

# Can ♥s Change Minds? Social Media Endorsements and Policy Preferences

Pierluigi Conzo<sup>1,✉</sup>, Laura K. Taylor<sup>2,✉</sup>, Juan S. Morales<sup>3,✉</sup>, Margaret Samahita<sup>2,✉</sup>, and Andrea Gallice<sup>1</sup>

Social Media + Society  
April-June 2023: 1–25  
© The Author(s) 2023  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/20563051231177899  
journals.sagepub.com/home/sms



## Abstract

We investigate the effect of social media endorsements (likes, retweets, shares) on individuals' policy preferences. In two pre-registered online experiments ( $N = 1,384$ ), we exposed participants to non-neutral policy messages about the COVID-19 pandemic (emphasizing either public health or economic activity as a policy priority) while varying the level of endorsements of these messages. Our experimental treatment did not result in aggregate changes to policy views. However, our analysis indicates that active social media users did respond to the variation in engagement metrics. In particular, we find a strong positive treatment effect concentrated on a minority of individuals who correctly answered a factual manipulation check regarding the endorsements. Our results suggest that though only a fraction of individuals appear to pay conscious attention to endorsement metrics, they may be influenced by these social cues.

## Keywords

social media, social conformity, political polarization, COVID-19

## Introduction

Social media has been hypothesized to have broad effects on politics (Zhuravskaya et al., 2020). However, the magnitude of these effects and the mechanisms through which they arise remain debated. This article studies how social media affects individuals' policy preferences. In particular, we study endorsements, as evinced by common metrics of engagement: *likes*, ♥s, ↻s, *retweets*, and *shares*. Social endorsements are a central feature of social media and are observed by billions of individuals around the world, thus even small effects may have important consequences for policy-making and political dynamics. Can the perceived support of social media messages affect how individuals evaluate policies?

To answer this question, we conducted two pre-registered online experimental studies in Europe (Ireland,  $n = 305$ , and Italy,  $n = 300$ ) and the United States ( $n = 779$ ) in the context of the COVID-19 pandemic and its policy trade-offs (public health vs. economic activity, Settele & Shupe, 2022).<sup>1</sup> The experiment allows us to isolate the effects of perceived support for policy choices in a controlled environment different from individuals' own social media. We study endorsements, a specific but important feature of social media, but without conflating issues of social image, peer effects, or selective exposure. Instead, we exposed individuals to strangers' tweets and endorsements, and examined their effects on

individuals' policy preferences in an anonymous survey. More specifically, we exposed participants to non-neutral policy messages about the COVID-19 pandemic, manipulated the perceived level of endorsements of these messages, and examined how this affected their policy attitudes.

We find no overall effects of endorsements on policy views. On aggregate, we estimate a precise zero effect of our treatment on the policy views of our participants. A factual manipulation check suggests that most individuals pay little attention to endorsement metrics, with only one-third of participants correctly answering a post-treatment question about these metrics.

One particular focus of our work studies the effects of endorsement metrics on active social media users, which we define as those who use Facebook or Twitter for at least 1 hr a day. These are individuals who are frequently exposed to engagement metrics and are thus likely to be sensitive and

<sup>1</sup>University of Turin and Collegio Carlo Alberto, Italy

<sup>2</sup>University College Dublin, Ireland

<sup>3</sup>Wilfrid Laurier University, Canada

## Corresponding Author:

Juan S. Morales, Department of Economics, Lazaridis School of Business and Economics, Wilfrid Laurier University, 75 University Ave West, N2L3C5, Waterloo, Ontario, Canada.

Email: jmorales@wlu.ca



responsive to likes and retweets on social media posts. In addition, it is most important to understand how the behavior and attitudes of these individuals are affected by the online environment, as they are the ones who use the platforms most frequently.

We find that our experimental treatment shifted the policy views of active social media users by about 0.12 standard deviations, with the effect further concentrated on the minority of individuals who correctly answered the manipulation check. These results suggest that though only a small share of the population appears to pay conscious attention to likes or retweet metrics, they may be influenced by these social cues. These findings can have further implications for policy decision-making, since social media users are known to be more engaged with politics and can have a disproportionate influence on policy agendas (Barberá et al., 2019; Vaccari et al., 2015; Vaccari & Valeriani, 2021).

Our work is related to a growing literature on the relationship between social media and politics (Zhuravskaya et al., 2020). Social media has been shown to affect electoral outcomes (Fujiwara et al., 2021), legislative processes (Barberá et al., 2019), political knowledge (Munger et al., 2022), and protest participation (Enikolopov et al., 2020; Fergusson & Molina, 2019). However, there is still little understanding of how different features of social media affect behavior.

Studies have emphasized how social media exposes individuals to echo-chambers of predominantly like-minded information (Bakshy et al., 2015; Barberá, 2015; Halberstam & Knight, 2016) and amplifies political polarization (Allcott, Braghieri, et al., 2020; Gorodnichenko et al., 2021; Levy, 2021; Settle, 2018; Sunstein, 2018). Media concerns about the influence of social media on elections are also common,<sup>2</sup> yet many contend that these concerns may be overblown (Allcott & Gentzkow, 2017; Boxell et al., 2017; Eady et al., 2019; Gentzkow & Shapiro, 2011; Guess, 2021; Guess et al., 2020; Scharkow et al., 2020). As a way to sharpen our understanding of these issues, we study one precise mechanism, endorsements, through which social media may affect political dynamics.

Given that social pressure is known to shape behavior and views (Bursztyn & Jensen, 2017; Carlson & Settle, 2016; Cialdini & Goldstein, 2004), online social endorsements can be an important channel through which social media may affect policy preferences, especially in situations of evolving public opinion (Bursztyn et al., 2020; Casoria et al., 2021; Hensel et al., 2022). Furthermore, humans are known to rely on heuristics or mental shortcuts to make judgments and decisions (i.e., bounded rationality), especially in situations where information is scarce and uncertainty is high (Kahneman, 2003; Simon, 1955). In the context of our study, there was substantial uncertainty on how to balance minimizing the spread of COVID-19 while preserving economic activity. An opinion that is highly endorsed would appear to have a higher level of credibility (Luo et al., 2022; Shin et al., 2022), thus potentially making it more persuasive to the

audience—in what is termed the endorsement heuristic (Hilligoss & Rieh, 2008). Relatedly, bandwagon heuristics, whereby individuals are likely to follow what others do can also explain these social dynamics (Banerjee, 1992; Bikhchandani et al., 1992; Metzger et al., 2010; Sundar, 2008). The use of heuristics, and thus the effectiveness of social media endorsements in shaping opinion, is expected to be prevalent in settings involving new issues where individuals are yet to form fixed opinions.

Related work has documented that online social endorsements and perceptions of support affect whether individuals select to read content (Anspach, 2017; Messing & Westwood, 2014), *like* or *retweet* messages (Alatas et al., 2019; Egebark & Ekström, 2018), and self-report voting (Bond et al., 2012, 2017), and can have broader implications for online political dissent (Morales, 2020). We contribute to this work by studying how the perceived endorsements attached to social media messages affect policy attitudes, and we do so in the context of the COVID-19 pandemic.

## Background

Since we present results from studies conducted with nationally representative samples in Ireland, Italy, and the United States, here we briefly describe the social contexts at the time of our experimental intervention and data collection. Our surveys were conducted in July 2020, at the end of the first COVID-19 wave. Overall, there was substantial policy uncertainty and contentious debates regarding the trade-offs between public health and economic activity.

In Ireland, daily confirmed deaths per million had stabilized at 0.26 (Mathieu et al., 2020). The country was at the end of the first period of lockdown, which had a significant negative impact on the economy with unemployment rising up to 28% and gross domestic product (GDP) forecasted to decline by 10.5%.<sup>3</sup> By the time our European data collection started on 8 July, cafés, restaurants, and non-essential retail outlets were allowed to open with social-distancing measures, though there were still restrictions on social gatherings.

Italy, having been severely affected early during the pandemic, was the first country to enact a nationwide lockdown on 9 March 2020. By early July, however, restrictions had gradually been eased and freedom of movement across regions had been restored, as COVID-19 deaths per million had also stabilized at 0.26 (Mathieu et al., 2020). Concerning the economic impact, by this date real GDP was forecasted to fall by over 11% in 2020.<sup>4</sup>

In the United States, lockdown policies varied across states with California being the first to issue a statewide stay-at-home order on 19 March, though by early April, about 90% of the US population were living under stay-at-home orders. By May 2020, the unemployment rate had grown to 14.7%, the highest since the Great Depression. While daily confirmed deaths per million was down to 1.55 in early July,

by the time our US survey was sent out on 31 July, this metric had risen again to 3.27 (Mathieu et al., 2020). The number of confirmed cases had exceeded 3 million and many states postponed re-opening plans as case numbers rose.<sup>5</sup>

## Experimental Design

Our main hypothesis is that the policy attitudes of participants are affected by social media metrics. As users conform to others' preferences, social media affects policy attitudes by informing individuals about others' views. Specifically, we hypothesize:

*H1, Conformity:* Individuals conform to views which appear more popular (as evinced by social media support metrics, that is, likes and retweets).<sup>6</sup>

We also report here our results from investigating additional sources of heterogeneity. First, we are particularly interested in studying whether social conformity effects arising from endorsement metrics are larger for active social media users, who are frequently exposed to, are more likely to pay attention to, and understand the significance of these metrics. Our definition of active social media users consists of those who use Facebook or Twitter for 1 hr or more each day (combined).<sup>7</sup> Second, through the use of a factual manipulation check, we investigate the role of attention in our findings. In particular, if participants did not observe the endorsement metrics, we would not expect an effect to arise (i.e., this serves as a form of placebo check). We pre-registered these dimensions of potential heterogeneity along with a number of other variables, therefore, the results presented should be considered exploratory. These analyses help us understand the mechanisms through which the effects of endorsement metrics arise (or fail to do so).

Finally, we discuss two additional pre-registered hypotheses related to the order in which the messages appear (anchoring), and we present results on one other margin of heterogeneity, pre-treatment attitudes. This last margin of heterogeneity speaks to the literature on social media and political polarization. We find no strong patterns along other margins of heterogeneity; the estimates for all variables specified in our original pre-registration are shown in Appendix Figure A2.<sup>8</sup>

## Implementation and Design

Our first survey was conducted using nationally representative samples in terms of age, gender, and region in Ireland ( $n=305$ ) and Italy ( $n=300$ ), and it was sent out on 8 July 2020. Our second survey was conducted in the United States, using a nationally representative sample in terms of age, gender, and census regions ( $n=1,519$ ), and was sent out on 31 July 2020. Both surveys were programmed in Qualtrics. The

main analyses presented below pool the two surveys, but our main results are quantitatively similar when analyzing the samples separately.<sup>9</sup> Unless otherwise noted, we follow the pre-analysis plans.

To recruit our sample, we contracted Dynata, formerly Research Now SSI, a survey company often used to recruit participants for research in social science (Krupnikov et al., 2021; Snowberg & Yariv, 2021). The company recruits panel members through various marketing channels and collects their sociodemographic information. To obtain a nationally representative sample, survey invitations are sent to potential respondents whose sociodemographic distributions match the one in the latest census data of the country (Bol et al., 2021). Respondents are rewarded for participating in surveys depending on the length and content of the survey. Data quality is ensured by identifying and, after checking, potentially removing random responding, illogical or inconsistent responding, overuse of item non-response (e.g., “don't know”), and speeding (overly quick survey completion).<sup>10</sup>

We first measured participants' pre-treatment policy attitudes using statements about COVID-19 policy responses, for example “The government's highest priority should be saving as many lives as possible even if it means the economy will recover more slowly.”<sup>11</sup> Participants indicated their agreement with these statements on a 1 to 7 Likert-type scale. We standardized these responses and coded positive values as being pro-economy. In addition, we combined the questions into one index through principal components analysis.

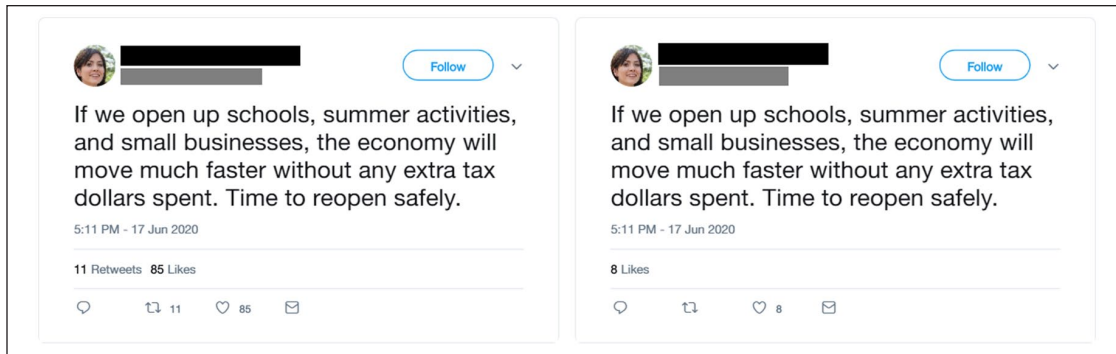
We next randomized participants into one of three treatments: control, pro-economy, or pro-health.<sup>12</sup> In each treatment, participants are shown six tweets from strangers about COVID-19 policies, of which three are pro-economy and three are pro-health.<sup>13</sup> In the control condition, all tweets have low endorsements (a low number of likes and retweets). In the pro-economy condition, the three pro-economy tweets are given high endorsements while the three pro-health tweets are given low endorsements. In the pro-health condition, the three pro-economy tweets are given low endorsements while the three pro-health tweets are given high endorsements.

