



**HAL**  
open science

## **Curtailing False News, Amplifying Truth**

Sergei Guriev, Emeric Henry, Théo Marquis, Ekaterina Zhuravskaya

► **To cite this version:**

Sergei Guriev, Emeric Henry, Théo Marquis, Ekaterina Zhuravskaya. Curtailing False News, Amplifying Truth. 2023. halshs-04315924

**HAL Id: halshs-04315924**

**<https://shs.hal.science/halshs-04315924>**

Preprint submitted on 30 Nov 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



WORKING PAPER N° 2023 – 34

## Curtailling False News, Amplifying Truth

Sergei Guriev  
Emeric Henry  
Théo Marquis  
Ekaterina Zhuravskaya

JEL Codes:  
Keywords:



# Curtailing False News, Amplifying Truth\*

Sergei Guriev

Emeric Henry

Théo Marquis

Ekaterina Zhuravskaya

November 29, 2023

## Abstract

We develop a comprehensive framework to assess policy measures aimed at curbing false news dissemination on social media. A randomized experiment on Twitter during the 2022 U.S. mid-term elections evaluates such policies as priming the awareness of misinformation, fact-checking, confirmation clicks, and prompting careful consideration of content. Priming is the most effective policy in reducing sharing of false news while increasing sharing of true content. A model of sharing decisions, motivated by persuasion, partisan signaling, and reputation concerns, predicts that policies affect sharing through three channels: (i) updating perceived veracity and partisanship of content, (ii) raising the salience of reputation, and (iii) increasing sharing frictions. Structural estimation shows that all policies impact sharing via the salience of reputation and cost of friction. Affecting perceived veracity plays a negligible role as a mechanism in all policies, including fact-checking. The priming intervention performs best in enhancing reputation salience with minimal added friction.

---

\*Guriev: Department of Economics, Sciences Po, 27 rue Saint Guillaume, 75007 Paris, France and CEPR (sergei.guriev@sciencespo.fr); Henry: Department of Economics, Sciences Po, 28 rue des Saints Peres, 75007 Paris, France and CEPR (emeric.henry@sciencespo.fr); Marquis: Department of Economics, Sciences Po, 28 rue des Saints Peres, 75007 Paris, France (theo.marquis@sciencespo.fr); Zhuravskaya: Paris School of Economics (EHESS) 48 boulevard Jourdan, 75014 Paris, France and CEPR (ezhuravskaya@gmail.com). This experiment obtained the approval of the internal Institutional Review Board of the Paris School of Economics; the approval number is IRB 2022-024. The experiment was preregistered at the AEA RCT Registry; registry number AEARCTR-0010219 (Guriev et al., 2022). We thank Project Liberty Institute for generous financial support. We are grateful to Charles Angelucci, David Atkin, Roland Bénabou, Leonardo Bursztyn, Micael Castanheira, Ruben Durante, Raymond Fisman, Matthew Gentzkow, Irena Grosfeld, Jeanne Haggenbach, Thierry Mayer, Andrea Prat, Gautam Rao, Carlo Schwarz, Evgeny Yakovlev, David Yanagizawa-Drott, and the participants of the 2nd CEPR Workshop on Media, Technology, Politics, and Society, conferences in Sciences Po and University of Chicago, a state-of-the-art session at the EEA Congress, and seminars at Boston University, CEU, CREST, and HEC for helpful comments.

# 1 Introduction

The spread of misinformation on social media is a major policy issue (e.g., Persily and Tucker, eds, 2020). There is a consensus among social scientists that the informativeness of voters is a key pillar of democracy (e.g., Besley and Prat, 2006). Yet, false political news disseminate widely on social media platforms (e.g., Allcott and Gentzkow, 2017; Vosoughi et al., 2018) and they can be highly persuasive (e.g., Barrera et al., 2020; Nyhan et al., 2019), while an increasing number of people worldwide rely on social media for political news.<sup>1</sup> This leads to is a growing concern that, if not curtailed, misinformation on social media could potentially result in distorted political outcomes, heightened affective polarization, political turmoil, an escalation of hate crimes, and the amplification of systemic health risks. Recent evidence supports the legitimacy of these concerns (e.g., Lazer et al., 2018; Allcott et al., 2020; Finkel et al., 2020; Bursztyjn et al., 2020; Guriev et al., 2021; Levy, 2021; Skaffe et al., 2022; Müller and Schwarz, forthcoming).<sup>2</sup>

These dangers, both potential and realized, have triggered a public debate on policy solutions. On the legislative front, the agenda of regulating content on social media platforms is advancing in many countries. Most notably, the European Union has adopted the Digital Services Act (EU, 2023). However, the extent of what can be achieved through regulation is severely limited by laws safeguarding free speech. In the United States, content moderation by means of regulation is unconstitutional, while in the European Union, it is restricted to addressing illegal content, whereas much of the misinformation not falling into that category. To address this challenge, a proposed policy solution involves implementing large-scale digital literacy training to enhance citizens’ ability to discern true information from false news (e.g., Guess et al., 2020).

Researchers have also proposed more modest, short-term, yet potentially highly effective measures. These include offering fact-checking, requiring confirmation clicks to access and share false content, as well as different nudges aimed at prompting users to consider the dangers of spreading false information or to think about the content they want to share. Some of these measures have been implemented by platforms, likely in response to social or political pressure. For instance, Facebook initiated its “Third Party Fact Checking program” in 2016, while Twitter introduced fact-checking and confirmation clicks, particularly for accessing selected content on

---

<sup>1</sup>As of April 2023, 60% of the world’s population actively used social media. Among adult social-media users, 34.6% named “reading news stories” as one of the main reasons to use social media, behind only “keeping in touch with friends and family” (49%) and “filling spare time” (37%) (see <https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/> and <https://datareportal.com/social-media-users>, accessed August 1, 2023.) Similarly, one out of five U.S. voters uses social media as the primary source of getting political news and, specifically, information about elections (see Pew Research Center’s survey conducted in 2020: [https://www.pewresearch.org/journalism/wp-content/uploads/sites/8/2020/07/PJ\\_2020.07.30\\_social-media-news\\_00-01.png?w=609](https://www.pewresearch.org/journalism/wp-content/uploads/sites/8/2020/07/PJ_2020.07.30_social-media-news_00-01.png?w=609), accessed August 1, 2023.)

<sup>2</sup>See also the anecdotal evidence on the role of social media in Brexit and in the January 6 siege of the U.S. Capitol: <https://firstdraftnews.org/articles/how-leave-eu-dominates-the-brexit-conversation-on-facebook/> and <https://www.propublica.org/article/facebook-hosted-surge-of-misinformation-and-insurrection-threats-in-months-leading-up-to-jan-6-attack-records-show>, accessed September 25, 2023.

Donald Trump’s account in 2020.<sup>3</sup> These interventions have been rigorously evaluated (see, e.g., Pennycook et al., 2020; Fazio, 2020; Yaqub et al., 2020; Henry et al., 2022; Arechar et al., 2023, a survey by Nyhan, 2020, and meta-analyses by Pennycook and Rand, 2022 and Kozyreva et al., 2022). The results show convincing and replicable evidence that, at least within the studied settings, each of these interventions reduces the circulation of false news on social media.

This significant body of research has several limitations. First, and most importantly, all papers thus far have exclusively relied on reduced-form analyses, limiting the ability to extrapolate beyond specific experimental contexts and to comprehend the mechanisms through which these different policies operate. Second, most contributions focus on evaluating one type of policy intervention per setting, making it hard to compare the impact of alternative policies. Third, only the studies focusing on priming users about misinformation have considered the dissemination of true news in addition to false news (see, Pennycook and Rand, 2022; Guay et al., 2023). However, it is crucial to assess the impact of all interventions on the dissemination of accurate news. This is essential not only for the welfare benefit of having informed voters but also in terms of enforceability considerations. Policies that enhance—or at the very least do not diminish—the circulation of true news align more closely with the platforms’ incentives. As social media platforms seek to maximize user engagement, they are more likely to adopt policies with a smaller negative impact on overall user engagement.<sup>4</sup>

In this paper, we address these limitations by developing a unified framework for assessing the impact of different policies on the circulation of both accurate and false news. A structural model of sharing allows us to analyze the mechanisms through which these policies operate. It also enables us to recover the indirect utility of sharing on social media and, therefore, characterize the key drivers of sharing. The model also allows us to evaluate the effect of counterfactual policies, such as digital literacy training.

During the U.S. mid-term election campaign in the Fall of 2022, we conducted a randomized controlled experiment designed to closely mimic an actual sharing experience on Twitter. We presented four tweets with political information to 3,501 American Twitter users. Two tweets contained misinformation, while the other two contained true facts. These tweets were posted on a Twitter account named “2022 Political News,” created for the experiment’s purpose. This account automatically posted news originating from liberal and conservative online media. The four tweets involved in the experiment were manually added to this account right at the beginning of the experiment. Participants viewed screenshots of the tweets within the survey environment and were provided with links to directly access the tweets on Twitter.

---

<sup>3</sup>In May 2020, Twitter accompanied Donald Trump’s false statement regarding the “fraudulent” mail vote with a blue link to a suggested fact-checking (Conger and Alba, 2020). During the same month, Twitter required users to confirm whether they wanted to view Trump’s tweets that glorified violence with the phrase “when looting starts, shooting starts” (Conger, 2020). In October 2020, Twitter introduced a new sharing interface that included a window with a prompt to add comments before sharing content. According to the company, the objective was to “encourage more thoughtful consideration” and reduce the spread of misinformation. See Ershov and Morales (2023) for the analysis of the effects of these changes.

<sup>4</sup>For the discussion of platforms’ incentives, see Tufekci (2018); Haidt and Rose-Stockwell (2019); Allcott et al. (2019); Henry et al. (2022); Acemoglu et al. (2021).

Subsequently, we subjected randomly selected subgroups of the experiment’s participants to treatments simulating policy measures aimed at limiting the circulation of false news. After the treatment phase, we asked the participants whether they would like to share one of these tweets on their Twitter accounts. If they agreed, the survey environment directed them to Twitter, where they could confirm the retweet of the chosen tweet.

Altogether, there are five treatment groups. The first group served as the control. In this group, participants did not receive any treatment and proceeded to the sharing decision directly after viewing the tweets. We refer to this treatment group as “No policy.” The second group received the “Extra click” treatment, which involved an additional confirmation click required for sharing, making sharing slightly more costly. Participants in the third group were subjected to the treatment, which we call “Prime fake news circulation.” Before sharing, they were shown a screen with a warning message in the spirit of the nudges surveyed by [Pennycook and Rand \(2022\)](#) and [Guay et al. \(2023\)](#): “Please think carefully before you retweet. Remember that there is a significant amount of false news circulating on social media.” The “Offer fact check” treatment, applied to the fourth treatment group, informed participants that the two false tweets had been fact-checked by PolitiFact.com, a reputable fact-checking NGO, and provided a link to access the fact-checking of these tweets. In the last treatment, “Ask to assess tweets,” we asked participants to provide their best assessments of the accuracy and partisan leaning of the four tweets, thus prompting concerns about accuracy and partisan biases related to the specific content.

