080		0.049		-0.006		-0.037
		(0.045)		(0.033)		(0.007)		(0.031)
Peer Treatment × |Signal − Prior|		0.241		0.101		0.007		0.027
		(0.053)		(0.037)		(0.007)		(0.034)
Constant	0.229	0.119	0.170	0.103	0.006	0.014	0.012	0.063
	(0.045)	(0.076)	(0.035)	(0.062)	(0.007)	(0.013)	(0.032)	(0.054)

Notes: This table explores the robustness of the treatment effects to sample characteristics. Panel A weights the observations to match the observable characteristics of the U.S. adult population, based on the observable characteristics reported in Table A-2. Panels B and C report estimates for the subsample of participants with low political interest (those who follow politics somewhat or not very closely) and high political interest, respectively. The first four columns examine the effects on changes in participants' beliefs about the average slant of the news diets of the accounts they follow (as in Panel A of Table 1), and the last four columns examine changes in the slant of participants' own news diets (as in Panel A of Table 2). All outcome variables are signal-oriented, as defined in Equation (3). Columns (1), (2), (5) and (6) report effects on the (continuous) change, whereas Columns (3), (4), (7) and (8) report effects on the discrete change (−1 for a leftward shift, 0 for no change, and 1 for a rightward shift). The even-numbered columns present heterogeneous treatment effects by the magnitude of the surprise, measured in standard-deviation units. Standard errors reported in parentheses are heteroskedasticity robust.

Table A-16: Instrumental Variables Estimates of the Disclosure and Peer Information Conditions

	(1)	(2)	(3)
A. Disclosure Condition			
	Any Change	|ΔFollowed Outlets|	|ΔSlant|
Posts News Diet	0.893	2.310	0.143
	(0.202)	(0.515)	(0.043)
Constant	0.320	0.607	0.029
	(0.015)	(0.036)	(0.003)
B. Peer Information Condition			
	Change in Participants' Own News Diets		
	Slant Change	sign(Slant Change)	
Change in Beliefs About Peers' News Diets	-0.017	-0.087	
	(0.023)	(0.149)	
Constant	0.008	0.022	
	(0.012)	(0.051)	

Notes: This table reports 2SLS estimates exploiting variation in the randomized assignment in the disclosure and peer information conditions as instruments in Panels A and B, respectively. Panel A reports 2SLS estimates of the effect of a participant disclosing their news diet through a post on X on the three outcomes shown in Figure 2: an indicator for whether the participant makes any change to their news diet in Column (1), the absolute change in the number of news outlets followed in Column (2), and the change in the absolute slant of their news diet in Column (3). Panel B reports 2SLS estimates of the effect of changes in beliefs about the average slant of peers' news diets on changes in the slant of participants' own news diets, measured both continuously in Column (1) and discretely in Column (2). The outcomes and endogenous variables for Panel B are signal-oriented, as defined in Equation (3). Standard errors reported in parentheses are heteroskedasticity robust.