The tweets were preceded by the following text:

The algorithms used on social media may sometimes present you with posts by complete strangers. You will now be shown 6 tweets. As if you were going through your own social media feed (eg Twitter or Facebook), please consider whether you would “like” or “retweet” each of the following 6 tweets.

Figure 1 shows an example of the experimental variation. The tweets were generated using <https://www.tweetgen.com/> using the following input:

- **Text:** We ran a search of COVID-19-related tweets on Twitter and selected six tweet messages, three pro-locking



**Figure 1.** Example of experimental variation.

Individuals are exposed to the same message with different levels of endorsements, depending on treatment. Left tweet appears more popular than right tweet.

down (which we term “pro-health”) and three pro-opening up (which we term “pro-economy”).

- **Metrics:** “Low” endorsement tweets have between 0 and 10 likes and 0 and 1 retweets. “High” endorsement tweets have between 50 and 100 likes and 10 and 20 retweets. While the metrics chosen for our “High” treatment may seem conservative, these numbers are more realistic for tweets by private persons (rather than famous people).<sup>14</sup>
- **User:** The profile pictures are generated by an algorithm using the website <https://thispersondoesnotexist.com/>. No username is shown.<sup>15</sup>
- **Time:** We randomly picked times and dates in the weeks before data collection.

An additional treatment dimension in the US sample exposed half the participants in each of the three above treatments to an *attention prime* prior to the six tweets. Participants were shown an unrelated tweet followed by three questions about the content, the timing of this tweet, and (importantly) the number of likes. We designed this manipulation to prime participants into paying careful attention to the subsequent six tweets and their endorsements, since absence of treatment effects can potentially be attributable to participants not noticing the metrics. We find that the prime did not reinforce the expected effect and in fact nullified the effect for social media users.<sup>16</sup> However, this treatment also allowed us to assert that the absence of an effect for non-social media users is not due to them not looking at the endorsement metrics.<sup>17</sup> Our analysis focuses on the non-primed group,  $n=779$  (out of 1,519) in the United States, and  $n=605$  in Europe.<sup>18</sup>

After the six tweets, we elicited participants’ post-treatment attitudes using a different set of questions about COVID-19 policy responses. Participants stated their agreement on a 7-point Likert-type scale to a number of policies, such as “Prohibiting gatherings” and “Closing the borders.”<sup>19</sup> We use the first principal component of these responses as an index measure of post-treatment policy attitudes, our main

outcome variable. We again defined positive values as being more pro-economy.

After the post-treatment attitude questions, we conducted a factual manipulation check by asking participants,

Views about COVID-19 policy response can be roughly split into two: (1) Pro-health: prioritise the elimination of COVID-19 over economic activities, for example by extending lockdown measures despite economic costs. (2) Pro-economy: prioritise economic activities over the elimination of COVID-19, for example by opening up the economy despite risks of a second wave. Which of these two views had more likes in the 6 tweets shown earlier?

Participants selected from “pro-health,” “pro-economy,” “neither (both had about the same number of likes),” or “don’t know.” Participants could not go back to the previous screen to check the number of “likes” on the tweets.<sup>20</sup>

Finally, we collected data on education, income, self-reported political ideology on a 0 to 10 left-right scale, party voted in the last election (or if they voted), experience of COVID-19, degree of stubbornness measured by the participants’ resistance to change (Oreg, 2003), media consumption, trust in the media and the government, and the frequency with which they discuss policy issues with family and friends (both on and outside of social media). We measured participants’ social media use by asking about time spent per day on the social media platforms Facebook and Twitter and define active social media users as those who spend more than 1 hr daily on Facebook or Twitter combined.<sup>21</sup>

## Empirical Analysis and Results

### Main Analysis

Key summary statistics are shown in Appendix Table A1, for the whole sample and split by social media use. Notably, active users are younger, hold a more right-wing ideology, and tend to support more pro-economy policies pre-treatment. They were also more likely to correctly answer the

factual manipulation check at the end of the survey. The proportion of active users is highest in Ireland (30%) and lowest in the US Midwest (17%); in all regions, Facebook use is more common than Twitter.

We estimate the effect of social media endorsements on participants' policy attitudes using ordinary least squares (OLS) as follows

$$\begin{aligned} PostAttitudes_i = & \beta_0 + \beta_1 Treatment_i \\ & + PreAttitudes_i \lambda + X_i' \delta + \varepsilon_i \end{aligned} \quad (1)$$

The dependent variable  $PostAttitudes_i$  is the standardized first principal component of the responses to the post-treatment policy questions, with a higher value representing a more pro-economy attitude.  $Treatment_i$  represents the assigned treatment and equals 1 for the pro-economy treatment,  $-1$  for the pro-health treatment, and 0 otherwise. The coefficient of interest is  $\beta_1$ , the effect of perceived endorsements on policy attitudes, which is expected to be positive. Hence, participants exposed to tweets where pro-economy views appear more popular are expected to show an increase in  $PostAttitudes_i$ , while participants exposed to tweets showing popular pro-health views are expected to show a decrease in  $PostAttitudes_i$ .  $PreAttitudes_i$  is taken from the responses to the pre-treatment policy questions.  $X_i'$  is a vector of control variables including gender (coded as a dummy for male), age, region (census regions for the United States, country for the European Union [EU] sample), household income (coded as the log of the mid-point of the interval specified by the subject), education (coded as a dummy for whether the subject has at least a 2 year college degree), and political ideology (self-reported response on a 0–10 left-right scale).<sup>22</sup> We include country fixed effects for all of our pooled analyses below, and we use robust standard errors in all specifications.<sup>23</sup>

We measure pre-treatment attitudes in two ways. First, we use the principal component of all the pre-treatment policy attitude questions (as pre-registered). Second, we use the question that has the highest correlation with the post-treatment attitude index to represent participants' pre-treatment attitudes. In the US sample, the question used is "The government's highest priority should be saving as many lives as possible even if it means the economy will recover more slowly. What do you think of this statement?." In the European sample, the question used is "Sweden's government has so far avoided implementing a lockdown in order to keep the economy going. What do you think of this policy?." Although this latter approach differs from the pre-registered specification, we find that the correlation between pre-treatment and post-treatment attitudes is substantially higher when using this measure, potentially better capturing the policy dimension of interest and thus maximizing the statistical gains from our quasi-pre-post design (Clifford et al., 2021).<sup>24</sup> We present results using both measures.

We estimate Model 1 for the whole sample and for active social media users below. The results are shown in Figure 2.

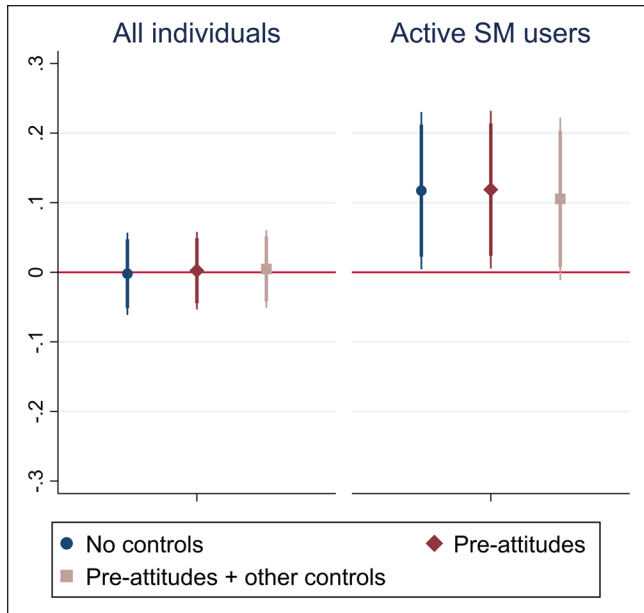
Estimates including a fully interacted model that tests for differences between the groups can be found in Appendix Table A2. We observe no overall treatment effect on participants' policy attitudes. Our estimates suggest a precisely estimated zero effect of endorsements on policy attitudes for our entire sample (left panel). However, we find a differential treatment effect for active social media users (right panel) that suggests these users shift their policy attitudes by about 0.12 standard deviations in response to our treatment. That is, we find evidence consistent with our main hypothesis only for active social media users: endorsement metrics have a significant effect in persuading active users to shift their views in the intended direction.

A factual manipulation check allows us to identify individuals who paid conscious attention to the endorsement counts and to study the extent to which the treatment effects are driven by them (Kane & Barabas, 2019). We asked participants—after they had submitted their policy preferences—about the relative levels of "likes" in the tweets they had seen. Overall, only 33.8% of participants answered the manipulation check correctly, with the rest answering incorrectly or "don't know."<sup>25</sup> The proportion of correct responders is higher in the group of active social media users than passive/non-users (38.7% vs. 32.1%,  $t$  test,  $p = .0225$ ). Importantly, this *post*-treatment attention check is endogenous to the extent to which attention (or correct reporting) of the endorsement metrics may be selective (Montgomery et al., 2018).<sup>26</sup>

To further explore these findings, we split our sample by both social media use and whether they answered this manipulation check correctly. The results are shown in Figure 3. Estimates are also presented in Appendix Table A6, and a fully interacted model that tests for differences between the groups can be found in Appendix Table A7. We observe that the treatment effect is concentrated on social media users who correctly answered the manipulation check. The coefficients are robust to the addition of controls, suggesting that selective attention is unlikely to explain these findings (Oster, 2019). Instead, the results suggest that a relatively small fraction of participants (about 10%) were sensitive to the social cues provided by the engagement metrics in our experiment; the treatment shifted their policy views by about 0.38 standard deviations ( $p$  value  $< .001$ ).

We also observe patterns suggestive of heterogeneous treatment effects for passive/non-social media users; but these are not robust to the addition of controls, revealing instead that these users may pay selective attention to social cues which match their policy attitudes. In particular, passive/non-social media users who correctly answered the manipulation check were more likely to hold views which aligned with the assigned treatment, while those who did not correctly answer the question were more likely to hold views which differed from their assigned treatment.

As further evidence of potential selection in these subsamples, regressing the treatment assignment on *pre-treatment*



**Figure 2.** Treatment effects.

The figure shows the main treatment effects for all users ( $N = 1,384$ ) and active social media users ( $n = 359$ ) separately. Active social media users are defined as individuals who spend more than 1 hr daily on Facebook or Twitter combined. Estimates including a fully interacted model that tests for differences between the groups can be found in Appendix Table A2.

attitudes highlights that correctly answering the manipulation check is potentially endogenous (in Appendix Table A8). The patterns appear particularly stark for passive/non-social media users and suggest that they are more prone to selective attention. To evaluate the extent to which the heterogeneity in Figure 3 may be driven by selective attention, as a further robustness check, we test the sensitivity of our estimates to the addition of controls in a selection on unobservables framework (following Oster, 2019). Our results (shown in Figure 4) corroborate our reading of the results presented here, revealing that the estimates for passive/non-social media users are sensitive to unobservable selection, while those for active social media users are not.<sup>27</sup>

The analyses presented here suggest that the (relatively small) subset of active social media users who tend to pay conscious attention to endorsement metrics are indeed influenced by these social cues. On the contrary, passive/non-social media users are more likely to notice endorsement metrics which reinforce their pre-existing attitudes, but they are on average not influenced by these metrics.

As two additional tests, in Appendix Table A9, we split our sample by countries and show that, though somewhat noisier, the patterns we documented are largely consistent across countries, with the largest effects concentrated on social media users who correctly answered the manipulation check (row 3, columns 4–6). We also find that our results are stronger when excluding the top and bottom 5% respondents in terms of study duration (i.e., those that were “rushing

through” or taking too long to respond, see Appendix Figure A5 and Appendix Table A18). Finally, the results are shown separately for each post-treatment question in Appendix Figure A3, revealing a general shift in attitudes, and that our results are not driven by any particular question.

### Additional Analyses

**Anchoring.** In addition to our main hypothesis on endorsement-driven social conformity, we pre-registered two additional hypotheses. First, the nature of the experiment allows us to evaluate the presence of *anchoring*, the idea that individuals may be disproportionately affected by the views that they *first* see (Furnham & Boo, 2011; Tversky & Kahneman, 1974). We hypothesize that this heuristic extends to social media settings in which individuals are exposed to different views.

*H2, Anchoring:* Individuals are anchored (or primed) by what they are first exposed to, so they tend to conform to the first views they observe.

Second, we explore the *complementarity* between our two hypotheses. In particular, we hypothesize that there are positive complementarities of the two treatments, such that higher endorsement metrics have a differential effect on attitudes when users are exposed to these views first.