An essential element of our study, which is central in the structural analysis, is the collection of information on individuals’ perceptions of tweet characteristics. After making the sharing decision, participants from the first four treatment groups were asked to evaluate the accuracy and partisan leaning of each tweet. The answers were incentivized using state-of-the-art experimental methods ([Danz et al., 2022](#)). Participants in the fifth group, i.e., ask to assess tweets treatment, conducted these assessments before sharing. Comparing the assessments between this treatment and the no policy group allows us to test whether sharing itself affects people’s assessment of the content.

Randomization achieved balance across treatment groups across a wide range of pre-treatment characteristics. This allows us to measure the reduced-form impact of the treatments on the sharing of true and false news by comparing the average sharing rates across treatment groups. In the no policy group, 28% of participants shared one of the false news tweets, while 30% of participants shared one of the true tweets. Consistent with findings from previous literature, all the treatments reduced the sharing of false tweets. Sharing rates of false news in the (i) extra click treatment, (ii) priming fake news circulation treatment, (iii) offering fact-check treatment, and (iv) the treatment that asks participants to assess the content prior to sharing are 3.6, 11.5, 13.6, and 14.1 percentage points lower than in the no policy group, respectively.

However, the treatments have markedly different effects on the sharing of true tweets. Requiring an extra click and asking participants to assess tweets prior to sharing have no effect.

Offering a fact-check reduces the sharing of true tweets by 7.8 percentage points relative to the no policy group. In contrast, the priming treatment increases the average sharing rate of true tweets by 8.1 percentage points. These results offer a clear ranking of the policies in terms of their effectiveness in enhancing the accuracy of shared political content. The priming fake news circulation treatment emerges as the most effective strategy, as it increases the “sharing discernment” advocated by [Pennycook and Rand \(2022\)](#).

To understand the mechanism behind the differential effects of the treatments on the sharing of true and false news, we develop and structurally estimate a game-theoretic model of sharing political information on social media. In this model, a potential sharer of such information (a “sender”) aims to influence the beliefs and actions of their audience (the “receivers”). Both the sender and the receivers have incomplete information about the state of the world which can either be such that a Democratic or a Republican political action is optimal.<sup>5</sup> The sender has access to political news that can be shared. This news may be true or false. With a certain probability, the sender and the receivers perfectly know the veracity of the news, those who do are considered “informed.”

The sender decides whether to share the news with the receivers. When the receivers receive political news from the sender, they update their beliefs about the world, which in turn affects their political action, as well as their beliefs about whether the sender is informed and what the sender’s partisan beliefs are. The sharing decision of the sender is driven by three motives: (i) The *reputation motive*: the sender aims to maintain their reputation as an informed and credible source. (ii) The *partisan persuasion motive*: the sender benefits from persuading receivers to adopt their own perspective on the correct state of the world. (iii) The *signaling partisanship* motive: the sender benefits from signaling their partisan beliefs—i.e., their priors about the state of the world—to the receivers. This could, for instance, be because they want to appear loyal to their party, although we do not explicitly model this aspect.

We demonstrate that, in equilibrium, the utility derived from sharing a given piece of news is a function of the news’ perceived veracity, perceived partisan alignment between the sender and the news, and the interaction between these two characteristics. The effects of veracity and partisan alignment are positive. Veracity positively affects the utility of sharing due to the reputation motive: informed receivers, upon receiving false news, are more likely to perceive the sender as uninformed. Both partisan motives—persuasion and signaling—imply that the sender gets higher payoff when they share aligned news. However, the effect of the interaction of veracity and alignment on sharing utility is a priori ambiguous and depends on which partisan motive dominates. Persuasion is facilitated when the news is more likely to be true since receivers are more likely to modify their actions accordingly. In contrast, signaling partisanship is more effective when the news is likely to be false. This is because a likely-false partisan message conveys a more credible signal of partisanship than a true partisan message.

We use the experimental data to estimate the model using the random utility discrete

---

<sup>5</sup>Political action is a broad concept that may encompass various behaviors, such as voting, making political donations, participating in protests, or sharing political content on social media.



choice framework (McFadden, 1973, 1974). Without loss of generality, we decompose the sharing choices of the experiment’s participants into two nested decisions: (i) whether to share anything or share nothing in the upper nest, and (ii) which tweet to share in the lower nest, conditional on the upper-nest decision to share something. The choice in the lower nest depends on the assessment of veracity and partisan alignment, as predicted by our model. In the upper nest, it depends on the sharing frictions as well as the inclusive utility derived from the optimal choice in the lower nest. We structurally estimate these relationships.

First, the estimation of the lower-nest relationship enables us to assess the sender’s motives for sharing. We find that the reputation motive is a crucial driver of sharing, as participants’ perception of content veracity has a substantial, positive, and robust impact on the decision to share a specific tweet. Partisan motives also play an important role because we observe a positive effect of partisan alignment on sharing. Moreover, among the two partisan motives, partisan persuasion dominates signaling partisanship as a motive for sharing because the interaction between veracity and partisan alignment also has a strong consistently positive effect on sharing. Taking this analysis further, we numerically solve the system of nonlinear equations that characterize the equilibrium using the estimates obtained from the lower nest as inputs. This allows us to recover the actual parameters of the sender’s utility function. We determine the relative weights assigned to the three motives within the sender’s utility function and demonstrate how these weights are influenced by the policy treatments. We show that all treatments increase the weight the senders place on reputation relative to partisan motives and that this effect is particularly strong for the priming treatment. To the best of our knowledge, our paper is the first to estimate the parameters of the utility function of sharing political content on social media.

Second, we use structural analysis to uncover the mechanisms through which various policy measures affect sharing. Overall, anti-misinformation policies can influence sharing through three channels, which we refer to as Updating, Salience, and Cost. The updating channel operates by treatments potentially prompting the sender to revise their beliefs about the content’s veracity and partisan alignment. For example, fact-checking aims to alter users’ perceptions of content accuracy. We directly measure this effect by estimating the impact of treatments on the participants’ assessment of content characteristics. We find that several treatments significantly affect estimates of veracity and partisanship, but these effects are small in magnitude. The salience channel works through treatments altering the relative salience of reputation concerns compared to the partisan motives in the sharer’s utility function. The nudges, for instance, are designed to affect salience. We estimate the salience channel in the lower nest, when we solve for the treatment-specific relative weights placed on different motives in the sender’s utility function. Finally, the treatments also affect sharing frictions. This is the cost of sharing channel. For instance, adding an extra confirmation click increases the sharing cost for all types of content, whether true or false. The estimation of the upper-nest relationship provides the measure of the cost of sharing by treatment.

To calculate the contribution of each of the three channels to the overall effects of treat-



ments, we conduct a series of counterfactual exercises. In particular, we simulate the counterfactual effects of each treatment consecutively shutting down some channels and allowing others to be at play. The most surprising finding of this simulation exercise is the negligible contribution of the updating channel to the effect of any of the treatments. Even though several treatments affect the sender’s estimates of the veracity and partisan alignment of the news, this has little effect on the sharing of false and true news. This non-result is particularly striking in the case of the fact-checking treatment, which does statistically significantly decrease the estimates of the veracity of false news. However, the magnitude of this updating is not substantial for any of the treatments, including fact-checking.

Instead, the overall effect of each treatment comes from the combination of how they influence the salience of reputation and the cost of sharing. The combined contribution of these two channels is one order of magnitude larger than that of the updating channel. In particular, the salience channel is what drives the difference in the effect of treatments on sharing false and true news: raising salience of reputation affects sharing true news positively and false news negatively. All treatments, to varying degrees, increase the salience of reputation.<sup>6</sup> Priming fake news circulation has the largest salience effect. At the same time, costs of sharing associated with different treatments drive sharing of both true and false news down. The additional costs in the priming treatment are relatively low and comparable to adding an extra click. They are substantially smaller than for the offer fact-check and ask to assess tweets treatments, which require more time and mental effort.<sup>7</sup> Overall, the priming treatment stands out as the most cost-effective policy.

The structural estimation also enables us to simulate the effects of digital literacy training, a policy that cannot be evaluated within a short-term survey experiment. In the model, this policy corresponds to an increase in the share of informed senders and receivers. This change affects all parameters determined in equilibrium. Consequently, we solve the model separately for each counterfactual value of the share of informed individuals. Our results reveal that digital literacy training has a strong positive effect on the sharing of true news and a more modest negative effect on the sharing of false tweets. Increasing the share of informed individuals by 50% (from 4% to 6%) results in a 13.1 percentage point increase in the sharing of true tweets and a 6.6 percentage point decrease in the sharing of false tweets. We can decompose this effect into the impacts of two channels: sender’s knowledge and receivers’ reactions. The first channel is direct: informed senders, whose share increases by digital literacy training, are less likely to share false news. The second channel is indirect and works through the expectation of the sender that receivers will react differently to news shared by them because receivers are more likely to be informed. We find that the overall effect of digital literacy training is mostly

---

<sup>6</sup>This is true even for the extra click treatment, which introduces a pause before sharing, prompting users to consider their reputation. Unsurprisingly, the effect of the extra click on the salience of reputation is the smallest among the policy interventions.

<sup>7</sup>The costs of fact-checking in our setup only include the costs of (potentially) accessing already existing fact-checking information. The full social cost of fact-checking is substantially larger, as it also involves the identification of news that should be fact-checked and the fact-checking process itself.

driven by the second, indirect channel. Senders are less likely to share false news as a result of digital literacy training primarily because they expect that better-informed receivers would not be persuaded, and they would negatively update their view of the sender’s knowledge if false news is shared with them.

Our analysis has important policy implications. First, it demonstrates that the most effective short-term policy is the priming of fake news circulation advocated by [Pennycook and Rand \(2022\)](#). This policy decreases the sharing of false news and increases the sharing of true news without reducing the overall engagement of social media users. Second, we show that fact-checking operates through the salience, not the updating mechanism. This implies that providing accurate and expensive fact-checking by professional fact-checkers is less effective than a faster, cheaper, but more error-prone algorithmic fact-checking, in which users are simply informed that the content is flagged as suspicious by the algorithm. Third, we show that digital literacy training is not only effective ([Guess et al., 2020](#)), but also—due to the mechanism via which it operates—complements the cheaper and easier to implement short-term policies, particularly, the priming of fake news circulation.