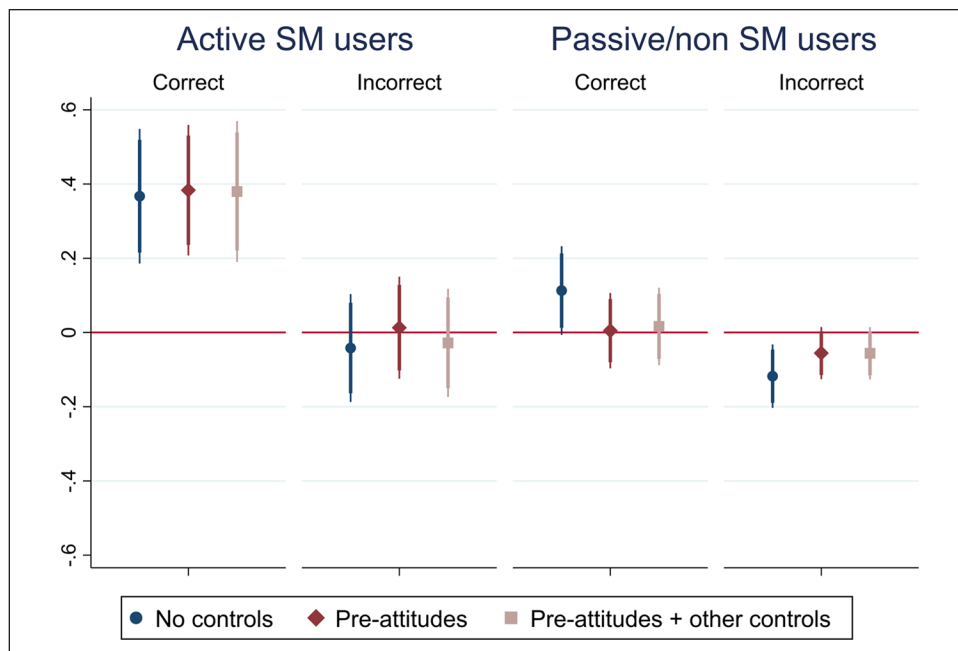
*H3, Complementarities:* Both anchoring and popularity affect individuals’ policy views, and there are positive complementarities between the two: individuals first exposed to popular messages conform most to these views.

The order in which tweets were shown to participants was randomized, which allows us to evaluate our *anchoring* and *complementarities* hypotheses. In particular, this randomization allows us to evaluate whether the *first* post observed affects policy attitudes, by estimating the following model

$$\begin{aligned} PostAttitudes_i = & \beta_0 + \beta_1 FirstMessageEcon_i \\ & + PreAttitudes_i \lambda + X_i' \delta + \varepsilon_i \end{aligned} \quad (2)$$

The indicator variable  $FirstMessageEcon_i$  equals 1 if participants were first exposed to a pro-economy message, and 0 otherwise. Our parameter of interest,  $\beta_1$ , therefore captures the post-treatment attitudes of participants who were first exposed to a pro-economy message, relative to participants who were first exposed to a pro-health message. In other words, both groups are “treated,” and we estimate the differential treatment effect.

In addition, we investigate whether this *anchoring* effect of the first tweet, and our main treatment, are “complementary.” In particular, we estimate this model



**Figure 3.** Treatment effects by social media use and manipulation check.

The figure shows the main treatment effects separately depending on whether individuals correctly responded to the factual manipulation check question. In particular, we split our sample in four groups: active social media users who correctly answered the manipulation check ( $n = 139$ ), active social media users who incorrectly answered the manipulation check ( $n = 220$ ), passive/non-social media users who correctly answered the manipulation check ( $n = 329$ ), and passive/non-social media users who incorrectly answered the manipulation check ( $n = 696$ ). Estimates are also presented in Appendix Table A6, and a fully interacted model that tests for differences between the groups can be found in Appendix Table A7.

$$\begin{aligned}
 PostAttitudes_i = & \beta_0 + \beta_1 FirstMessageEcon_i \\
 & + \beta_2 FirstMessageHigh_i \\
 & + \beta_3 FirstMessageEcon_i \\
 & \times FirstMessageHigh_i \\
 & + PreAttitudes_i \lambda + X_i' \delta + \varepsilon_i
 \end{aligned} \quad (3)$$

The indicator variable  $FirstMessageHigh_i$  equals 1 if the first tweet was in the “high-popularity” category. Our parameter of interest,  $\beta_3$ , captures the differential effect of anchoring when the *first* tweet also had these high social endorsement metrics.

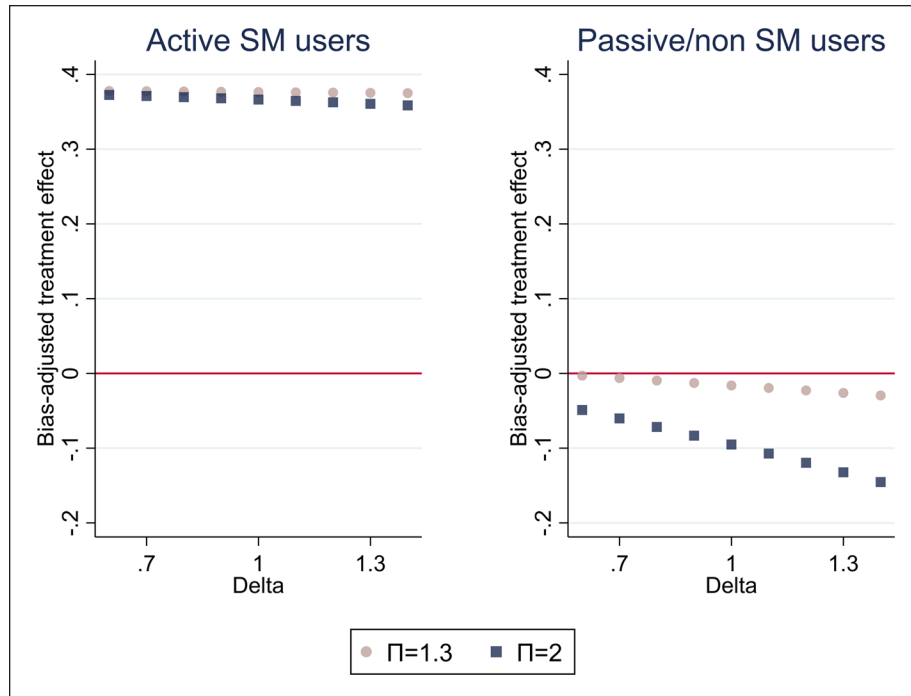
Our results from these analyses are presented in Figure 5. We find weak and marginally significant anchoring effects that appear to be concentrated on passive/non-social media users. These findings on anchoring have therefore unclear policy implications. In particular, potential policy interventions from social media platforms hoping to exploit these results (e.g., by manipulating the top message shown in a news feed) may prove ineffective, since the effect appears to be only present in passive/non-social media users.

In addition, we find no overall complementarity between the content of the first message (pro-econ vs. pro-health) and the engagement metrics, but (conditional on controls) we do find a complementarity for active social media users. This pattern suggests that, for this subset of participants, the content of the first tweet did matter when it had the

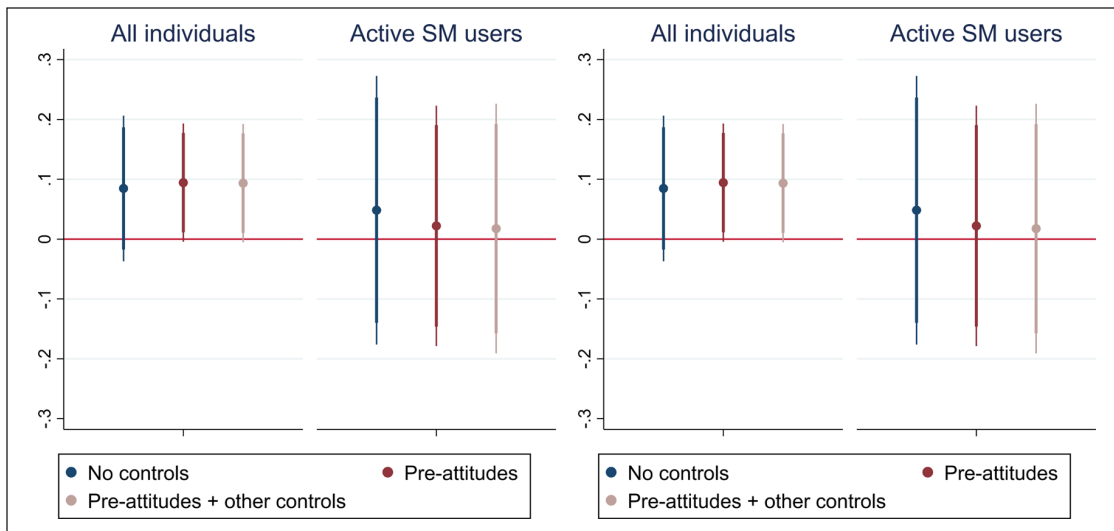
high endorsement metrics. Although these results are more imprecisely estimated, they appear to confirm our previous findings and suggest once more that active social media users can be influenced by engagement metrics.

**Political Polarization.** Social media has commonly been associated with an increase in political polarization (Zhuravskaya et al., 2020). In concordance with these worries, active social media users in our survey were less likely to consider themselves politically moderate (Figure 6) and were less likely to hold (pre-treatment) moderate policy views with respect to COVID-19 (Appendix Figure A4). However, these patterns could well be the result of *selection* into social media use: individuals who hold more polar views tend to be more active on social media, perhaps as an outlet for their extreme opinions. Although recent work documents that deactivating Facebook can indeed reduce individuals’ political polarization (Allcott, Braghieri, et al., 2020), the extent to—and the precise mechanisms through—which social media causes polarization remains debated.

We explored whether our treatment varied with participants’ pre-treatment policy attitudes: Is the effect of endorsements larger for congenial views? This margin of heterogeneity was pre-registered among others, but we do not have enough statistical power to perform multiple hypothesis corrections. For this reason, the results from this section should be viewed as merely exploratory.



**Figure 4.** Selection-bias-adjusted treatment effects for participants with correct manipulation check. Following Oster (2019), the figure shows the estimated bias-adjusted treatment effects for a range of values of  $\delta$  and two values of  $\Pi$  ( $\Pi = 1.3$  is suggested, and  $\Pi = 2$  is conservative). Controls include pre-treatment attitudes (both first principal component and single question) as well as age, gender, region (fixed effects), education, income, and political position.



**Figure 5.** Anchoring effects and complementarities. The figure shows the anchoring effects (left) and the complementarity between the anchoring and the high-popularity metrics effect (right). Estimates are also presented in Appendix Tables A10 and A11, respectively.

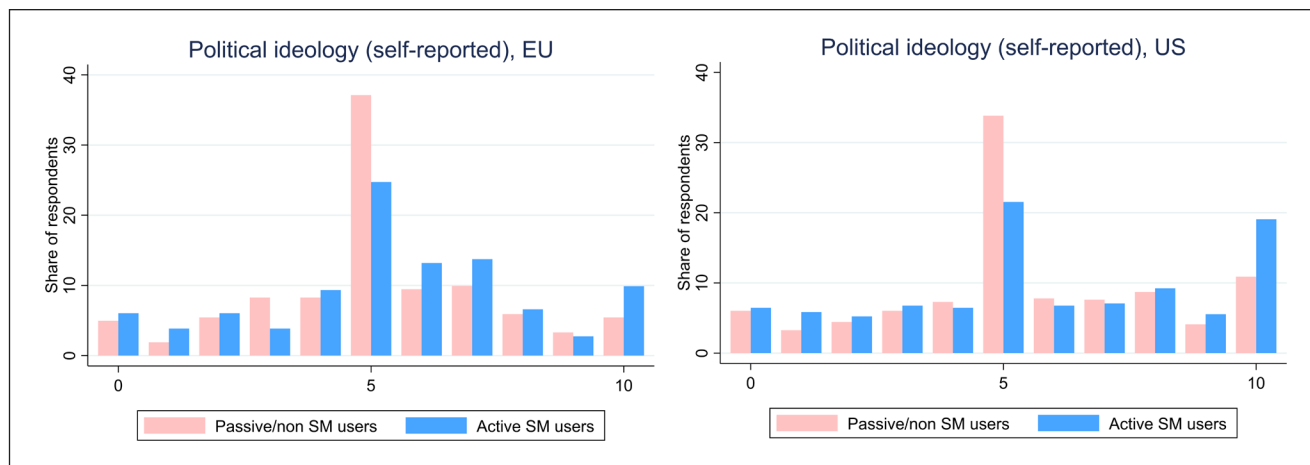
We investigate the heterogeneity of our main treatment by pre-treatment attitudes with the following empirical model

$$PostAttitudes_i = \beta_0 + \beta_1 Treatment_i + \beta_2 PreAttitudes_i + \beta_3 Treatment_i \times PreAttitudes_i + X_i' \delta + \varepsilon_i \quad (4)$$

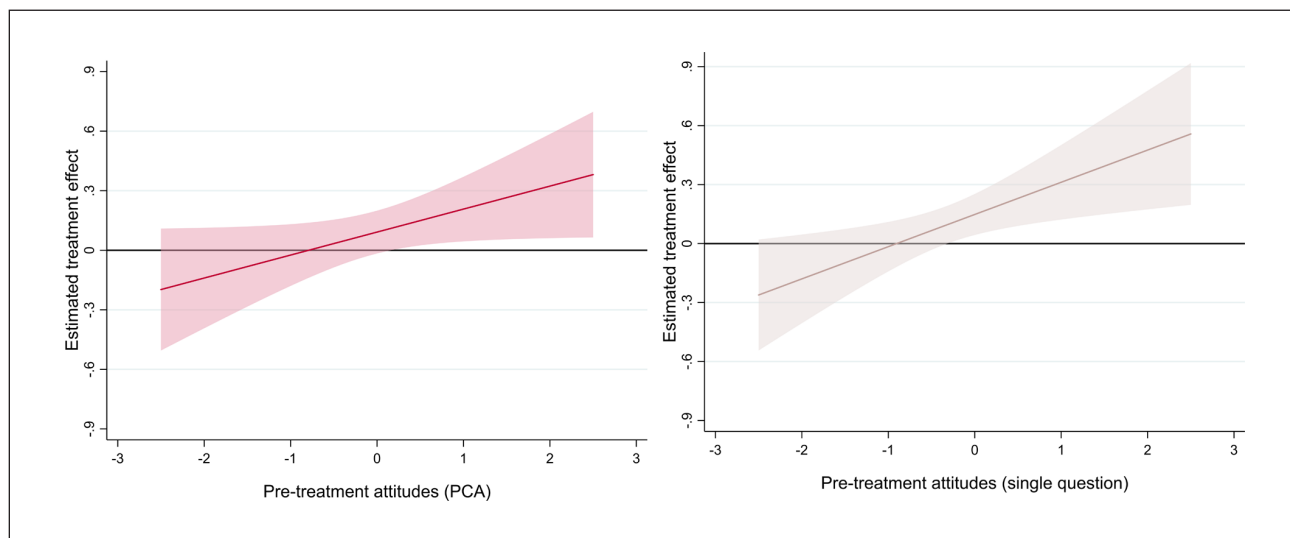
and do so specifically for active social media users.

The estimated marginal treatment effects from estimating Model 4 are shown in Figure 7. We find that our treatment differentially affects social media users depending on their pre-existing attitudes. A one standard deviation increase in pre-treatment attitudes leads to an additional change of between 0.11 and 0.16 standard deviations in policy attitudes (depending on how we measure pre-treatment attitudes). Put





**Figure 6.** Self-reported political ideology and social media use. The figure shows the distribution of responses to the question “In political matters, people talk of ‘the left’ and ‘the right’. How would you place your views on this scale, generally speaking?” separately for active and for passive/non–social media users, and for our European (left,  $n=605$ ) and US (right,  $n=1,519$ ) samples.



**Figure 7.** Heterogeneity by pre-treatment attitudes. The figure shows estimated marginal treatment effects by pre-treatment attitudes. The results are also shown in table form in Appendix Table A12, including additional specifications.

differently, individuals who held more extreme pre-treatment attitudes were more responsive to the treatment, and the treatment reinforced their pre-treatment attitudes. This pattern is particularly driven by individuals who held relatively more pro-economy views before the treatment. We also present models in which we include triple-interactions of treatment, active social media use, and pre-treatment attitudes (Appendix Table A12). The results suggest that public endorsement metrics may be a mechanism through which social media affects individuals’ policy preferences and could also contribute to polarization, and especially since individuals may be more likely to endorse congenial views (Garz et al., 2020).

### Discussion

We hypothesized that online endorsements could affect the formation of policy preferences. Our experiment, in contrast, revealed a precisely estimated null effect of perceived endorsements on our representative samples in the United States and Europe. Furthermore, we found that only about one-third of the experiment participants appeared to pay conscious attention to these engagement metrics. At the same time, we found suggestive evidence of large treatment effects concentrated on a small share of individuals: active social media users who did appear to pay attention to the endorsement metrics (about 10% of all participants). For these


individuals, a higher sensitivity to popularity metrics means that the perceived popularity of (even) strangers' opinion is enough to sway their attitude. Finally, given our conservative measure of "high" metrics, our results potentially underestimate the impact of endorsement for viral tweets or tweets by famous people, which may attract thousands or even more likes and retweets—this would be an interesting avenue for future research.

That the effects appeared concentrated on only a fraction of participants perhaps suggests that the broader effects of endorsement metrics on politics may be limited. However, social media dynamics could further propagate across society in different ways (Margetts et al., 2015; Tufekci, 2017). Social media engagement is also associated with other forms of political engagement; as such, these individuals could exert disproportionate influence in political processes (Barberá et al., 2019; Vaccari et al., 2015; Vaccari & Valeriani, 2021) and have a broad impact on public opinion (Centola et al., 2018). In our survey, active social media users are significantly more likely to (say they) have voted in the previous election. They also report more frequently discussing policy issues with friends or family members both on and outside of social media (Appendix Table A16).<sup>28</sup>

Our micro-level study identifies endorsement metrics as one channel through which social media affects users' policy attitudes. However, there is a trade-off between isolating a precise causal mechanism in a controlled setting versus external validity, and future work should aim to study this relationship in real social media settings. Improved understanding of these effects can inform social media platforms in the design of appropriate interventions to address issues of polarization, misinformation, and foreign influence in politics, among others (since platforms are unlikely to promote account deactivation, as in Allcott, Braghieri, et al., 2020). One further implication is that these social cues may reinforce the effects of selective exposure (as emphasized in Zhuravskaya et al., 2020). If individuals with more extreme preferences are also more likely to "like" content, then not only will social media algorithms expose users to more polarized opinions (Levy, 2021), but such content may also *appear* to have broader support. In particular, our results may underestimate the true effect of endorsement metrics on social media platforms where individuals are exposed to posts and endorsements by people they know and whom they choose to follow—whose opinions the individual likely agrees with—which, as our exploratory analysis suggests, may further contribute to political polarization. How these different features of social media interact to influence political views and the persistence of the effects remain an important avenue for future work. Finally, substantial uncertainty surrounding the topic of COVID-19 in the early stages of the pandemic likely made policy attitudes in this respect highly malleable. Future work should also

examine whether endorsements can shape policy attitudes of social media users in deeper entrenched topics for which views are likely to be more rigid.

### Acknowledgements

We thank Anna Dreber Almenberg, Karen Arulsamy, Stefan Müller, and Johannes Wohlfart as well as seminar participants at UCD and Collegio Carlo Alberto for comments. The experiments were approved by University College Dublin Office of Research Ethics (reference numbers: HS-E-20-110-Samahita and HS-E-20-134-Samahita). Authors' order has been randomized using the AEA Author Randomization Tool (reference: eTjBUATa\_zKY), denoted by . All errors are our own.

### Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funding from UCD, Collegio Carlo Alberto, and the Einaudi Institute for Economics and Finance is gratefully acknowledged.

### ORCID iD

Margaret Samahita  <https://orcid.org/0000-0002-8693-1185>

### Supplemental material

Supplemental material for this article is available at [https://osf.io/3xkaw/?view\\_only=0a7f56532cb6440fb0a8d9e960e9f6a6](https://osf.io/3xkaw/?view_only=0a7f56532cb6440fb0a8d9e960e9f6a6).

### Notes

1. Pre-registration is available at AEARCTR-0006254 and <https://aspredicted.org/blind.php?x=5t367e>.
2. See, for instance, <https://www.nytimes.com/2020/03/29/technology/russia-troll-farm-election.html> and <https://www.scientificamerican.com/article/how-twitter-bots-help-fuel-political-feuds/>
3. See the Government of Ireland's 2020 "July Jobs Stimulus," <https://assets.gov.ie/81556/d4fa4cc4-7e9f-4431-8540-a9ecb7126505.pdf>, accessed 3 February 2023.
4. See the European Commission's macroeconomic forecast for Italy, Summer 2020: [https://ec.europa.eu/economy\\_finance/forecasts/2020/summer/ecfin\\_forecast\\_summer\\_2020\\_it\\_en.pdf](https://ec.europa.eu/economy_finance/forecasts/2020/summer/ecfin_forecast_summer_2020_it_en.pdf), accessed 4 February 2023.
5. See <https://www.cdc.gov/museum/timeline/covid19.html>, accessed 4 February 2023.
6. Banerjee et al. (2020) document the presence of community spillovers on behavior from a randomized controlled trial delivering YouTube COVID-19 information messages; Alatas et al. (2019) show that celebrity endorsed tweets about (non-COVID-related) immunization received higher engagement and had some effects on health knowledge, suggesting that social dynamics play an important role in the formation of health attitudes and beliefs; and Ho et al. (2022) document that celebrity endorsements promote pro-environmental behavior.

7. Social media use was pre-registered in our first experiment as one of many heterogeneity dimensions to explore. After finding significant effects in our first experiment, it was pre-registered as the most important dimension of interest in our second experiment.
8. The complete set of results, including separate analyses for each survey wave, are provided in the supplementary materials at [https://osf.io/3xkaw/?view\\_only=0a7f56532cb6440fb0a8d9e960e9f6a6](https://osf.io/3xkaw/?view_only=0a7f56532cb6440fb0a8d9e960e9f6a6).
9. We show the main results separated by country in Appendix Table A5, and in the supplementary materials, we show the full analyses for each survey.
10. See [http://sigs.researchnow.com/EU\\_Emails/UK/14Apr/Panel%20IE%20Landing%20Page/ESOMAR\\_28\\_IE.pdf](http://sigs.researchnow.com/EU_Emails/UK/14Apr/Panel%20IE%20Landing%20Page/ESOMAR_28_IE.pdf) for more details.
11. The EU survey included two pre-treatment statements, and the US survey included five statements. See Appendix section “COVID-19 Policy Questions” for the full list of pre-treatment questions.
12. Our European sample was only exposed to the pro-health and pro-economy treatments.
13. Users are more likely to encounter posts by strangers on Twitter rather than Facebook, where posts shown in the news-feed are more likely to have come from someone known to the user (Oz et al., 2018). The COVID-19 messages were pre-classified by us and are shown in Appendix Figure A1.
14. A limitation of our study is that we are not able to disentangle the effect of “likes” and “retweets.” While in our study a higher number of retweets is used to indicate higher popularity (retweets increase with likes), future work could study how different types of metrics influence the perception of the user.
15. We did not randomize the gender in the profile pictures, the same text is always assigned to the same profile picture (two males and one female for the pro-economy tweets, and two females and one male for the pro-health tweets). However, since every respondent sees the same set of tweets and profile pictures, our effect is not driven by the gender assignment of the tweets.
16. Perhaps due to these participants realizing that the metrics were manipulated (shown in Appendix Table A14).
17. Passive/non-social media users in the primed group were more likely to correctly answer a manipulation check post-treatment (Appendix Table A15).
18. All pre-registered analyses of the primed group are shown in supplementary materials. Including this group in our analysis somewhat weakens our results but the main effect remains statistically significant.
19. The EU sample included seven post-treatment policies, and the US sample included eight. See Appendix section “COVID-19 Policy Questions” for the full list of policies asked about.
20. Although the question was not incentivized, we do not see strong reasons for participants to misreport (consistent with findings in Allcott, Boxell, et al., 2020).
21. Our results are robust to defining active social media users as those who use Facebook for at least 30 min a day as shown in Appendix Table A13. In addition, for the US sample, we asked participants to measure their own level of social media activity on a scale from 0 to 100. Our main result also holds when using this alternative measure of social media use. In particular, the effect of endorsements is concentrated on those who report the highest levels of activity.
22. Controlling for COVID-19 case numbers and a stringency index of government response yields similar results.
23. Our specification differs from the pre-analysis plan in two ways. First, instead of including two separate treatment coefficients and estimating their pooled average effect, we pool our treatments by defining a negative treatment for the pro-health group, a specification which is statistically equivalent to the one pre-registered which uses a dummy for each treatment (as shown in Appendix Tables A3 and A4 for the subsample who answered the manipulation check correctly) while being easier to interpret and implement. Second, because we are pooling our studies, we define region as the respondent’s country for the Irish and Italian samples (when region fixed effects are included as an additional control, the country fixed effects are absorbed), and we use the self-reported political ideology instead of the party affiliation. The analysis is robust to using a “right-wing” dummy, which equals 1 for US participants voting Republican and Italian participants voting Lega Nord in the last election, and 0 for all others (including Irish participants since there is no Irish right-wing party).
24. This “highest correlation” question is the same in treatment and control groups.
25. Note that since there were four alternatives, the correct answer would be chosen by 25% of respondents if they were selecting their answer randomly.
26. See Iyengar et al. (2008) and Wang et al. (2014) for evidence of motivated selective attention. Another possible explanation for incorrect reporting is consensus bias (Ross et al., 1977): respondents guess the answer in the direction of their own attitudes, thus failing the manipulation check if the treatment is not aligned with their views.
27. Interestingly, correctly answering the manipulation check question is not significantly correlated with baseline demographic variables (though there are significant differences across countries, Appendix Table A17).
28. These patterns are in line with findings in Guess (2021) of homogeneously partisan information consumption among only a minority of US citizens, but who nonetheless were on average more likely to vote.