We make several contributions to the recent experimental literature on the effects of policies designed to mitigate social media misinformation ([Nyhan, 2020](#); [Pennycook et al., 2020](#); [Fazio, 2020](#); [Yaqub et al., 2020](#); [Henry et al., 2022](#); [Arechar et al., 2023](#); [Pennycook and Rand, 2022](#); [Kozyreva et al., 2022](#); [Pillai and Fazio, 2023](#)). First, by developing and structurally estimating a model of information sharing, we shed light on the mechanisms through which different policies operate, which cannot be done with reduced-form analyses. Second, in contrast to the existing literature, our paper investigates the effects of different policy types within the same experimental framework, allowing for direct comparisons. Third, unlike most papers in this field, which typically inquire about hypothetical intentions to share, we delve into real decisions to share on Twitter within a high-stakes political context. Fourth, our paper bridges the gap between the empirical literature and the theoretical literature on the determinants of misinformation spread on social media (e.g., [Acemoglu et al., 2010, 2021](#); [Papanastasiou, 2020](#)). Fifth, our theoretical model and its estimation provide new insights into the motivation behind the sharing of political content on social media. Physiologists and media scholars have relied solely on survey evidence on the motivation for sharing (for a survey, see [Melchior and Oliveira, 2023](#)). Recent political economy research has considered two motives for generating social-media content systematically: attracting attention ([Srinivasan, 2023](#)) and finding justification for a stigmatized opinion ([Bursztyn et al., 2023](#)). Our paper highlights the importance of reputation and partisan motives for sharing. Finally, our analysis delivers direct policy implications about the design of optimal interventions and, due to its theoretical foundations, it arguably has validity outside the analyzed empirical setting. In an important recent study, [Guess et al. \(2023b\)](#) found that removing all reshared content from Facebook feeds for three months significantly decreased users’ knowledge but did not affect political attitudes. We show that the priming intervention changes the composition of reshared content toward accurate news and away from falsehoods, having little effect on the total amount of sharing. One should

expect this policy to have an even greater effect on the informativeness of voters than shutting down all reshares altogether and it could also influence attitudes. More research is needed to confirm this premise empirically.<sup>8</sup>

Our paper also relates to a growing literature on salience (for a survey, see [Bordalo et al., 2022](#)). We conceptualize participants’s decisionmaking as weighing different attributes of each tweet when deciding to share, similar to the approach taken by [Bordalo et al. \(2013\)](#). In contrast to [Bordalo et al. \(2013\)](#), in our setting these weights are endogenously determined as a function of fundamental preferences. The policy treatments influence these weights. Our contribution is in structurally estimating the effect of increased salience.

The rest of the paper is organized as follows. In [Section 2](#), we describe the design of our experiment. [Section 3](#) calculates the average treatment effects using reduced-form analysis. In [Section 4](#), we develop a game-theoretic model of sharing political information on social media. [Section 5](#) presents the results of the structural estimation of the model and counterfactual analyses. [Section 6](#) discusses policy implications. [Section 7](#) concludes.

## 2 Experimental Design

The experiment took place during the 2022 mid-term elections campaign in the U.S. The CINT online survey platform recruited participants among U.S. Twitter users for an online survey regarding political information online.<sup>9</sup> Initially, participants were contacted by email, and those who agreed to participate were provided with a link to the survey. Out of the 10,870 people who responded to the initial CINT invitation, only about 35% met the necessary criteria to participate in the survey, including having an active Twitter account and being a U.S. voter. Ultimately, 3,501 participants completed the entire survey experiment. As the sample was drawn from the subscribers of the CINT online platform who participate in surveys for pay, it cannot be considered representative of the overall population of social media users.

The survey experiment consisted of four parts: (1) the pre-treatment phase, (2) the presentation of four tweets with political content, (3) treatments and the sharing decision, and (4) the assessment of the tweets phase. We describe each of them below.

### 2.1 Pre-treatment phase

At the beginning of the survey, participants received a brief introduction that indicated the survey’s focus was on news consumption and circulation on Twitter. It also mentioned that only aggregate results would be published. The introductory page allowed participants to opt out at this stage. The authors’ institutional affiliations were not specified to prevent the appearance of potential ideological biases on the part of the experiment designers.

---

<sup>8</sup>More generally, we contribute to the literature on political and social effects of the internet and social media (for surveys of this literature, see [Zhuravskaya et al., 2020](#); [Persily and Tucker, 2020](#)). Particularly important recent contributions include: [Allcott et al. \(2020\)](#); [Guriev et al. \(2021\)](#); [Levy \(2021\)](#); [Allcott et al. \(2022\)](#); [Braghieri et al. \(2022\)](#); [Angelucci and Prat \(2023\)](#); [Nyhan et al. \(2023\)](#); [Guess et al. \(2023a\)](#).

<sup>9</sup>Our experiment took place before Twitter was rebranded as X.

Following this brief introduction, participants answered a series of questions regarding socioeconomic characteristics, including age, gender, education, employment status, and income, as well as political affiliation, past voting behavior, questions about what participants perceived to be the most important issues in the upcoming election, and information about the participants' Twitter usage, including the number of followers and the number of people they follow. Summary statistics for these pre-treatment characteristics are presented in the last two columns of Online Appendix Table A1.

## 2.2 The presentation of the tweets with political content

Participants were then shown, sequentially and in a random order, screenshots of four tweets containing political information. Two of these tweets contained true facts, while the other two contained false news. The texts of the true tweets were as follows: “The Biden administration has opened the application process for Americans seeking student debt relief.” and “Florida schools are ordered to provide bathrooms separated by biological sex.” The texts of the false tweets were: “Biden is adding more IRS agents to investigate your taxes than we have detectives investigating every crime in the country.” and “The Supreme Court has just voted to ban condoms.” Each tweet also included a picture, as shown in Figure 1. In addition to the screenshots, participants were provided with a link to each tweet on Twitter. 47% of the participants clicked on at least one of these links and observed that the tweets appeared on Twitter exactly as they did in the screenshots within the survey environment.

The four tweets were posted on a Twitter account named “2022 Political News” (*@News-Election22*) just before the experiment. We created this account in September 2022 for the purpose of the experiment. This account automatically reposts news from Twitter accounts belonging to the most popular liberal and conservative news media, as defined by Godel et al. (2021). Every three hours, the account retweeted the four most viewed political news posts from liberal media and the four most viewed political news posts from conservative media, ensuring there was no repetition. We promoted the account at its inception, and at the beginning of the experiment on October 31, 2022, at 12:00 AM, the account had 185 followers and 114 posts. Online Appendix Figure A1 presents how the account appeared on Twitter.

We selected four tweets for the experiment with the objective of covering both cultural and economic issues relevant to the election campaign. To begin, we chose two false political news from all the news that had been fact-checked and rated as false by PolitiFact in the past month. Our focus was on selecting brief news articles with associated pictures that could be used in a tweet without any modifications. Next, we selected the true news: one from those posted by conservative and the other by liberal media within the past month. We aimed to select news on topics similar to those of the two chosen false tweets. The four tweets included in the experiment were posted on the “2022 Political News” account without mentioning the source of these news.

## 2.3 Treatments and the sharing decision

The treatment phase, in which participants were subjected to different policy interventions, occurred after they had seen the four tweets. Following the treatment phase, participants in all treatment groups made their sharing decisions.

**Sharing decision:** Participants were asked if they wanted to share one of these tweets on their Twitter account. Specifically, within the survey environment, participants were presented with a screen listing the four tweets and asked if they wished to share one of them. They could choose one of the four tweets or opt not to share any. Those who chose one of the tweets were directed to a new survey page, where they had to confirm their choice by clicking a button that resembled Twitter’s retweet button. Clicking on it would take them to the selected tweet within the Twitter environment, where they could directly retweet it on their account.<sup>10</sup>

**Treatments:** Participants were randomly allocated to one of five treatment groups. This determined the type of screens they saw before making their sharing decision.

1. The **no policy** treatment group did not receive any special treatment. Participants in this group proceeded directly to making the sharing decision.
2. The **extra click** treatment differed from the no policy treatment in two ways. First, on the screen with the sharing question, it was indicated that participants would subsequently be asked to confirm their retweet choice on an additional screen. After the sharing screen, they were indeed asked for confirmation before being directed to the Twitter button.
3. The **prime fake news circulation** treatment presented participants with an additional screen before the sharing decision compared to the no policy treatment. This screen simply stated: “Please think carefully before you retweet. Remember that there are a lot of false news circulating on social media.”
4. The **offer fact check** treatment made participants view another screen before their sharing decision. The message on this screen read: “For some of these tweets, a fact-checking of the content has been done by PolitiFact, an independent fact checker. You can select the fact-checks you want to see. You can also choose to view none at all.” It listed the options: “See the fact check on ‘Supreme Court just voted to ban condoms;’” “See the fact check on ‘Biden is adding more IRS agents to investigate your taxes than we have detectives investigating every crime in the country;’” “Do not see any fact-check.” Participants could then choose to see one, both, or none of these fact checks. Those who decided to view a fact-check of a particular statement were directed to a separate page, which informed them that the statement had been judged false and provided a summary of the ruling by PolitiFact and a link to the full fact-checking article.

---

<sup>10</sup>The screens that participants had to view before they could share these tweets on their Twitter account were introduced to minimize the real-life impact of the experiment. In our previous work (Henry et al., 2022), we demonstrated that each additional click substantially reduces sharing, and we did not want to risk any of the false news involved in our experiment going viral. The responses to the sharing question within the survey environment were directly observed by us. The subsequent two clicks could be matched to the participants with certain probability. In this paper, we focus on the first sharing decision click as the main outcome variable.

5. The **ask to assess tweets** treatment did not add any screens or questions to the survey experiment compared to the no policy treatment but changed the order of the questions. Specifically, participants in this group were asked to evaluate the veracity and partisanship of each of the four tweets before making the sharing decision, whereas everyone else did this after making the sharing decision. We describe how the assessment of the tweets was performed in what follows.

## 2.4 Assessment of tweets phase

Participants were instructed to assess the content of each tweet in this last phase of the survey experiment.

**Veracity of tweets.**—The participants were asked to give their best guess of the probability that the content of each tweet is true, selecting a number between 0 and 100. To incentivize accurate estimates, we implemented the Binarized Scoring rule, which ensures maximum payments when reporting the true belief, regardless of risk aversion. Participants had the opportunity to earn an additional dollar for their accuracy. Following [Danz et al. \(2022\)](#), participants were informed that the payment rule is designed to maximize their expected earnings when reporting their most accurate guess, without showing the details of the rule.<sup>11</sup>

**Partisanship of tweets.**—Participants were also asked to assess whether each tweet, if true, favored Democrats or Republicans on a scale from 1 to 5. If participants were within one point of the average answer given by other participants for a randomly selected tweet, they obtained an additional 50 cents reward.