## References

- Alatas, V., Chandrasekhar, A. G., Mobius, M., Olken, B. A., & Paladines, C. (2019). *When celebrities speak: A nationwide Twitter experiment promoting vaccination in Indonesia* [NBER working paper no. 25589]. [https://www.nber.org/system/files/working\\_papers/w25589/w25589.pdf](https://www.nber.org/system/files/working_papers/w25589/w25589.pdf)
- Allcott, H., Boxell, L., Conway, J., Gentzkow, M., Thaler, M., & Yang, D. Y. (2020). Polarization and public health: Partisan differences in social distancing during the Coronavirus pandemic. *Journal of Public Economics*, 191, 104254.
- Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, 110(3), 629–676.
- Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), 211–236.

- Anspach, N. M. (2017). The new personal influence: How our Facebook friends influence the news we read. *Political Communication*, 34(4), 590–606.
- Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239), 1130–1132.
- Banerjee, A., Alsan, M., Breza, E., Chandrasekhar, A. G., Chowdhury, A., Duflo, E., Goldsmith-Pinkham, P., & Olken, B. A. (2020). *Messages on COVID-19 prevention in India increased symptoms reporting and adherence to preventive behaviors among 25 million recipients with similar effects on non-recipient members of their communities* [NBER working paper no. 27496]. <https://web.stanford.edu/~arungc/BABCCDGO.pdf>
- Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3), 797–817.
- Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, 23(1), 76–91.
- Barberá, P., Casas, A., Nagler, J., Egan, P. J., Bonneau, R., Jost, J. T., & Tucker, J. A. (2019). Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review*, 113(4), 883–901.
- Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5), 992–1026.
- Bol, D., Giani, M., Blais, A., & Loewen, P. J. (2021). The effect of COVID-19 lockdowns on political support: Some good news for democracy? *European Journal of Political Research*, 60(2), 497–505.
- Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415), 295–298.
- Bond, R. M., Settle, J. E., Fariss, C. J., Jones, J. J., & Fowler, J. H. (2017). Social endorsement cues and political participation. *Political Communication*, 34(2), 261–281.
- Boxell, L., Gentzkow, M., & Shapiro, J. M. (2017). Greater internet use is not associated with faster growth in political polarization among US demographic groups. *Proceedings of the National Academy of Sciences*, 114(40), 10612–10617.
- Bursztyn, L., Egorov, G., & Fiorin, S. (2020). From extreme to mainstream: The erosion of social norms. *American Economic Review*, 110(11), 3522–3548.
- Bursztyn, L., & Jensen, R. (2017). Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure. *Annual Review of Economics*, 9, 131–153.
- Carlson, T. N., & Settle, J. E. (2016). Political chameleons: An exploration of conformity in political discussions. *Political Behavior*, 38(4), 817–859.
- Casoria, F., Galeotti, F., & Villeval, M. C. (2021). Perceived social norm and behavior quickly adjusted to legal changes during the COVID-19 pandemic. *Journal of Economic Behavior & Organization*, 190, 54–65.
- Centola, D., Becker, J., Brackbill, D., & Baronchelli, A. (2018). Experimental evidence for tipping points in social convention. *Science*, 360(6393), 1116–1119.
- Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, 55, 591–621.
- Clifford, S., Sheagley, G., & Piston, S. (2021). Increasing precision without altering treatment effects: Repeated measures designs in survey experiments. *American Political Science Review*, 115(3), 1048–1065.
- Eady, G., Nagler, J., Guess, A., Zilinsky, J., & Tucker, J. A. (2019). How many people live in political bubbles on social media? Evidence from linked survey and Twitter data. *SAGE Open*, 9(1), 2158244019832705.
- Egebark, J., & Ekström, M. (2018). Liking what others “Like”: Using Facebook to identify determinants of conformity. *Experimental Economics*, 21(4), 793–814.
- Enikolopov, R., Makarin, A., & Petrova, M. (2020). Social media and protest participation: Evidence from Russia. *Econometrica*, 88(4), 1479–1514.
- Fergusson, L., & Molina, C. (2019). *Facebook causes protests* [Documento CEDE no. 41]. <https://thedocs.worldbank.org/en/doc/f73fcca90d36718b63430887535b1eb6-0050022021/original/Facebook-Causes-Protests.pdf>
- Fujiwara, T., Müller, K., & Schwarz, C. (2021). *The effect of social media on elections: Evidence from the United States* [NBER working paper no. 28849]. <https://www.princeton.edu/~fujiwara/papers/SocialMediaAndElections.pdf>
- Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40(1), 35–42.
- Garz, M., Sörensen, J., & Stone, D. F. (2020). Partisan selective engagement: Evidence from Facebook. *Journal of Economic Behavior & Organization*, 177, 91–108.
- Gentzkow, M., & Shapiro, J. M. (2011). Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4), 1799–1839.
- Gorodnichenko, Y., Pham, T., & Talavera, O. (2021). Social media, sentiment and public opinions: Evidence from #Brexit and #USElection. *European Economic Review*, 136, 103772.
- Guess, A. M. (2021). (Almost) everything in moderation: New evidence on Americans’ online media diets. *American Journal of Political Science*, 65(4), 1007–1022.
- Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behavior*, 4(5), 472–480.
- Halberstam, Y., & Knight, B. (2016). Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of Public Economics*, 143, 73–88.
- Hensel, L., Witte, M., Caria, A. S., Fetzer, T., Fiorin, S., Götz, F. M., Gomez, M., Haushofer, J., Ivchenko, A., & Kraft-Todd, G. (2022). Global behaviors, perceptions, and the emergence of social norms at the onset of the COVID-19 pandemic. *Journal of Economic Behavior & Organization*, 193, 473–496.
- Hilligoss, B., & Rieh, S. Y. (2008). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing & Management*, 44(4), 1467–1484.
- Ho, T. Q., Nie, Z., Alpizar, F., Carlsson, F., & Nam, P. K. (2022). Celebrity endorsement in promoting pro-environmental behavior. *Journal of Economic Behavior & Organization*, 198, 68–86.
- Iyengar, S., Hahn, K. S., Krosnick, J. A., & Walker, J. (2008). Selective exposure to campaign communication: The role of anticipated agreement and issue public membership. *The Journal of Politics*, 70(1), 186–200.
- Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58(9), 697–720.

- Kane, J. V., & Barabas, J. (2019). No harm in checking: Using factual manipulation checks to assess attentiveness in experiments. *American Journal of Political Science*, 63(1), 234–249.
- Krupnikov, Y., Nam, H. H., & Style, H. (2021). Convenience samples in political science experiments. In J. N. Druckman & D. P. Green (Eds.), *Advances in experimental political science* (pp. 165–183). Cambridge University Press.
- Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, 111(3), 831–870.
- Luo, M., Hancock, J. T., & Markowitz, D. M. (2022). Credibility perceptions and detection accuracy of fake news headlines on social media: Effects of truth-bias and endorsement cues. *Communication Research*, 49(2), 171–195.
- Margetts, H., John, P., Hale, S., & Yasseri, T. (2015). *Political turbulence: How social media shape collective action*. Princeton University Press.
- Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Giattino, C., Hasell, J., Macdonald, B., Dattani, S., Beltekian, D., Ortiz-Ospina, E., & Roser, M. (2020). Coronavirus pandemic (COVID-19). *Our World in Data*. <https://ourworldindata.org/coronavirus>
- Messing, S., & Westwood, S. J. (2014). Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication Research*, 41(8), 1042–1063.
- Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of Communication*, 60(3), 413–439.
- Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, 62(3), 760–775.
- Morales, J. S. (2020). Perceived popularity and online political dissent: Evidence from Twitter in Venezuela. *The International Journal of Press/Politics*, 25(1), 5–27.
- Munger, K., Egan, P. J., Nagler, J., Ronen, J., & Tucker, J. (2022). Political knowledge and misinformation in the era of social media: Evidence from the 2015 UK election. *British Journal of Political Science*, 52(1), 107–127.
- Oreg, S. (2003). Resistance to change: Developing an individual differences measure. *Journal of Applied Psychology*, 88(4), 680.
- Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2), 187–204.
- Oz, M., Zheng, P., & Chen, G. M. (2018). Twitter versus Facebook: Comparing incivility, impoliteness, and deliberative attributes. *New Media & Society*, 20(9), 3400–3419.
- Ross, L., Greene, D., & House, P. (1977). The “false consensus effect”: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13(3), 279–301.
- Scharkow, M., Mangold, F., Stier, S., & Breuer, J. (2020). How social network sites and other online intermediaries increase exposure to news. *Proceedings of the National Academy of Sciences*, 117(6), 2761–2763.
- Settle, S., & Shupe, C. (2022). Lives or livelihoods? Perceived trade-offs and policy views. *The Economic Journal*, 132(643), 1150–1178.
- Settle, J. E. (2018). *Frenemies: How social media polarizes America*. Cambridge University Press.
- Shin, I., Wang, L., & Lu, Y.-T. (2022). Twitter and endorsed (fake) news: The influence of endorsement by strong ties, celebrities, and a user majority on credibility of fake news during the COVID-19 pandemic. *International Journal of Communication*, 16, 2573–2595.
- Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1), 99–118.
- Snowberg, E., & Yariv, L. (2021). Testing the waters: Behavior across participant pools. *American Economic Review*, 111(2), 687–719.
- Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. Metzger & A. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 73–100). MIT Press.
- Sunstein, C. R. (2018). *#Republic: Divided democracy in the age of social media*. Princeton University Press.
- Tufekci, Z. (2017). *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131.
- Vaccari, C., & Valeriani, A. (2021). *Outside the bubble: Social media and political participation in Western democracies*. Oxford University Press.
- Vaccari, C., Valeriani, A., Barberá, P., Bonneau, R., Jost, J. T., Nagler, J., & Tucker, J. A. (2015). Political expression and action on social media: Exploring the relationship between lower- and higher-threshold political activities among Twitter users in Italy. *Journal of Computer-Mediated Communication*, 20(2), 221–239.
- Wang, Z., Morey, A. C., & Srivastava, J. (2014). Motivated selective attention during political ad processing: The dynamic interplay between emotional ad content and candidate evaluation. *Communication Research*, 41(1), 119–156.
- Zhuravskaya, E., Petrova, M., & Enikolopov, R. (2020). Political effects of the internet and social media. *Annual Review of Economics*, 12, 415–438.

### Author Biographies

**Pierluigi Conzo** (PhD, University of Rome “Tor Vergata”) is an Associate Professor of Economics at the University of Turin and a Research Fellow of the Collegio Carlo Alberto. His recent research interests include the analysis of the determinants (and the outcomes) of prosocial preferences as well as the socio-political effects of anti-immigration campaigns.

**Laura K. Taylor** (PhD/MA, University of Notre Dame) is an Associate Professor of Psychology at University College Dublin. Her research interests include constructive intergroup relations, particularly in conflict-affected settings.

**Juan S. Morales** (PhD, University of Toronto) is an Assistant Professor of Economics at Wilfrid Laurier University. His research interests include political economy, development economics, applied microeconomics, conflict, and media.

**Margaret Samahita** (PhD, Lund University) is an Assistant Professor of Economics at University College Dublin. Her research interests include behavioral economics, experimental economics, and political economy.

**Andrea Gallice** (PhD, European University Institute) is Associate Professor of Economics at the University of Turin and a Research Fellow of the Collegio Carlo Alberto. His research interests include political economy and the analysis of social status concerns and social pressures.

## Appendix

### COVID-19 Policy Questions

Before the treatment, subjects answer the following questions on a 1 to 7 Likert-type scale:

- What do you think of [US: the federal government's response measures]/[EU: your national government's lockdown measures] in reaction to the COVID-19 pandemic? (*extremely insufficient—extreme overreaction*)
- \*What do you think of your state government's response measures in reaction to the COVID-19 pandemic? (*extremely insufficient—extreme overreaction*)
- Sweden's government has so far avoided implementing a lockdown in order to keep the economy going. What do you think of this policy? (*strongly disagree—strongly agree*)
- \*The government's highest priority should be saving as many lives as possible even if it means the economy will recover more slowly. What do you think of this statement? (*strongly disagree—strongly agree, reverse-coded*)

- \*It is becoming more important for the government to save jobs and restart the economy than to take every precaution to keep people safe. What do you think of this statement? (*strongly disagree—strongly agree*)

After the treatment, participants stated their agreement on a 7-point Likert-type scale to these policies:

- Closing the borders
- Prohibiting gatherings
- Prohibiting non-essential travels
- Closing daycares, schools, colleges, and universities
- Closing non-essential businesses (bars, stores that are not food or health related, etc.)
- Handing out USD/EUR 1,000 fines to those who do not comply with social-distancing rules
- General lockdown of the population with a ban on leaving the home (except for medical reasons)
- \*Mandatory use of face-coverings in public places
- \*Note that these questions were only asked for the US sample.

**Table A1.** Summary Statistics.

	All individuals		Active SM users		Passive/non-SM users		Difference
	M	SD	M	SD	M	SD	
Age	45.105	16.089	39.916	14.841	46.922	16.120	7.006***
Male	0.467	0.499	0.493	0.501	0.458	0.498	-0.035
Education	0.626	0.484	0.602	0.490	0.635	0.482	0.033
Income	6.700	2.670	6.806	2.411	6.663	2.756	-0.144
Political ideology	5.316	2.632	5.680	2.856	5.189	2.539	-0.490***
Ireland	0.220	0.415	0.256	0.437	0.208	0.406	-0.048*
Italy	0.217	0.412	0.251	0.434	0.205	0.404	-0.046*
US West	0.212	0.409	0.271	0.446	0.194	0.396	-0.077**
US Midwest	0.231	0.422	0.169	0.376	0.249	0.433	0.080**
US Northeast	0.175	0.380	0.169	0.376	0.176	0.381	0.007
US South	0.215	0.411	0.192	0.395	0.223	0.417	0.031
Manipulation check (correct)	0.338	0.473	0.387	0.488	0.321	0.467	-0.066**
Duration (min)	14.156	76.402	11.932	16.477	14.935	88.241	3.002
Active SM users	0.259	0.438	1.000	0.000	0.000	0.000	-1.000
Facebook use	1.256	1.043	2.532	0.727	0.809	0.717	-1.723***
Twitter use	0.719	0.960	1.663	1.151	0.388	0.599	-1.275***
Observations	1,384		359		1,025		1,384

SM: social media.

Summary statistics of age, gender (coded as a dummy for male), education (dummy for having at least a 2-year college degree), household income (log of the midpoint of the interval specified by the subject), political ideology (self-reported response on a 0–10 left-right scale), region (dummy for country for EU, census region for the US), ActiveSMuser (dummy for spending more than 1 hr a day on Facebook or Twitter combined), Facebook use (daily time spent on Facebook: 0 = never/no account, 1 = less than 30 min, 2 = from 30 min to 1 hr, 3 = more than 1 hr), Twitter use (daily time spent on Twitter: 0 = never/no account, 1 = less than 30 min, 2 = from 30 min to 1 hr, 3 = more than 1 hr).

Significance levels indicated \* $p < .10$ , \*\* $p < .05$ , \*\*\* $p < .01$ .

**Table A2.** Main Treatment Effects.

	All individuals								Active SM users			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treatment	-0.002 (0.030)	0.002 (0.028)	0.005 (0.028)	0.010 (0.026)	-0.039 (0.035)	-0.032 (0.031)	-0.025 (0.031)	-0.036 (0.028)	0.117** (0.058)	0.119** (0.058)	0.105* (0.059)	0.148*** (0.056)
Treatment × Active SM user					0.161** (0.067)	0.162** (0.067)	0.145** (0.068)	0.195*** (0.063)				
<i>n</i>	1,384	1,384	1,384	1,384	1,384	1,384	1,384	1,384	359	359	359	359
<i>R</i> <sup>2</sup>	.003	.115	.126	.285	.019	.143	.156	.303	.027	.044	.075	.177
Pre-attitudes (PCA)	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Other controls	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes
Pre-attitudes (Single Q)	No	No	No	Yes	No	No	No	Yes	No	No	No	Yes

SM: social media; PCA: principal components analysis.

Ordinary least square estimates using post-treatment attitudes index (first principal component of the responses to the post-treatment policy questions) as outcome. The treatment variable equals 1 for the pro-economy treatment, -1 for the pro-health treatment, and 0 otherwise. Pre-attitude controls include the first principal component of the pre-treatment policy questions and the single question with the highest correlation with post-treatment attitudes. Columns 5 to 8 present estimates of models of this form

$$PostAttitudes_i = \beta_0 + \beta_1 Treatment_i + \beta_2 ActiveSMuser_i + \beta_3 Treatment_i \times ActiveSMuser_i + X_i' \delta + \varepsilon_i$$

ActiveSMuser is a dummy variable which equals 1 if the subject spends more than 1 hr a day on Facebook or Twitter (combined) and zero otherwise. Controls include age, gender, region (fixed effects), education, income, and political position. All specifications include country fixed effects. Robust standard errors in parentheses.

Significance levels indicated \**p* < .10, \*\**p* < .05, \*\*\**p* < .01.

**Table A3.** Main Effects for Pro-Economy and Pro-Health Treatments Separately.

	All individuals				Active SM users			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment Econ	-0.007 (0.085)	0.002 (0.079)	-0.001 (0.079)	0.047 (0.067)	0.172 (0.168)	0.159 (0.166)	0.142 (0.165)	0.233 (0.150)
Treatment Health	-0.003 (0.084)	-0.002 (0.079)	-0.010 (0.079)	0.026 (0.067)	-0.062 (0.159)	-0.079 (0.157)	-0.069 (0.155)	-0.063 (0.142)
TE: (β <sub>1</sub> - β <sub>2</sub> ) / 2	-0.002 (0.030)	0.002 (0.028)	0.005 (0.028)	0.011 (0.026)	0.117** (0.058)	0.119** (0.058)	0.105* (0.060)	0.148*** (0.056)
<i>n</i>	1,384	1,384	1,384	1,384	359	359	359	359
<i>R</i> <sup>2</sup>	.003	.115	.126	.285	.027	.044	.074	.178
Pre-attitudes (PCA)	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Other controls	No	No	Yes	Yes	No	No	Yes	Yes
Pre-attitudes (Single Q)	No	No	No	Yes	No	No	No	Yes

SM: social media; PCA: principal components analysis.

Ordinary least square regressions with the post-treatment attitudes index (first principal component of the responses to the post-treatment policy questions) as outcome. Treatment Econ (Health) equals 1 for the pro-economy (pro-health) treatment and 0 otherwise. TE equals the average treatment effect of the pro-economy and pro-health treatments, calculated as (β<sub>1</sub> - β<sub>2</sub>) / 2. Pre-attitude controls include the first principal component of the pre-treatment policy questions and the single question with the highest correlation with post-treatment attitudes. Controls include age, gender, region (fixed effects), education, income, and political position. All specifications include country fixed effects. Robust standard errors in parentheses.

Significance levels indicated \**p* < .10, \*\**p* < .05, \*\*\**p* < .01.

**Table A4.** Main Effects for Subsample With Correct Manipulation Check.

	All individuals				Active SM users			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment Econ	0.124 (0.161)	0.066 (0.151)	0.073 (0.148)	0.095 (0.140)	0.199 (0.286)	0.180 (0.281)	0.247 (0.281)	0.284 (0.262)
Treatment Health	-0.249* (0.147)	-0.249* (0.141)	-0.247* (0.137)	-0.159 (0.129)	-0.526** (0.241)	-0.545** (0.233)	-0.470** (0.235)	-0.473** (0.216)
TE: $(\beta_1 - \beta_2) / 2$	0.187*** (0.052)	0.157*** (0.050)	0.160*** (0.050)	0.127*** (0.049)	0.363*** (0.094)	0.363*** (0.095)	0.359*** (0.099)	0.378*** (0.097)
SE								
n	468	468	468	468	139	139	139	139
R <sup>2</sup>	.033	.134	.154	.249	.131	.144	.208	.274
Pre-attitudes (PCA)	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Other controls	No	No	Yes	Yes	No	No	Yes	Yes
Pre-attitudes (Single Q)	No	No	No	Yes	No	No	No	Yes

SM: social media; PCA: principal components analysis.

Ordinary least square regressions with the post-treatment attitudes index (first principal component of the responses to the post-treatment policy questions) as outcome. Treatment Econ (Health) equals 1 for the pro-economy (pro-health) treatment and 0 otherwise. TE equals the average treatment effect of the pro-economy and pro-health treatments, calculated as  $(\beta_1 - \beta_2) / 2$ . Pre-attitude controls include the first principal component of the pre-treatment policy questions and the single question with the highest correlation with post-treatment attitudes. Controls include age, gender, region (fixed effects), education, income, and political position. All specifications include country fixed effects. Robust standard errors in parentheses. Significance levels indicated \* $p < .10$ , \*\* $p < .05$ , \*\*\* $p < .01$ .

**Table A5.** Heterogeneity by SM Activity.

	All individuals				Correct manipulation check			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment Econ	-0.069 (0.095)	-0.047 (0.086)	-0.045 (0.085)	0.008 (0.075)	0.062 (0.187)	0.009 (0.176)	0.018 (0.170)	0.080 (0.168)
Treatment Health	0.010 (0.095)	0.017 (0.087)	0.005 (0.086)	0.080 (0.076)	-0.158 (0.178)	-0.133 (0.170)	-0.139 (0.165)	0.015 (0.163)
Treatment Econ × Active SM	0.310* (0.176)	0.285* (0.167)	0.263 (0.170)	0.188 (0.150)	0.248 (0.285)	0.228 (0.266)	0.233 (0.255)	0.076 (0.241)
Treatment Health × Active SM	-0.009 (0.175)	-0.037 (0.170)	-0.025 (0.173)	-0.202 (0.156)	-0.288 (0.278)	-0.377 (0.267)	-0.350 (0.255)	-0.585** (0.244)
TE: $(\beta_3 - \beta_4) / 2$	0.159** (0.067)	0.161** (0.067)	0.144** (0.068)	0.195*** (0.063)	0.268** (0.105)	0.303*** (0.107)	0.291*** (0.109)	0.330*** (0.103)
SE								
n	1,384	1,384	1,384	1,384	468	468	468	468
R <sup>2</sup>	.020	.144	.156	.303	.052	.161	.187	.278
Pre-attitudes (PCA)	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Other controls	No	No	Yes	Yes	No	No	Yes	Yes
Pre-attitudes (Single Q)	No	No	No	Yes	No	No	No	Yes

SM: social media; PCA: principal components analysis.

Ordinary least square regressions with the post-treatment attitudes index (first principal component of the responses to the post-treatment policy questions) as outcome. The model estimated is

$$PostAttitudes_i = \beta_0 + \beta_1 TreatmentEcon_i + \beta_2 TreatmentHealth_i + \beta_3 TreatmentEcon_i \times ActiveSMuser_i + \beta_4 TreatmentHealth_i \times ActiveSMuser_i + \beta_5 ActiveSMuser_i + X_i'\delta + \varepsilon_i$$

Treatment Econ (Health) equals 1 for the pro-economy (pro-health) treatment and 0 otherwise. ActiveSMuser is a dummy variable which equals 1 if the subject spends more than 1 hr a day on Facebook or Twitter (combined) and zero otherwise. TE equals the average treatment effect of the pro-economy and pro-health treatments for active SM users, calculated as  $(\beta_3 - \beta_4) / 2$ . Pre-attitude controls include the first principal component of the pre-treatment policy questions and the single question with the highest correlation with post-treatment attitudes. Controls include age, gender, region (fixed effects), education, income, and political position. All specifications include country fixed effects. Robust standard errors in parentheses. Significance levels indicated \* $p < .10$ , \*\* $p < .05$ , \*\*\* $p < .01$ .



**Table A6.** Heterogeneity by Manipulation Check.

	Active SM users				Passive/non-SM users			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.367*** (0.092)	0.380*** (0.096)	-0.042 (0.074)	-0.028 (0.074)	0.113* (0.061)	0.016 (0.053)	-0.118*** (0.043)	-0.056 (0.036)
<i>n</i>	139	139	220	220	329	329	696	696
<i>R</i> <sup>2</sup>	.129	.273	.021	.216	.013	.335	.012	.389
Correct m. check	Yes	Yes	No	No	Yes	Yes	No	No
Pre-attitudes	No	Yes	No	Yes	No	Yes	No	Yes
Other controls	No	Yes	No	Yes	No	Yes	No	Yes

SM: social media.

Ordinary least square regressions with the post-treatment attitudes index (first principal component of the responses to the post-treatment policy questions) as outcome. The treatment variable equals 1 for the pro-economy treatment, -1 for the pro-health treatment, and 0 otherwise. Pre-attitude controls include the first principal component of the pre-treatment policy questions and the single question with the highest correlation with post-treatment attitudes. The sample is split both between Active SM users and Passive/non-SM users and by those who correctly answered a post-treatment manipulation check asking participants which view had more likes in the six tweets shown. Controls include age, gender, region (fixed effects), education, income, and political position. All specifications include country fixed effects. Robust standard errors in parentheses. Significance levels indicated \**p* < .10, \*\**p* < .05, \*\*\**p* < .01.