## 2.5 Balance across treatments and descriptive statistics

The randomization into treatment groups occurred at the time of participant recruitment through the CINT online platform. An equal number of invitations was sent to subscribers of the platform in the first four treatment groups, while only half of that number was sent in the fifth treatment group, which was the “Ask to Assess Tweets” treatment. This decision was based on two factors: a tight budget constraint and our expectation that this treatment would be particularly intellectually demanding, and therefore, unlikely to be very cost-effective. In each of the five treatment groups, 794, 783, 785, 786, and 353 subjects respectively completed the entire survey experiment.

**Balance.**—In Online Appendix Tables [A1](#) and [A2](#), we examine balance across treatments. Table [A1](#) presents the results of the balance tests, where we regress each of the pre-treatment characteristics one by one on the dummies for each treatment, with no policy as the comparison group. Table [A2](#) reports the results of an omnibus test of randomization quality. We regress the dummies for the treatment status in four subsamples that combine participants of the no policy (i.e., control) group with the participants of each of the four treatments on all the pre-treatment characteristics and test for their joint significance. We find that the predictive

---

<sup>11</sup>We provided a link to the full description of the payment rule. Only 4% of participants clicked on the link to view the rule.



power of the pre-treatment characteristics in determining the treatment status is very small. Overall, randomization worked very well, despite some imbalances across treatment groups. We cannot reject that these occasional imbalances are a result of random error: we find statistical significance at the 10% level in less than 10% of the estimated coefficients.

**Sharing.**—A significant number of participants found the content of the four tweets worth sharing. In the no policy group, 57.7% of respondents chose to share one of the four tweets in the first sharing decision screen. Furthermore, 13.9% clicked on the button that resembled the Twitter button, and 11.3% completed all the required steps to reshare one of the tweets on their Twitter account. In the absence of interventions, the engagement of participants with the false and true content in our experiment was similar. Specifically, 28.0% chose to retweet one of the two false tweets, and 29.7% chose one of the two true tweets.<sup>12</sup> However, it is important to note that, in addition to accuracy, there are many other qualities of news that affect how widely they are shared. We will take both the observed and potential unobserved determinants of sharing of news into account in our structural estimation.

**Veracity of tweets.**—Participants, on average, can distinguish between the veracity of true and false statements, but only a few are certain in their veracity assessments. The mean assessment of the probability that a false tweet is true is 36.3%, whereas the corresponding figure for a true tweet is 77.2%.<sup>13</sup> Only 4.14% of respondents indicated the veracity of all four tweets correctly with certainty. However, less than one percent of respondents deemed both false statements to be true and both true statements to be false with a probability greater than one-half. The confidence of respondents in the veracity of statements depends on the specific content. In the group with no policy, 52% of respondents are certain that the tweet about student debt relief is true, while only 24% believe the tweet about separate bathrooms is true. Similarly, 45% of respondents are confident that the tweet about a condom ban is false, but only 9% are certain that the tweet about IRS agents is false.

**Partisanship of tweets.**—Participants also seem to be rather effective at assessing the partisanship of true political content. The true tweets were selected so that one came from a liberal media source and the other from a conservative-leaning media source. Participants do rank them accordingly. On a scale from  $-2$  (very pro-Democrat) to  $2$  (very pro-Republican), the tweet from the liberal media about student debt relief received an average score of  $-1.15$ , which is between pro-Democrat and very pro-Democrat, and closer to pro-Democrat. In contrast, the tweet from the conservative media about bathrooms separated by biological sex received

---

<sup>12</sup>Specifically, 16.62% of respondents in the no policy group chose to retweet the tweet about the number of new IRS agents and 11.34% – the tweet about the ban on condoms, both of which are false. For the true tweets, 23% respondents in the no policy group chose to retweet the tweet about student debt relief and 6.7% – the tweet about bathrooms separated by biological sex.

<sup>13</sup>This is consistent with the results of [Angelucci and Prat \(2023\)](#), who do not directly ask for veracity estimates but require participants to select the true news among a set of news items. They found that the probability of choosing the true news correctly when facing one true and one false news item is 81%. When we use the veracity estimates of participants to do a similar exercise, we find a comparable figure of 74%. For all participants, we consider the four possible combinations of one true and one false tweet. For each combination, we measure whether the participant assigned a higher veracity score to the true tweet, which occurs in 74% of the cases.



an average score of 0.68, which is between neutral and pro-Republican, and closer to pro-Republican. Interestingly, when it comes to the two false tweets, participants do not agree on the political orientation of these tweets. On average, false tweets are judged as neutral, but this average masks a significant amount of disagreement among participants regarding their perceived partisanship.<sup>14</sup>

**Partisan alignment.**—We define partisan alignment between the participants and tweets as the product of the tweet’s partisanship (as perceived by the participant) and the score of the participant’s partisan orientation. The partisan orientation of the respondents is calculated by aggregating their responses to ten questions about their policy positions, designed to assess their partisan leanings. These questions were administered during the pre-treatment phase. On average, the respondents are slightly liberal-leaning. The median respondent scores a  $-5$  on a scale ranging from  $-30$  (very liberal) to  $30$  (very conservative). (We present the distribution of this variable across respondents in Online Appendix Figure A2.) By construction, partisan alignment is positive when the respondent leans conservative and the tweet is pro-Republican or very pro-Republican, or when the respondent leans liberal and the tweet is pro-Democrat or very pro-Democrat. This variable is negative when there is no alignment between the tweet and the respondent. We normalize the partisan alignment variable to range between  $-2$  (very non-aligned) and  $2$  (indicating strong alignment). Overall, in 38.5% of pairs of respondents and tweets, the alignment is greater than zero, in 44.7% it is below zero, and in the remaining 16.8%, it is exactly zero.

In the Online Appendix, we present summary statistics for the main variables of interest among respondents in the full sample and in the subsample of the no policy treatment group (see Table A3) and display the distributions of the veracity and partisanship estimates for each of the four tweets separately within the subsample of the no policy group (see Figure A3).

### 3 Average Treatment Effects

We begin by examining how different policies, emulated by the treatments, impact sharing behavior. Table 1 presents the estimation of average treatment effects (ATE) on sharing false and true tweets separately, as well as together. The first three columns consider the decision to share one of the two false tweets as the outcome, the next three columns focus on the decision to share one of the two true tweets, and the last two columns address the decision to share any one of the four tweets. For the first two outcomes, we provide three sets of results: without any controls, with socio-economic controls including race, gender, age, education, and employment status, and with additional controls for the respondent’s political affiliation, along with measures of Twitter usage intensity, the number of followers, and the number of people the respondent follows, in addition to the socio-economic controls. For the last outcome, which is essentially the sum of the first two outcomes in the corresponding specification, for brevity,

---

<sup>14</sup>We also asked participants to rank tweets based on their expected virality. We do not include this variable in the analysis because the participants were unable to predict the circulation of content. Their virality estimates appeared as good as pure noise and do not correlate with any other factors.

we present the results without any controls and with the full set of controls. We illustrate the results for the ATE on sharing false and true tweets in Figure 2.

The first three columns of Table 1 show that, compared to the no policy group, each treatment results in a reduction in the sharing of false tweets, although to varying degrees. In the no policy group, 28.0% of participants chose to share one of the false tweets. We arrange the treatments in order of their impact on the sharing of false tweets. The effects of the extra click treatment, priming fake news circulation treatment, offering fact-check treatment, and asking participants to assess tweets are as follows: 3.6, 11.5, 13.6, and 14.1 percentage-point decreases in the sharing of false tweets, respectively, relative to the no policy group. These magnitudes come from the specification with socio-economic controls. When comparing the coefficients across the first three columns, it is evident that the point estimates hardly change from one specification to another, confirming the effectiveness of randomization. Importantly, the differences in the magnitude of the effects for the treatments of priming fake news circulation, offering fact-checks, and asking participants to assess tweets are not statistically significant. However, the effect of each of these treatments is significantly greater in magnitude than that of the extra click treatment.

The following three columns reveal that the impact of treatments on the sharing of true tweets is very different. The sharing rate of true tweets is 29.7% in the no policy group. The effect of the extra click treatment on the sharing of true tweets is a precisely estimated zero. In contrast, the priming treatment significantly increases the sharing of true tweets by 8.1 percentage points relative to the no policy group. The offer fact-check treatment, in contrast, significantly reduces the sharing of true tweets by 7.8 percentage points. The ask to assess tweets treatment has a small, statistically insignificant negative effect on the sharing of true tweets.

The last two columns of the table depict the effects of treatments on aggregate sharing. We find that the extra click and priming treatments do not exert a significant impact on overall engagement with the political content, whereas the other treatments lead to a significant decrease in engagement. Importantly, the reasons behind the null effect of the extra click and priming treatments on aggregate sharing differ markedly. The extra click treatment has a minimal effect, whereas the priming of fake news circulation has substantial and opposing effects on the sharing of false and true news. Therefore, we conclude that the priming fake news circulation treatment is the most effective in altering the balance of shared news toward the true news without significantly affecting overall sharing.

These findings align with existing literature on the effects of individual policies. According to Henry et al. (2022), offering access to fact-checking reduces the sharing of false news on Facebook by 45%. In our experiment, we observe a similar impact of the fact-check treatment, which reduces sharing by 48.5% (a 13.6 percentage point decrease from the baseline level of 28%). Our ask to assess tweets treatment is conceptually similar to interventions studied by Fazio (2020) and Pillai and Fazio (2023), who consider a prompt “Please explain how you know that the headline is true or false” before asking whether people wants to share a news headline.

These papers find a reduction in sharing of false headlines by about 30% and no effect on true headlines.<sup>15</sup> Our intervention asks individuals to assess the content and provide estimates of accuracy and partisanship, resulting in a slightly larger 49% reduction in the sharing of false content and no effect on sharing true content. Our priming fake news circulation treatment reduces the sharing of false tweets by 41% (an 11.5-percentage-point reduction compared to the 28.0% sharing rate in the no policy group). Simultaneously, it increases the sharing of true tweets by 27% (an 8.1-percentage-point increase from the baseline of 29.7%). [Pennycook and Rand \(2022\)](#) provide an overview of the effects of twenty different studies on the use of “accuracy prompts,” which are conceptually similar to our priming treatment. In all of them, accuracy prompts shift the balance of circulated content toward true news. In all cases, these priming interventions led to reduced sharing of false news, and in about half of them, they had a positive, albeit often imprecise, effect on the sharing of true news. [Pennycook and Rand \(2022\)](#) advocate the use of accuracy prompts because they are effective and cost-efficient, i.e., they do not require identifying true and false content. Our unified approach to studying different policies in the same setting adds another compelling reason in favor of priming interventions: social media platforms, concerned with overall engagement, should be more inclined to adopt such policies.