**Table A7.** Heterogeneity by Manipulation Check Using Interactions.

	All individuals			Active SM users			Passive/non-SM users		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment	-0.116*** (0.043)	-0.067* (0.036)	-0.064* (0.035)	-0.047 (0.073)	0.005 (0.069)	-0.014 (0.071)	-0.118*** (0.043)	-0.055 (0.035)	-0.055 (0.035)
Treatment × Active SM user	0.062 (0.083)	0.091 (0.077)	0.084 (0.077)						
Treatment × Correct metrics	0.226*** (0.074)	0.082 (0.062)	0.087 (0.061)	0.417*** (0.114)	0.372*** (0.110)	0.396*** (0.113)	0.231*** (0.074)	0.055 (0.062)	0.068 (0.062)
Treatment × Active SM user × Correct metrics	0.215 (0.135)	0.267** (0.127)	0.260** (0.127)						
<i>n</i>	1,384	1,384	1,384	359	359	359	1,025	1,025	1,025
<i>R</i> <sup>2</sup>	.037	.304	.311	.061	.189	.208	.016	.354	.367
Pre-attitudes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Other controls	No	No	Yes	No	No	Yes	No	No	Yes

SM: social media.

Ordinary least square regressions with the post-treatment attitudes index (first principal component of the responses to the post-treatment policy questions) as outcome. The treatment variable equals 1 for the pro-economy treatment, -1 for the pro-health treatment, and 0 otherwise. Pre-attitude controls include the first principal component of the pre-treatment policy questions and the single question with the highest correlation with post-treatment attitudes. Controls include age, gender, region (fixed effects), education, income, and political position. All specifications include country fixed effects. Robust standard errors in parentheses. Significance levels shown \**p* < .10, \*\**p* < .05, \*\*\**p* < .01.

**Table A8.** Manipulation Check and Pre-Treatment Attitudes.

	Active SM users				Passive/non-SM users			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	-0.004 (0.100)	-0.040 (0.094)	-0.027 (0.081)	-0.145* (0.076)	0.122** (0.055)	0.198*** (0.058)	-0.085** (0.042)	-0.095** (0.042)
<i>n</i>	139	139	220	220	329	329	696	696
<i>R</i> <sup>2</sup>	.007	.072	.006	.037	.057	.068	.012	.012
Correct m. check	Yes	Yes	No	No	Yes	Yes	No	No
Pre-attitudes measure	PCA	Single Q	PCA	Single Q	PCA	Single Q	PCA	Single Q

SM: social media; PCA: principal components analysis.

Ordinary least square regressions with the pre-treatment attitudes as outcome. Pre-attitudes are measured both as the first principal component of the pre-treatment policy questions and the single question with the highest correlation with post-treatment attitudes. The treatment variable equals 1 for the pro-economy treatment, -1 for the pro-health treatment, and 0 otherwise. The sample is split both between Active SM users and Passive/non-SM users and by those who correctly answered a post-treatment manipulation check asking participants which view had more likes in the six tweets shown. Controls include age, gender, region (fixed effects), education, income, and political position. All specifications include country fixed effects. Robust standard errors in parentheses.

Significance levels indicated \**p* < .10, \*\**p* < .05, \*\*\**p* < .01.

**Table A9.** Treatment Effects by Country.

	All individuals			Active SM users			Passive/non-SM users		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment	-0.088 (0.062)	-0.012 (0.060)	-0.012 (0.035)	-0.319** (0.145)	-0.028 (0.173)	0.086 (0.074)	-0.143* (0.086)	0.049 (0.085)	-0.058 (0.042)
Treatment × Active SM user	0.115 (0.122)	0.241* (0.140)	0.154** (0.074)						
Treatment × Correct metrics				0.593*** (0.195)	0.494* (0.249)	0.270* (0.156)	0.112 (0.125)	-0.113 (0.127)	0.176** (0.082)
<i>n</i>	305	300	779	92	90	177	213	210	602
<i>R</i> <sup>2</sup>	.178	.144	.471	.190	.201	.440	.250	.220	.491
Pre-attitudes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country	IRE	ITA	USA	IRE	ITA	USA	IRE	ITA	USA

SM: social media; IRE: Ireland; ITA: Italy; USA: United States.

Ordinary least square regressions with the post-treatment attitudes index (first principal component of the responses to the post-treatment policy questions) as outcome. The treatment variable equals 1 for the pro-economy treatment, -1 for the pro-health treatment, and 0 otherwise. Pre-attitude controls include the first principal component of the pre-treatment policy questions and the single question with the highest correlation with post-treatment attitudes. Controls include age, gender, region (fixed effects), education, income, and political position. Results shown separately by country. Robust standard errors in parentheses.

Significance levels shown \**p* < .10, \*\**p* < .05, \*\*\**p* < .01.

**Table A10.** Anchoring Effects.

	All individuals			Active SM users			Passive/non-SM users		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
First-message econ	0.085 (0.062)	0.094* (0.050)	0.094* (0.050)	0.048 (0.114)	0.022 (0.102)	0.018 (0.106)	0.098 (0.073)	0.125** (0.057)	0.117** (0.057)
<i>n</i>	1,384	1,384	1,384	359	359	359	1,025	1,025	1,025
<i>R</i> <sup>2</sup>	.005	.280	.286	.015	.142	.160	.004	.353	.365
Pre-attitudes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Other controls	No	No	Yes	No	No	Yes	No	No	Yes

SM: social media.

Ordinary least square regressions with the post-treatment attitudes index (first principal component of the responses to the post-treatment policy questions) as outcome. The "First-message econ" variable equals 1 if participants were first exposed to a pro-economy tweet. Pre-attitude controls include the first principal component of the pre-treatment policy questions and the single question with the highest correlation with post-treatment attitudes. Controls include age, gender, region (fixed effects), education, income, and political position. All specifications include country fixed effects. Robust standard errors in parentheses.

Significance levels shown \**p* < .10, \*\**p* < .05, \*\*\**p* < .01.

**Table A11.** Complementarity Effects: Anchoring+ High-Popularity.

	All individuals			Active SM users			Passive/non-SM users		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
First-message econ × First-message high	0.017 (0.131)	0.037 (0.106)	0.043 (0.107)	0.377 (0.263)	0.517** (0.235)	0.510** (0.237)	-0.087 (0.151)	-0.121 (0.117)	-0.107 (0.117)
<i>n</i>	1,384	1,384	1,384	359	359	359	1,025	1,025	1,025
<i>R</i> <sup>2</sup>	.006	.281	.287	.040	.175	.190	.007	.353	.366
Pre-attitudes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Other controls	No	No	Yes	No	No	Yes	No	No	Yes

SM: social media.

Ordinary least square regressions with the post-treatment attitudes index (first principal component of the responses to the post-treatment policy questions) as outcome. The “First-message econ” variable equals 1 if participants were first exposed to a pro-economy tweet, and the “First-message high” variable equals 1 if participants were first exposed to a tweet with high-popularity metrics. Pre-attitude controls include the first principal component of the pre-treatment policy questions and the single question with the highest correlation with post-treatment attitudes. Controls include age, gender, region (fixed effects), education, income, and political position. All specifications include country fixed effects. Robust standard errors in parentheses.

Significance levels shown \**p* < .10, \*\**p* < .05, \*\*\**p* < .01.

**Table A12.** Heterogeneity by Pre-Treatment Attitudes.

	All individuals		Active SM users		All individuals		Active SM users	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	-0.033 (0.032)	-0.026 (0.032)	0.092* (0.055)	0.079 (0.057)	-0.041 (0.029)	-0.038 (0.029)	0.148*** (0.053)	0.141*** (0.054)
Treatment × Active SM user	0.148** (0.065)	0.130** (0.065)			0.205*** (0.061)	0.198*** (0.061)		
Pre-treatment attitudes	0.359*** (0.030)	0.352*** (0.032)	0.107* (0.055)	0.149** (0.060)	0.529*** (0.027)	0.516*** (0.028)	0.352*** (0.057)	0.351*** (0.058)
Treatment × Pre-attitudes	-0.008 (0.040)	-0.014 (0.040)	0.116* (0.060)	0.114* (0.059)	-0.035 (0.034)	-0.042 (0.034)	0.164*** (0.062)	0.162** (0.063)
Treatment × Active SM user × Pre-attitudes	0.076 (0.076)	0.087 (0.077)			0.185** (0.074)	0.197*** (0.074)		
<i>n</i>	1,384	1,384	359	359	1,384	1,384	359	359
<i>R</i> <sup>2</sup>	.144	.158	.057	.087	.298	.307	.183	.199
Pre-attitudes measure	PCA	PCA	PCA	PCA	Single Q	Single Q	Single Q	Single Q
Other controls	No	Yes	No	Yes	No	Yes	No	Yes

SM: social media; PCA: principal components analysis.

Ordinary least square regressions with the post-treatment attitudes index (first principal component of the responses to the post-treatment policy questions) as outcome. The treatment variable equals 1 for the pro-economy treatment, -1 for the pro-health treatment, and 0 otherwise. ActiveSMuser is a dummy variable which equals 1 if the subject spends more than 1 hr a day on Facebook or Twitter (combined) and zero otherwise. Controls include age, gender, region (fixed effects), education, income, and political position. All specifications include country fixed effects. Robust standard errors in parentheses.

Significance levels indicated \**p* < .10, \*\**p* < .05, \*\*\**p* < .01.

**Table A13.** Defining Active SM Users as Those Using Facebook for At Least 30 Min a Day.

	Active SM users				Active SM users		Passive/non-SM users		Active SM users			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treatment	0.102** (0.049)	0.097** (0.049)	0.091* (0.050)	0.104** (0.046)	0.066 (0.046)	0.060 (0.047)	0.100** (0.044)	0.096** (0.045)	0.265*** (0.077)	0.259*** (0.078)	-0.002 (0.065)	0.008 (0.060)
Pre-treatment attitudes					0.166*** (0.052)	0.192*** (0.056)	0.420*** (0.049)	0.424*** (0.050)				
Treatment × Pre-attitudes					0.141** (0.057)	0.143** (0.057)	0.136** (0.056)	0.133** (0.056)				
<i>n</i>	517	517	517	517	517	517	517	517	202	202	315	315
<i>R</i> <sup>2</sup>	.016	.050	.064	.205	.066	.080	.211	.218	.064	.204	.012	.261
Correct m. check									Yes	Yes	No	No
Pre-attitudes (PCA)	No	Yes	Yes	Yes					No	Yes	No	Yes
Other controls	No	No	Yes	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Pre-attitudes (Single Q)	No	No	No	Yes					No	Yes	No	Yes
Pre-attitudes					PCA	PCA	Single Q	Single Q				

SM: social media; PCA: principal components analysis.

Ordinary least square regressions with the post-treatment attitudes index (first principal component of the responses to the post-treatment policy questions) as outcome. The treatment variable equals 1 for the pro-economy treatment, -1 for the pro-health treatment, and 0 otherwise. Pre-attitudes are measured both as the first principal component of the pre-treatment policy questions and/or the single question with the highest correlation with post-treatment attitudes, as indicated. Controls include age, gender, region (fixed effects), education, income, and political position. All specifications include country fixed effects. Robust standard errors in parentheses.

Significance levels shown \* $p < .10$ , \*\* $p < .05$ , \*\*\* $p < .01$ .

**Table A14.** Attention Prime Treatment.

	Active SM users			Passive/non-SM users		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.124 (0.084)	0.145** (0.064)	0.153** (0.063)	-0.011 (0.052)	-0.017 (0.035)	-0.012 (0.035)
Attention prime	-0.083 (0.096)	-0.050 (0.076)	-0.057 (0.076)	0.000 (0.059)	-0.001 (0.044)	0.002 (0.043)
Treatment × Attention prime	-0.265** (0.118)	-0.194** (0.087)	-0.183** (0.085)	0.053 (0.072)	0.024 (0.052)	0.017 (0.052)
<i>n</i>	325	325	325	1,194	1,194	1,194
<i>R</i> <sup>2</sup>	.017	.396	.438	.001	.445	.449
Pre-attitudes	No	Yes	Yes	No	Yes	Yes
Other controls	No	No	Yes	No	No	Yes

SM: social media.

Ordinary least square regressions with the post-treatment attitudes index (first principal component of the responses to the post-treatment policy questions) as outcome. The treatment variable equals 1 for the pro-economy treatment, -1 for the pro-health treatment, and 0 otherwise. Pre-attitude controls include the first principal component of the pre-treatment policy questions and the single question with the highest correlation with post-treatment attitudes. The Attention Prime treatment showed participants a non-COVID-related tweet and asked questions about this (including number of likes), before the treatment. Controls include age, gender, region (US Midwest, US Northeast, US South, US West), education, income, and political position. Only the US sample was subject to this treatment. See the supplementary materials for more details. Robust standard errors in parentheses.

Significance levels shown \* $p < .10$ , \*\* $p < .05$ , \*\*\* $p < .01$ .

**Table A15.** Correct Manipulation Check and the Attention Prime Treatment.

	Active SM users			Passive/non-SM users		
	(1)	(2)	(3)	(4)	(5)	(6)
Attention prime	0.038 (0.052)	0.031 (0.052)	0.027 (0.052)	0.087*** (0.026)	0.086*** (0.026)	0.089*** (0.026)
<i>n</i>	325	325	325	1,194	1,194	1,194
<i>R</i> <sup>2</sup>	.002	.022	.043	.009	.011	.028
Pre-attitudes	No	Yes	Yes	No	Yes	Yes
Other controls	No	No	Yes	No	No	Yes

SM: social media.