The core of the paper is dedicated to understanding the mechanisms underlying the results above. The treatments can influence outcomes through three distinct channels. First, they may impact the potential sharer’s assessments of pertinent content characteristics, such as accuracy and partisan leaning. Second, they can alter the prominence of various motives for sharing, particularly by increasing the salience of reputation concerns. Lastly, by introducing frictions, they may directly raise the sharing cost. To shed light on the role of each of these channels, we develop and estimate a structural model.

## 4 Theoretical Model

We build a model of sharing in Section 4.1 which, starting from general assumptions about preferences of senders, determines how the utility of sharing depends on the key variables we measure experimentally, namely veracity and partisan alignment. In Section 4.2 we use the framework of our experiment to develop a structural approach to estimation of the factors influencing the equilibrium level of sharing in our model. In Section 4.3 we describe how the structural estimation sheds light on the mechanisms underlying the treatment effects.

### 4.1 A model of sharing

The binary state of the world  $\theta \in \{0, 1\}$  is unknown.  $\theta = 1$  is a state of the world in which Republican actions are optimal and  $\theta = 0$  is a state of the world in which Democratic actions are optimal. A sender (indexed by  $i$ ) has access to a piece of content (tweet) indexed by  $j$ . The

---

<sup>15</sup>In [Fazio \(2020\)](#), 57% of participants in the control group indicated they would be “likely,” “somewhat likely,” or “extremely likely” to share at least one false headline, as opposed to 39% in the treatment group. For the category “extremely likely,” the effect dropped from 24% to 17%.

tweets are potentially informative about  $\theta$ . The sender decides whether to reshare content  $j$  to receivers indexed by  $k$ .

#### 4.1.1 Characteristics of content

From the point of view of sender  $i$ , each content  $j$  is characterized by veracity and partisanship.

**Content Veracity** ( $\nu_{ij}$ ).— Content  $j$  can be either truthful or misinformation; the actual veracity of the content is denoted as  $\omega_j \in \{0, 1\}$ . Initially, both the sender and the receivers share the same prior belief about the veracity of  $j$ :  $P[\omega_j = 1] = \nu_j^0$ . This prior belief may have a common component applicable to all news, such as the prevalence of fake news on social media, and a tweet-specific component, which relates to the appearance and wording of the content. Before individuals take actions, they may receive information and update their initial beliefs regarding the veracity of the content, as explained below. Veracity is, therefore, individual- and content-specific; we denote it as  $\nu_{ij}$ .

**Content Partisanship** ( $\pi_{ij}$ ).— Content  $j$ , conditional on being true, can be more or less aligned with a particular worldview. Specifically, we assume that sender  $i$  perceives that tweet  $j$  will be interpreted by a given receiver with probability  $\pi_{ij}$  as suggesting that the state of the world is 1 (favoring Republican action), while with probability  $1 - \pi_{ij}$ , the tweet will be viewed by the receiver as a signal of state 0 (favoring Democratic action). We assume that  $\pi_{ij}$  is distributed according to a cumulative distribution function  $F(\cdot)$  with the mean of  $1/2$  and a symmetric density.

#### 4.1.2 Characteristics of individuals

**Individual Information** ( $\psi_l$ ).— Senders and receivers can be of two types: “informed” ( $I$ ) with probability  $q$  and “uninformed” ( $U$ ) with probability  $1 - q$ . We denote the type of each individual  $l$  as  $\psi_l \in \{I, U\}$ . Informed individuals know the veracity  $\omega_j$  of content  $j$ . For instance, they may consistently follow information from mainstream media to stay informed or have a habit of fact-checking the information they receive. Consequently, for an informed sender  $i$ , if the content is true,  $\nu_{ij} = 1$ , and if it is false,  $\nu_{ij} = 0$ . In other words, informed individuals have a discerning judgment: they know exactly whether the news is true or false. Uninformed individuals, in contrast, do not receive any information about the veracity of news. They keep their prior  $\nu_j^0$ .

**Individual Partisanship** ( $\pi_l$ ).— Senders and receivers are characterized by their partisanship, which we define as the prior belief regarding the state of the world, denoted as  $\pi_l$  for an individual  $l$ . As demonstrated below, a key determinant of the sender’s behavior is the partisan alignment between the partisanship of sender  $i$  ( $\pi_i$ ) and their perceived partisanship of the news  $j$  ( $\pi_{ij}$ ). We define the partisan alignment between  $i$  and  $j$  as:

$$\pi_{ij}^a = \left( \pi_i - \frac{1}{2} \right) \left( \pi_{ij} - \frac{1}{2} \right).$$

The partisan alignment is positive if and only if the news  $j$  and sender  $i$  have the same partisanship (both Democrat or both Republican). To simplify the notation, we assume that  $\pi_i \in \{0, 1\}$ . In other words, the sender either believes that the state of the world calls for Democratic action or believes that the state of the world is such that Republican action is optimal. We further assume that both of these outcomes are equally likely.<sup>16</sup>

### 4.1.3 Preferences of senders and receivers

**Receivers.**—Receiver  $k$  chooses action  $a_k \in [0, 1]$  to maximize  $-(a_k - \theta)^2$ , so in equilibrium, they choose the action that conforms with their belief about the state of the world,  $a_k = E[\theta]$ . The belief about  $\theta$  is the updated belief of the receiver based on any information they might have received.

**Senders.**—We describe the utility of sender  $i$  from sharing content  $j$ . Let us denote the decision to share  $s_{ij} \in \{0, 1\}$ . If they share ( $s_{ij} = 1$ ), they receive a content-specific payoff  $\xi_j$ , which can be positive or negative, capturing, for instance, the entertainment or surprise value of content  $j$ . They also gain three additional sources of payoffs, each with its own weight in the sender's utility function:

- (1) A persuasion payoff comes from influencing others' actions to align with the sender's partisan beliefs. The weight of the persuasion payoff is denoted as  $\mu_p$  in the utility function. Here,  $p$  stands for persuasion.
- (2) A signaling partisanship payoff comes from others correctly identifying the sender's partisan beliefs. The signaling partisanship payoff has a weight of  $\mu_s$  in the utility function. In this case,  $s$  stands for signaling.
- (3) A reputation payoff represents the benefit of having an image of being informed. Its weight in the utility function is denoted as  $\mu_r$ . Here,  $r$  stands for reputation.<sup>17</sup>

Overall, the sender's utility is as follows:

$$\begin{aligned}
 V_{ij} = \xi_j &+ \mu_p (2\pi_i - 1) \left( E[a_k | s_{ij} = 1] - \frac{1}{2} \right) \\
 &+ \mu_s (2\pi_i - 1) \left( E[R | s_{ij} = 1] - \frac{1}{2} \right) \\
 &+ \mu_r (E[\psi_i = I | s_{ij} = 1] - q). \tag{1}
 \end{aligned}$$

The first two components in the formula above correspond to the two partisan motives: partisan persuasion and signaling partisanship, respectively. The third component represents reputation concerns.

We define the persuasion payoff in this way because a Republican sender aims to maximize the action  $a_k$  taken by the receivers, while a Democrat sender seeks to maximize  $1 - a_k$ . Therefore, sender  $i$  with partisanship  $\pi_i$  aims to maximize  $(2\pi_i - 1)a_k$ . Similarly, the signaling

<sup>16</sup>The results presented in Proposition 1 below also extend to a case with continuous sender's partisanship  $\pi_i$  as long as its distribution is symmetric with a mean of  $1/2$ .

<sup>17</sup>We compute these images following Benabou and Tirole (2011).

partisanship payoff for a Republican is positive if the receivers become more confident that the sender is a Republican ( $E[R|s_{ij} = 1] - \frac{1}{2} > 0$ ) and negative if they perceive the sender as a Democrat. It is worth noting that all payoffs are defined as deviations from their mean values:  $\frac{1}{2}$  for the receivers' actions and beliefs and  $q$  for the sender's reputation.

We assume that the receivers are aware of all parameters in the utility function, except for the values of  $\pi_{ij}$ ,  $\nu_{ij}$ ,  $\pi_i$ , and  $\psi_i$ .

#### 4.1.4 Equilibrium

The partisan persuasion, signaling partisanship, and reputation payoffs are determined in equilibrium. Receivers are aware that the decision to share depends on whether the sender is informed or uninformed and, in the case that the sender is informed, it depends on whether the message is true ( $\omega_j = 1$ ) or false ( $\omega_j = 0$ ). We define the state of the sender as  $\sigma \in \{0, U, 1\}$ , such that  $\sigma = U$  if the sender is uninformed and  $\sigma = \omega_j$  if they are informed. The following proposition characterizes the sender's utility in equilibrium.

**Proposition 1.** *In all Perfect Bayesian Nash Equilibria of the game, senders share if and only if the partisan alignment between them and the news ( $\pi_{ij}^a$ ) is high. Republican leaning senders share if and only if  $\pi_{ij} > \hat{\pi}_\sigma$ . Democratic leaning senders share if and only if  $\pi_{ij} < 1 - \hat{\pi}_\sigma$ . The cutoff  $\hat{\pi}_\sigma$  depends on the state  $\sigma$ .*

Moreover, in equilibrium, sender  $i$ 's indirect utility derived from sharing  $j$  can be expressed as:

$$V_{ij} = \alpha_0^j + \alpha_1 \nu_{ij} + \alpha_2 \pi_{ij}^a + \alpha_3 \nu_{ij} \pi_{ij}^a, \quad (2)$$

where:

$$\begin{aligned} \alpha_1 &= \mu_r q r_1 \\ \alpha_2 &= \mu_p 2(1 - q) \tilde{\nu}_j + \mu_s (-2 + 4((1 - q)s_U + qs_0)) \\ \alpha_3 &= \mu_p 2q - \mu_s 4q(s_0 - s_1) \end{aligned}$$

and  $\alpha_0^j$  is a news-specific constant.

The indirect utility is increasing in perceived veracity  $\nu_{ij}$  and partisan alignment  $\pi_{ij}^a$ :  $\alpha_1 > 0$  and  $\alpha_2 > 0$ . Furthermore, if the partisan persuasion motive is sufficiently important compared to the signaling partisanship motive ( $\mu_p > \mu_s 2(s_0 - s_1)$ ), then the indirect utility is also increasing in the interaction  $\nu_{ij} \pi_{ij}^a$ , namely,  $\alpha_3 > 0$ , and  $\hat{\pi}_0 \geq \hat{\pi}_U \geq \hat{\pi}_1$ .

**Proof:** See Online Appendix B, where we characterize the equilibrium conditions, derive the equilibrium values of  $\tilde{\nu}_j$ , the receivers' beliefs about the sender's perception of veracity given the content that is shared,  $s_\sigma$ , the receivers' beliefs about the sender's partisanship conditional on the state of the receiver  $\sigma \in \{0, U, 1\}$ ,  $r_1$ , the receivers' beliefs about whether the sender is informed conditional on the state of the receiver being 1, and discuss the existence of equilibria.

Proposition 1 characterizes the sender's utility of sharing as a function of their estimates of veracity and partisan alignment, both of which we measure in the data for all combinations

of participants and tweets, and a tweet fixed effect ( $\alpha_0^j$ ), which is a measure of the unobserved appeal and shareability of each specific piece of news. For instance,  $\alpha_0^j$  gauges how well the tweet  $j$  is presented or how sensational it is. In what follows, we provide the intuition for these results. Even though different payoffs are interconnected in equilibrium, we discuss them separately.

The reputation payoff, corresponding to the updated belief of the receiver that the sender is informed given the content they shared, increases with  $\nu_{ij}$ . If a receiver is informed, they are more likely to realize that the shared news was false when the perceived veracity is low. Since false news is less likely to be shared by an informed sender, the informed receiver will update negatively their belief that the sender is informed when they receive false news. This, in turn, decreases the reputation payoff.

The sharer's persuasion payoff increases with the partisan alignment between the sender and the news,  $\pi_{ij}^a$ , because aligned news are more likely to influence receivers in the desired direction. Perceived veracity also matters for persuasion. If the receiver is informed, they will act on information  $j$  if and only if the shared content is true. If the content is false, the receiver will not update their belief about the state of the world. If the sender believes that the news is more likely to be true, they expect it to be more persuasive for receivers. Therefore, the persuasion payoff increases with both partisan alignment  $\pi_{ij}^a$  and the interaction between partisan alignment and perceived veracity  $\nu_{ij}\pi_{ij}^a$ .

The signaling partisanship payoff also increases with partisan alignment,  $\pi_{ij}^a$ , because the receiver updates their beliefs about the sender's partisan affiliation based on the partisanship of the received news. The strength of this updating effect also depends on perceived veracity,  $\nu_{ij}$ . In contrast to persuasion, higher veracity makes the effect of partisan alignment weaker for signaling partisanship, i.e., the interaction  $\nu_{ij}\pi_{ij}^a$  has a negative effect on the signaling partisanship payoff. The reason for this is that if the news is less likely to be true, it sends a more credible signal of the sender's partisanship. If the receiver is informed and they receive a very partisan false message, then they are more likely to conclude that the sender is partisan. This is because sending a false message is otherwise costly for the sender. Thus, both partisan motives imply that sender's indirect utility increases in partisan alignment, but the sign of the effect of the interaction between partisan alignment and perceived veracity depends on which of the two partisan motives is stronger.

Senders only share content that is sufficiently aligned with their partisan beliefs. Proposition 1 establishes the existence of a cutoff value for partisan alignment, above which the news is shared ( $\hat{\pi}_\sigma$ ). This cutoff value depends on whether the sender is informed and, if so, on the veracity of the news (i.e., on the state  $\sigma$ ). For instance, when the sender is informed and the content is false ( $\sigma = 0$ ), the sender knows that the news is likely to be less persuasive. This means that it needs to be more aligned with the sender's beliefs for them to want to share it, compared to the situation when the sender is uninformed ( $\sigma = U$ ). Therefore, a Republican sender has the following ranking of cutoffs:  $\hat{\pi}_0 \geq \hat{\pi}_U \geq \hat{\pi}_1$ . (A Democratic sender's ranking is the opposite.)



## 4.2 Structural model: The nested random utility model of sharing

The purpose of our structural analysis is to estimate the parameters of the indirect utility of sharing, as derived in Proposition 1. In our experiment, participants could either share one of the four tweets or choose not to share anything. We decompose the participant’s choice into two sequential (“nested”) decisions. First, in the upper nest, the decision  $m \in \{s, n\}$  pertains to whether they want to share any of the tweets available ( $B_s = \{1, 2, 3, 4\}$ ) or opt not to share anything ( $B_n = \{0\}$ ). Second, if the decision in the upper nest is to share something ( $m = s$ ), then, in the lower nest, the individual selects which content  $j \in \{1, 2, 3, 4\}$  to share.

The utility derived by the individual when deciding whether to share content on social media and which content to share can be expressed as the sum of two terms: the upper-nest component  $W_{im}$ , which depends solely on the decision  $m \in \{s, n\}$  of whether to share at all, and the lower-nest component  $V_{ij}$ , which varies across alternatives  $j$  within the lower nest if  $m = s$ . The sender’s total utility is thus given by:

$$W_{im} + V_{ij} + \epsilon_{ijm}, \quad (3)$$

where  $\epsilon_{ijm}$  represents the unobserved shock to the utility experienced by individual  $i$  when choosing alternative  $j$  from the set  $\{1, 2, 3, 4\}$  if  $m = s$ , and when deciding whether to share anything or not, i.e., choosing  $m$  from the set  $\{s, n\}$ .

The lower-nest decision regarding which content to share depends on the characteristics of the individual tweets. In contrast, the upper-nest decision is determined by the comparison between the overall cost of sharing and the maximum potential gain that can be achieved by making the optimal choice in the lower nest.<sup>18</sup> The sender’s utility is defined up to a constant, which means that only the relative levels of the determinants of sharing matter when making the decision to share. Therefore, without loss of generality, we normalize  $W_{is} = 0$  and  $V_{i0} = 0$ . If the sender chooses not to share at all, their utility is  $W_{in}$ . If they decide to share one of the tweets,  $j \in \{1, 2, 3, 4\}$ , their utility is  $V_{ij}$ .

Given that the recipients of content shared by the participants in our experiment were unaware that the senders were involved in an experiment, it is reasonable to assume that they did not modify the inferences they made about the sender based on the fact a tweet was shared. Thus, our model is applicable to the choices made by the experiment’s participants. Therefore, equation (2) is a suitable representation of  $V_{ij}$  for  $j \neq 0$ .

In the course of the experiment, we collected participants’ assessments of the tweets’ veracity and partisanship. We also inquired about the partisan inclinations of the participants themselves in the pre-treatment set of questions. Consequently, we have the measurements

---

<sup>18</sup>This structural modeling approach is similar to [Bugge et al. \(2023\)](#), who study the nested decision of a refugee considering emigration. In their model, the lower-nest decision involves choosing a destination and depends on pull factors, namely, the characteristics of the destination. Conversely, the upper-nest decision revolves around the choice of whether to leave or stay, and it depends on push factors, such as the threat of persecution at home. In the upper nest, the individual evaluates the overall cost of leaving their home against the benefits that can be obtained in the best possible destination.

of  $\nu_{ij}$  and  $\pi_{ij}^a$  for all participants and all tweets. This allows us to estimate the parameters  $\alpha_1$ ,  $\alpha_2$ , and  $\alpha_3$  in equation (2) from the experimental data. To accomplish this, we follow the methodology outlined by [Anderson et al. \(1992\)](#) and [Train \(2009\)](#), assuming that  $\epsilon_{ijm}$  follows a Generalized Extreme Value (GEV) distribution. Under this assumption, the probability of selecting alternative  $j$  can be represented as the product of two probabilities: the probability of sharing content  $j$ , given that sharing occurs, and the probability of sharing anything at all (see [Anderson et al., 1992](#); [Cameron and Trivedi, 2005](#); [Train, 2009](#); [Bugge et al., 2023](#)). In particular, for  $j \neq 0$ , we derive an expression for the probability of sharing content  $j$ :

$$P_{ij} = P(j|s)P_s, \quad (4)$$

where  $P_s$  is the probability of sharing anything. As in [Train \(2009\)](#), the conditional probability  $P(j|s)$  is given by:

$$P(j|s) = \frac{e^{V_{ij}/\lambda_w}}{\sum_{l \in \{1, \dots, \mathcal{J}\}} e^{V_{il}/\lambda_w}}, \quad (5)$$

where the parameter  $\lambda_w$  captures the within-lower-nest heterogeneity of the error term  $\epsilon_{ijm}$  and  $V_{ij}$  is defined by equation (2). Equation (5) has a straightforward and intuitive interpretation: it informs us that the probability of sharing a specific content  $j$  increases when the utility of sharing that content, which depends on the content characteristics  $\nu_{ij}$  and  $\pi_{ij}^a$ , is higher compared to the utility of sharing of other, alternative content. The denominator in equation (5) represents the inclusive utility, which is the expected utility of sharing the best alternative among the  $\mathcal{J} = 4$  available pieces of news, specifically the four tweets in the lower nest:

$$I_{is} \triangleq \ln \left( \sum_{l \in \{1, \dots, \mathcal{J}\}} e^{V_{il}/\lambda_w} \right). \quad (6)$$

The total probability of sharing can then be expressed as:

$$P_s = \frac{1}{1 + \exp \left( \frac{W_{in}}{\lambda_b} - \frac{\lambda_w}{\lambda_b} I_{is} \right)}, \quad (7)$$

where the parameter  $\lambda_b$  represents the between-nest component of the error term  $\epsilon_{ijm}$ .<sup>19</sup> Equation (7) demonstrates that the probability of sharing decreases with the utility of non-sharing ( $W_{in}$ ). For instance, sharing should decrease if the cost of sharing, such as sharing friction, increases. In addition, sharing increases with  $I_{is}$ , which represents the payoff from the best possible alternative among  $j > 0$ .

This characterization enables us to bring the model to the data. In the lower nest, we estimate equation (5) using a Conditional Multinomial Logit ([McFadden, 1973, 1974](#)). Within the sample of those who opted to share something, we regress the discrete choice among the

---

<sup>19</sup>For a detailed discussion of the distributional assumptions on  $\lambda_w$  and  $\lambda_b$ , refer to Appendix H.1 of [Bugge et al. \(2023\)](#). In their notation, these parameters are called  $\lambda_2$  and  $\lambda_1$ , respectively.

four tweets on the veracity and partisan alignment estimates as well as their interaction for each participant, as guided by equation (2). Subsequently, we use this estimation to predict  $V_{ij}$  for each  $i$  and  $j$  in this sample. This step allows us to compute the inclusive utility  $I_{is}$  by employing (6) and to estimate the determinants of the lower-nest choice, as represented by equation (7). Conveniently, this equation exhibits a logistic functional form, which we can estimate using a logistic regression. Although we lack a direct measure of the cost of sharing, denoted as  $W_{in}$  in equation (7), we can infer differences in  $W_{in}$  across treatments. This is possible because, unlike the cost of sharing, which is affected by the treatments, the unobserved parameter  $\lambda_b$  should not differ systematically across treatments. This is due to the randomization and balance.