Ordinary least square regressions with an indicator (1/0) for whether individuals correctly answered the manipulation check as outcome. Pre-attitude controls include the first principal component of the pre-treatment policy questions and the single question with the highest correlation with post-treatment attitudes. The Attention Prime treatment showed participants a non-COVID-related tweet and asked questions about this (including number of likes), before the treatment. Controls include age, gender, region (US Midwest, US Northeast, US South, US West), education, income, and political position. Only the US sample was subject to this treatment. See the supplementary materials for more details. Robust standard errors in parentheses. Significance levels shown \**p* < .10, \*\**p* < .05, \*\*\**p* < .01.

**Table A16.** Political Engagement and SM Use.

	Voted				Discuss policy on SM		Discuss policy off SM	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Active SM user	0.069*** (0.025)	0.086*** (0.024)	0.044** (0.020)	0.059*** (0.020)	1.237*** (0.081)	1.039*** (0.078)	0.424*** (0.074)	0.370*** (0.072)
<i>n</i>	605	605	1,519	1,519	1,519	1,519	1,519	1,519
<i>R</i> <sup>2</sup>	.010	.166	.003	.060	.139	.235	.023	.084
Sample	EU	EU	USA	USA	USA	USA	USA	USA
Other controls	No	Yes	No	Yes	No	Yes	No	Yes

SM: social media; EU: European Union; USA: United States.

The table shows the correlation between political engagement and SM use. The dependent variable in columns 1 to 4 is a dummy equal to 1 if the individual reports having voted in the previous elections. The dependent variable in columns 5 to 8 is a numerical value (0–4) to the question “How often do you discuss policy issues with your friends or family members on SM (columns 5–6) / outside of SM (columns 7–8)? [never, rarely, sometimes, often, always].” This question was only asked for the US sample. Controls include age, gender, region (fixed effects), education, income, and political position. Robust standard errors in parentheses.

Significance levels indicated \**p* < .10, \*\**p* < .05, \*\*\**p* < .01.

**Table A17.** Determinants of Correct Manipulation Check.

	All individuals	Active SM users	Passive/non-SM users
	(1)	(2)	(3)
Age	−0.000 (0.001)	−0.001 (0.002)	−0.000 (0.001)
Male	−0.007 (0.025)	0.037 (0.052)	−0.023 (0.029)
Education	−0.044 (0.028)	−0.070 (0.060)	−0.034 (0.032)
Income	0.008* (0.005)	0.005 (0.011)	0.009* (0.005)
Political ideology	0.004 (0.005)	0.001 (0.009)	0.006 (0.006)
Ireland	0.206*** (0.034)	0.174*** (0.066)	0.219*** (0.040)
Italy	0.173*** (0.036)	0.142*** (0.069)	0.184*** (0.042)

(Continued)

**Table A17.** (Continued)

	All individuals	Active SM users	Passive/non-SM users
	(1)	(2)	(3)
Active SM users	-0.036 (0.048)		
Facebook use	0.030* (0.018)	0.035 (0.040)	0.028 (0.021)
Twitter use	0.021 (0.017)	0.022 (0.027)	0.020 (0.026)
Constant	0.175*** (0.059)	0.189 (0.173)	0.154** (0.066)
<i>n</i>	1,384	359	1,025
<i>R</i> <sup>2</sup>	.055	.040	.057

SM: social media.

Ordinary least square estimates of the correlation between correctly answering the manipulation check question (as dependent variable), and various covariates: age, gender (coded as a dummy for male), education (dummy for having at least a 2 year college degree), household income (log of the midpoint of the interval specified by the subject), political ideology (self-reported response on a 0–10 left-right scale), country dummies (US as the omitted variable), ActiveSMuser (dummy for spending more than 1 hr a day on Facebook or Twitter combined), Facebook use (daily time spent on Facebook: 0 = never/no account, 1 = less than 30 min, 2 = from 30 min to 1 hr, 3 = more than 1 hr), Twitter use (daily time spent on Twitter: 0 = never/no account, 1 = less than 30 min, 2 = from 30 min to 1 hr, 3 = more than 1 hr). Robust standard errors in parentheses.

Significance levels indicated \**p* < .10, \*\**p* < .05, \*\*\**p* < .01.

**Table A18.** Main Treatment Effects Excluding Top and Bottom 5% in Study Duration.

	All individuals									Active SM users		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treatment	-0.008 (0.032)	0.001 (0.030)	0.002 (0.030)	0.010 (0.027)	-0.054 (0.037)	-0.042 (0.033)	-0.036 (0.033)	-0.038 (0.030)	0.138** (0.058)	0.141** (0.058)	0.130** (0.061)	0.152*** (0.058)
Treatment × Active SM user					0.197*** (0.068)	0.200*** (0.069)	0.185*** (0.070)	0.205*** (0.066)				
<i>n</i>	1,246	1,246	1,246	1,246	1,246	1,246	1,246	1,246	324	324	324	324
<i>R</i> <sup>2</sup>	.002	.130	.140	.295	.024	.165	.178	.316	.031	.047	.086	.165
Pre-attitudes (PCA)	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Other controls	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes
Pre-attitudes (Single Q)	No	No	No	Yes	No	No	No	Yes	No	No	No	Yes

SM: social media; PCA: principal components analysis.

Ordinary least square estimates using post-treatment attitudes index (first principal component of the responses to the post-treatment policy questions) as outcome. The treatment variable equals 1 for the pro-economy treatment, -1 for the pro-health treatment, and 0 otherwise. Pre-attitude controls include the first principal component of the pre-treatment policy questions and the single question with the highest correlation with post-treatment attitudes. Columns 5 to 8 present estimates of models of this form

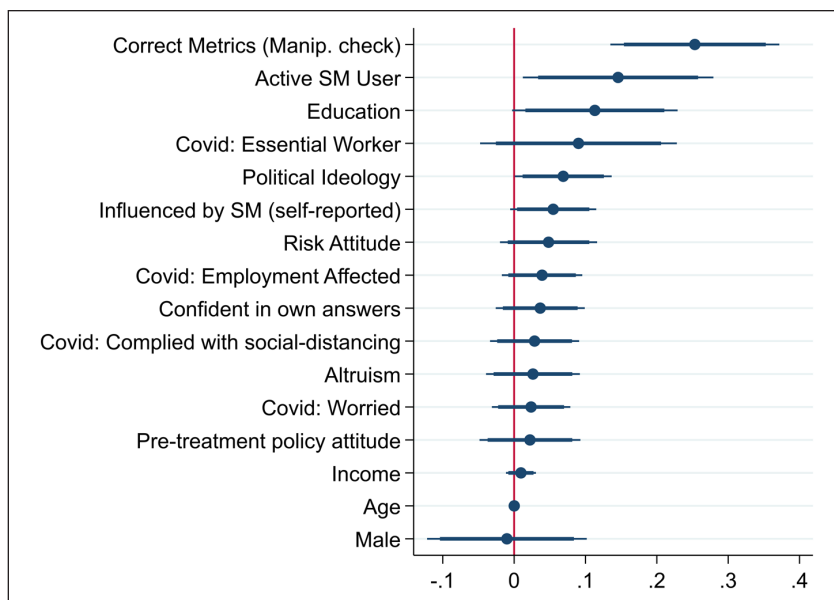
$$PostAttitudes_i = \beta_0 + \beta_1 Treatment_i + \beta_2 Var_i + \beta_3 Treatment_i \times Var_i + X_i' \delta + \varepsilon_i$$

ActiveSMuser is a dummy variable which equals 1 if the subject spends more than 1 hr a day on Facebook or Twitter (combined) and zero otherwise. Controls include age, gender, region (fixed effects), education, income, and political position. All specifications include country fixed effects. Robust standard errors in parentheses.

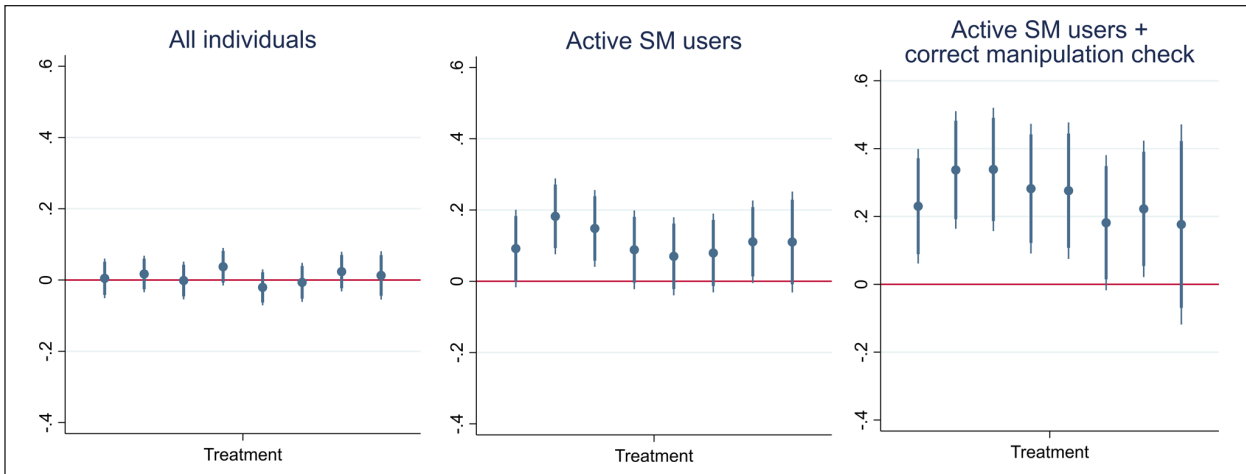
Significance levels indicated \**p* < .10, \*\**p* < .05, \*\*\**p* < .01.



**Figure A1.** Tweets used in the experiment. These tweets were shown to all participants. Tweets were classified as pro-health (left) or pro-economy (right). The number of likes/retweets varied depending on the randomly assigned treatment arm.



**Figure A2.** Treatment effects along various margins of heterogeneity. The figure shows the coefficient  $\beta_3$  from the following model, as also used in Appendix Table A2, column 7  $PostAttitudes_i = \beta_0 + \beta_1 Treatment_i + \beta_2 Var_i + \beta_3 Treatment_i \times Var_i + X_i' \delta + \epsilon_i$  where  $Var_i$  indicates the dimensions of interest in exploring heterogeneous effects. Controls include age, gender, region (fixed effects), education, income, and political position. All specifications include country fixed effects. Full estimate tables are provided in the Supplementary Materials.



**Figure A3.** Treatment effects separately for each post-treatment question.

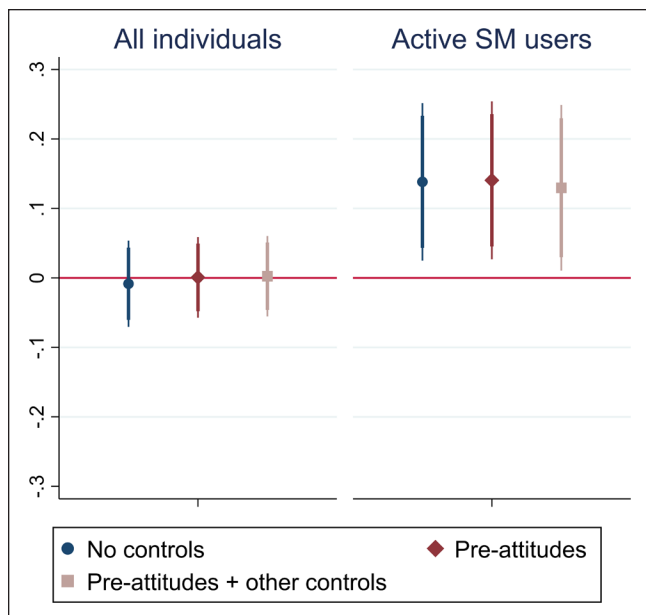
The figure shows our main results individually for each of the (standardized) post-treatment questions. From left to right, these are agreement with: “Closing the borders”; “Prohibiting gatherings”; “Prohibiting non-essential travels”; “Closing daycares, schools, colleges and universities”; “Closing non-essential businesses (bars, stores that are not food or health related, etc.)”; “Handing out USD/EUR 1,000 fines to those who do not comply with social-distancing rules”; “General lockdown of the population with a ban on leaving the home (except for medical reasons)”; and “Mandatory use of face-coverings in public places” (US only). The results are shown, respectively, for all participants, active social media users, and active social media users who correctly answered the factual manipulation check question.



**Figure A4.** Pre-treatment policy attitudes and social media use.

The figure shows the distribution of responses to the pre-treatment policy attitude questions, separately for active and for passive/non-social media users. Questions marked with a \* include participants in both Europe ( $n=605$ ) and the United States ( $n=1,519$ ). All others include only US participants.





**Figure A5.** Treatment effects excluding top and bottom 5% in study duration. The figure shows the main treatment effects for all users excluding those from the top and bottom 5% in study duration ( $n = 1,246$ ) and active social media users ( $n = 324$ ) separately. Active social media users are defined as individuals who spend more than 1 hr daily on Facebook or Twitter combined. Estimates including a fully interacted model that tests for differences between the groups can be found in Appendix Table A18.