Overall, the structural estimation of equilibrium sharing behavior using equations (2) and (5) to (7), which takes  $\nu_{ij}$  and  $\pi_{ij}^a$  as inputs, yields estimates of  $\alpha_1, \alpha_2, \alpha_3, I_{is}, \frac{W_{in}}{\lambda_b}, \frac{\lambda_w}{\lambda_b}$ . In the next section, we discuss how the experimental treatments affect the structural model's parameters.

### 4.3 Structural impact of the treatments

In this section, we consider the mechanisms behind the effects of different treatments on sharing behavior. Importantly, the receivers of the tweets shared by the participants in our experiment were unaware of the experiment's existence. In particular, they did not know that the senders of these tweets were subjected to any treatments. This is why we rule out the following two mechanisms. First, the treatment status of participants should not affect the functional form of the persuasion and reputation payoffs derived in Proposition 1. Second, the treatment effects should not result from a switch between equilibria, in case there are multiple equilibria.

We identify three potential channels through which treatments affect the sharing decision of senders: the updating channel, the salience channel, and the cost of sharing channel. We describe them below and, in the next section, we decompose the total average treatment effects into the contributions of these different channels.

**Updating channel:** The treatments can influence the senders' assessments of the veracity and partisanship of tweets. Specifically, providing fact-checking information is designed to modify social media users' beliefs about the accuracy of specific content. Priming fake news circulation may also lead to changes in beliefs about the average spread of false information, potentially causing users to update their estimates of the accuracy of individual tweets. A change in the veracity and partisan alignment of a particular tweet should consequently result in a change in the indirect utility of sharing of that content ( $V_{ij}$ ), as described in Proposition 1. As  $V_{ij}$  enters equations (5) and (7), this will impact the sharing behavior.

Consider a situation in which some participants verify the fact-checking information in the offer fact-check treatment and update their estimates of the veracity of the two false tweets. According to the updating channel, this should lead to a reduction in the sharing of these false tweets in the lower nest. This is because the indirect utility of sharing increases with perceived veracity. Furthermore, it should also decrease sharing in the upper nest as the inclusive utility is also negatively affected.

**Salience channel:** It is commonly believed that nudges, such as the one used in the priming fake news circulation treatment, function by increasing the salience of certain drivers of behavior (as demonstrated by Thaler and Sunstein, 2008; Bordalo et al., 2022). In our context, treatments could produce such behavioral effects by increasing the salience of reputation concerns in comparison to partisan motivations for sharing. This would imply an increase in  $\mu_r$  relative to  $\mu_p$  and  $\mu_s$  (change in weights as in Bordalo et al., 2013), leading to an increased sharing of news believed to be true by the participants and a reduced sharing of news believed to be false, provided that sharing occurs. The overall impact on sharing due to the salience effect will depend on how it influences the inclusive utility. One should expect the salience channel to be at work not only in the priming fake news circulation treatment, but in any treatment that prompts individuals to contemplate the consequences of sharing false news.

**Cost of sharing channel:** Finally, the treatments could introduce frictions that make sharing less appealing. In the model, this would translate into an increase in  $W_{in}$ , which represents the relative upper-nest utility of not sharing compared to sharing. For instance, treatments could induce fatigue, whether from processing the information contained in fact-checks or from the efforts required to assess content quality. This fatigue might lead individuals to opt for disengagement rather than sharing. A similar effect could occur if users feel that they have already invested too much time and energy online and decide to disengage as a result. Importantly, the cost of sharing channel only works through the upper nest, because it does not affect  $V_{ij}$ . We can estimate the relative size of  $W_{in}$  across treatments under the assumption that  $\lambda_b$  does not differ systematically across treatments: the coefficients on the treatment dummies in logistic estimation of equation (7) should give us the treatment-specific estimates of  $W_{in}/\lambda_b$ .

In the next section, we estimate these mechanisms separately and evaluate their relative importance.

## 5 Structural Estimation

### 5.1 Estimating the channels in the structural model

#### 5.1.1 Updating channel

The first channel through which treatments can affect sharing is the updating channel. Our treatments can impact participants' perceptions of the veracity and partisanship of tweets. We illustrate the average treatment effects on respondents' estimates of the veracity and partisan alignment of both false and true tweets in Figure 3. The figure presents the unconditional means of the respondents' estimates by treatment.

A detailed regression analysis with controls for socio-demographic characteristics of respondents is presented in Table 2. The unit of observation is a respondent-tweet pair. The first four columns use veracity estimates (scaled between 0 and 1 by dividing by 100) as the dependent variable; and in the last four columns, the dependent variable is partisan alignment. For each dependent variable, in the first column, we consider the sample of false tweets; the second column focuses on the sample of true tweets; and the last two columns present regressions on

the full sample. Our main explanatory variables are the treatment dummies. The regressions are estimated with OLS and we allow for clusters at the level of each respondent. At the bottom of the table, we present the means of dependent variables in the no policy treatment group as well as their standard deviations in the full sample.

We find that veracity estimates of false tweets are significantly negatively affected by two out of four treatments: the priming fake news circulation and the offer fact check treatments (see columns (1) and (4)). The priming treatment reduces the veracity estimates of false tweets by 5.3 percentage points and the offer fact check treatment decreases them by 6.2 percentage points from the mean level of 38.9% in the no policy group. These effects constitute 15 and 17 percent of the standard deviation of the veracity of false tweets, respectively. The extra click and ask to assess tweets have a precisely estimated zero effect on the veracity estimates of false tweets. As far as the effect of the treatments on the veracity of true tweets is concerned, only the priming fake news circulation treatment has a statistically detectable effect: this treatment reduces the veracity estimates of true tweets by 1.7 percentage points from the mean of 78% in the no policy group (see columns (2) and (4)). The effect, however, is very small in magnitude. The other treatments do not affect veracity estimates of true tweets: the point estimates of these effects are precisely estimated zeros.

What is the intuition behind these effects? The fact-checking treatment offered access to fact-checks for two tweets, showing that they were false. The negative updating of participants for the veracity of these two tweets is, therefore, natural.<sup>20</sup> It is interesting that this treatment did not make participants update the veracity of the other two tweets. Specifically, we find no evidence on the implied truth effect identified by [Pennycook et al. \(2022\)](#).

The fact that the priming treatment affects the veracity estimates of false tweets much more than of true tweets is remarkable because the treatment just warns the participants that false information circulates on social media, without any further details. Even though the average updating for false news is not statistically different between the offer fact-check and priming treatments (the p-value for the test of the difference in the effects of the two treatments is 0.45), [Online Appendix Table A4](#) shows that participants are significantly more likely to perfectly assess the veracity of all four tweets in the offer fact-check treatment group compared to no policy, and this is not the case for the priming fake news circulation treatment group. Naturally, the other treatments do not have any effect because they are not designed to make participants update their priors about the veracity of content.

Columns (5) to (8) in [Table 2](#) demonstrate that three treatments, namely, priming fake news circulation, offering fact-check, and asking participants to assess tweets, have statistically significant negative effects on political alignment. The magnitudes of the coefficients on the dummies for each of these treatments are as follows:  $-0.058$ ,  $-0.047$ , and  $-0.052$  (column (5)).

---

<sup>20</sup>Out of 785 participants in the offer fact-check group, 187 chose to access both available fact-checks, 144 chose to access only the fact-check of the tweet about IRS agents, and 243 clicked to access only the fact-check of the tweet about ban on condoms. We do not use this information in the analysis because the decision to see the fact-check is endogenous and we want to focus on the exogenous average effect of the offering fact-check treatments (see [Henry et al. \(2022\)](#) for the analysis of the determinants of the decision to access fact-checking).

These values are relatively small compared to the range of alignment, which spans from  $-2$  to  $2$ , corresponding to 10%, 8%, and 9% of the standard deviation of this variable, respectively.<sup>21</sup> In contrast, all treatments have precisely estimated zero effects on the political alignment of true tweets. The exact reason why participants subjected to these treatments perceive false tweets as less politically aligned is not clear. It is possible that, as a result of the treatments, participants want to distance themselves from false information.

### 5.1.2 Salience channel (lower nest)

In Table 3, we present the results of the structural estimation for the lower-nest choice. We employ a Conditional Logit Choice Model to estimate the relationship between sharing a specific tweet and the perceived veracity and political alignment of that tweet, allowing for an interaction effect between them. More precisely, we estimate equation (5) with  $V_{ij}$  substituted from equation (2). The unit of observation, once again, is a participant-tweet pair. Since the lower-nest relationship is conditional on sharing, we focus our analysis on participants who shared one of the four tweets. The first five columns of the table present the results separately for each treatment group, while column (6) presents the results for all treatment groups combined. As above, we control for socio-economic characteristics of respondents.

First, we observe that the parameters  $\alpha_1$ ,  $\alpha_2$ , and  $\alpha_3$  are consistently positive. This holds true for each parameter within every sample. All of these estimated parameters are statistically significant, with two exceptions (namely,  $\alpha_2$  in the priming treatment subsample and  $\alpha_3$  in the ask to assess tweets subsample). In accordance with Proposition 1, the fact that perceived veracity, partisan alignment, and the interaction between these two variables all increase sharing indicates the importance of the reputation motives and at least one of the partisan motives. Furthermore, while signaling partisanship may play a role as a driver of sharing, it is outweighed by the partisan persuasion motive (i.e., the condition  $\mu_p > \mu_s 2(s_0 - s_1)$  in Proposition 1 is satisfied). This is because  $\alpha_3$ —the coefficient on the interaction between veracity and partisan alignment—is positive and significant. These findings provide new insights into the motivations behind sharing on the social media.

One could also imagine other motives for sharing that are not considered by our model. For instance, some people may share to attract attention, meaning that tweets which more likely to achieve this would be shared more (Srinivasan, 2023). Attention-grabbing of tweets, along with other unobserved characteristics that may also affect sharing, is absorbed by the tweet fixed effects in our estimation. Due to randomization, we expect the share of people who have these and other unobserved motivations for sharing to be balanced across treatments.

Columns (1) to (5) show that the treatments affect the relative sizes of  $\alpha_1$ ,  $\alpha_2$ , and  $\alpha_3$ , which, in turn, are functions of the parameters of the utility function, as stated in Proposition 1. We derive and solve numerically the system of equations, which allows us to recover the

---

<sup>21</sup>Due to randomization, there are no significant differences in the scores of political orientation among respondents across treatments, which are multiplied by partisanship of the tweets to calculate political alignment. This implies that these are causal effects of the treatments.



relative weights of different motives in the utility function of the sender ( $\mu_r$ ,  $\mu_p$ , and  $\mu_s$ ) for each treatment separately using the estimates of  $\alpha_1$ ,  $\alpha_2$ , and  $\alpha_3$  from Table 3. We show that treatments significantly impact the values of  $\mu_r$ ,  $\mu_p$ , and  $\mu_s$ , providing strong evidence of a salience effect.

Specifically, the equilibrium is characterized by a system of ten nonlinear equations with constraints on parameters. We present this system in Online Appendix C. In particular, two equations in this system (C.1 and C.2) characterize  $\hat{\pi}_U$  and  $\hat{\pi}_1$ , i.e., the values of  $\pi_i$  such that senders of type  $U$  and 1 are indifferent between sharing and not sharing.<sup>22</sup> Another two equations (C.3 and C.4) characterize the image  $r_\sigma$  ( $\sigma \in \{U, 1\}$ ) projected by the sender as an informed individual when the state of the receiver is  $\sigma$ . Three equations (C.5 to C.7) characterize the image  $s_\sigma$  ( $\sigma \in \{0, U, 1\}$ ) projected by the sender as a partisan when the state of the receiver is  $\sigma$ . The rest of the system (equations C.8 to C.10) link the values of  $\alpha_1$ ,  $\alpha_2$ , and  $\alpha_3$  to  $\mu_p$ ,  $\mu_r$  and  $\mu_s$  as shown in the proof of Proposition 1. We solve this system and recover the parameters  $r_1, r_U, s_0, s_U, s_1, \hat{\pi}_U, \hat{\pi}_1, \hat{\xi}_j, \mu_p, \mu_r, \mu_s$ . The methodological details are provided in Online Appendix C.

The system is initially solved for the no-policy group, i.e., with values of  $\alpha_i$  as in column (1) of Table 3. This step determines the fundamental equilibrium parameter values of the model:  $r_1, r_U, s_0, s_U, s_1, \hat{\pi}_U, \hat{\pi}_1, \hat{\xi}_j$ . These values remain constant across treatments since they dictate the inferences made by the receivers, who are themselves unaware of the treatment assignment. Fixing these parameters, we derive the implied values of  $\mu_p, \mu_r$  and  $\mu_s$  in the different treatments using values of  $\alpha_i$  from columns (2)-(5) of Table 3.

Figure 4 plots, for each treatment, the ratios of  $\mu_r/\mu_p$ ,  $\mu_s/\mu_r$ , and  $\mu_p/\mu_s$  normalized by the respective values in the no policy group, along with their 90% confidence intervals.<sup>23</sup> We find that all policies significantly increase the importance of the reputation concern relative to the partisan persuasion motive, but the priming treatment is the most effective in doing so. The reputation concerns also increase on average relative to partisan signaling, but insignificantly so, as the signaling motive is imprecisely estimated and is dominated by partisan persuasion ( $\alpha_3 > 0$ ). At the same time, the relative importance of the two partisan motives is not significantly affected by the treatments. There is only a slight (and insignificant) decrease in the importance of partisan signaling in the ask to assess tweets treatment, which is the only treatment that explicitly demands respondents to think about the partisanship of the content.<sup>24</sup>

<sup>22</sup>In the proof of Proposition 1, we focus on an equilibrium where  $\hat{\pi}_0 = 1$ , i.e., a sender who knows that the content is false never shares it. This also implies that  $r_0 = 0$ , i.e., if the receiver is informed and knows that the state is 0, they know that a sender who shares has to be uninformed.

<sup>23</sup>To calculate the confidence intervals, we conduct Monte Carlo simulations, taking draws of  $\alpha_1, \alpha_2$  and  $\alpha_3$  from their estimated joint distribution in each treatment from Table 3. Using these draws, we solve the system of equations applying the methodology described above and detailed in Online Appendix C. We use the solution in the no policy group as the starting value of the algorithm.

<sup>24</sup>The key distinction between the ask to assess tweets treatment and the no policy treatment lies in the sequence of decision-making regarding sharing and evaluating tweet characteristics. This distinction allows us to investigate whether the decision to share impacts individuals' perceptions of content characteristics, which we consider fixed in the lower-nest estimation. For example, if respondents choose to believe a particular news item is true because they shared it to avoid feeling responsible for spreading misinformation, this might



### 5.1.3 Cost-of-sharing channel (upper nest)

As previously discussed, the lower-nest estimates allow us to calculate the inclusive utility of sharing, which is the expected benefit of sharing one of the tweets as described in equation (6). In the upper-nest decision, potential sharers compare their inclusive utility from sharing ( $I_{is}$ ) to the cost of sharing ( $W_{in}$ ). They are more inclined to share when the inclusive utility is higher and the cost is lower. We estimate the upper-nest choice using a logit regression on the entire sample of respondents. The unit of observation is the respondent in the upper nest. In this regression, the dependent variable is a dummy variable for sharing anything, and the main explanatory variables are the dummies for each treatment group and the inclusive utility. According to equation (6), the coefficients on the treatment dummies are the estimates of  $W_{in}/\lambda_b$ , while the coefficient on the inclusive utility is the estimate of  $\lambda_w/\lambda_b$ . The average differences in the probability of sharing anything across treatments ( $W_{in}/\lambda_b$ ), after accounting for the inclusive utility of sharing in the lower nest, reflect the cost of sharing associated with each treatment ( $W_{in}$ ), which is our main focus. This is because one should expect the parameter  $\lambda_b$  not to systematically vary across treatments, while  $\lambda_w$ , in contrast, could vary across treatments. As above, we control for social-economic characteristics of respondents.

The results of this estimation are presented in columns (1) and (2) of Table 4. Column (1) forces the coefficient on inclusive utility ( $\lambda_w/\lambda_b$ ) to be the same across treatments. Column (2) allows it to vary across treatments. Columns (3) and (4) present the marginal effects of the variables of interest on the probability of sharing anything for each of the two specifications presented in the first two columns of the table, respectively.

Our main finding is that two out of the four treatments, namely, the offer fact-check and ask to assess tweets treatments, significantly reduce sharing anything in comparison to the no policy group, holding the inclusive utility constant. Conversely, the difference between the average sharing in the extra click and priming fake news circulation treatments compared to the no policy group is small and statistically insignificant. This can be seen from the coefficients on the first four presented regressors, which are the treatment dummies. The point estimates of the marginal effects on the probability of sharing anything are  $-2.6$  percentage points for the extra click and priming treatments, and  $-19.8$  and  $-14.7$  percentage points for the offer fact-check and ask to assess tweets treatments (as reported in column (3)). This is in comparison to an average sharing rate of 49.5%. Importantly, the magnitude of the effects does not depend on whether we allow the coefficient on  $I_{is}$  to vary across treatments (i.e., columns (1) and (3) vs. columns (2) and (4)). The estimates of the cost of sharing, i.e., the coefficients  $W_{in}/\lambda_b$ , are

---

indicate a form of motivated reasoning tied to self-justification. In Online Appendix Table A5, we test for this motivated reasoning. Using the subsample of ask to assess tweets and no policy treatment groups, we analyze the relationship between veracity and alignment estimates and the decision to share. The coefficient on the interaction between the decision to share and the no policy treatment dummy indicates whether the decision to share affects the estimates of the content’s veracity and partisan alignment. We find that this coefficient is precisely estimated as zero for veracity estimates. For the partisan alignment, it is negative and marginally significant. However, its magnitude, at  $-0.06$ , is minimal when compared to the alignment variable’s range from  $-2$  to  $2$ . Thus, we conclude that motivated reasoning does not play an important role.

significantly different from each other for different treatments, except for the extra click and priming treatments, for which the coefficients are essentially the same (and not distinguishable from no policy). These results allow us to quantify the costs of sharing associated with each policy.

Furthermore, as expected, the coefficient on the inclusive utility is positive and statistically significant. Our model has an unambiguous prediction that  $\lambda_w/\lambda_b$  is positive in all treatment groups. Using estimates from column (2), we verify that, indeed, the slope of the inclusive utility ( $\lambda_w(T)/\lambda_b$ ) is positive and significant in each treatment subsample, as our model predicts.

## 5.2 Counterfactual analysis of treatment effects

In this section, we quantify the relative importance of the three channels. Our structural model allows us to run counterfactual simulations shutting down one or two of those channels at a time. Shutting down all three channels is equivalent to using the results for the no policy group. Shutting down zero channels is equivalent to the total average treatment effect (presented in Section 3).

In order to see what happens if we only activate the updating channel and shut down the other two channels, we carry out the following procedure. First, we use the coefficients from the lower-nest estimation for the no policy group to predict the counterfactual probabilities of sharing each of the tweets conditional on sharing and the counterfactual inclusive utility for each individual under the assumption that treatments only affect veracity and alignment estimates. These predictions are out of sample for participants of all treatment groups except for the no policy group. In these calculations, we use the actual estimates of veracity and alignment reported by respondents as they reflect the updating effect. Then, using these counterfactual estimates of the inclusive utility and the upper-nest coefficients (from column (1) of Table 4), we predict the counterfactual probability of sharing any tweet in the no policy treatment group. This allows us to predict the counterfactual probability of sharing each tweet for each participant, as this is the product of the conditional probability in the lower nest and the probability of sharing any tweet in the upper nest. Then, we aggregate these estimates by the actual veracity of tweets to calculate the counterfactual probability of sharing true and false tweets.

In order to simulate the counterfactual treatment effects allowing only the salience channel and shutting down the other two channels, we calculate the counterfactual veracity and alignment estimates as if they were unaffected by treatments from columns (4) and (8) of Table 2 by predicting the counterfactual values for the no policy group. Using these predicted estimates, we calculate the counterfactual inclusive utility and counterfactual conditional sharing using the actual lower-nest coefficients by treatment (columns (2) to (5) of Table 3). Then, in the upper nest, we use this inclusive utility but still, as above, calculate the upper-nest sharing for the no policy group to eventually compute the counterfactual probabilities of sharing true and false tweets, as above.

Finally, in order to simulate the effects allowing only the cost of sharing channel but

shutting down the updating and salience channels, we predict veracity and alignment estimates from Table 2 using coefficients for the no policy group. Then, we use these values and the lower-nest coefficients for the no policy group from Table 3 to predict the inclusive utility and conditional sharing probabilities. In the upper nest, in contrast, we use these counterfactual estimates of the inclusive utility and of veracity and alignment to predict counterfactual sharing of any tweet using the upper-nest estimation for each treatment group.

The results of these simulations are presented in Figure 5. The top row shows the simulated effects on false tweets and bottom row – on true tweets. In each row, we first re-state the average treatment effect (for comparison) and then present the simulated sharing by treatment for the simulations allowing for one channel at a time: updating channel only, salience channel only, and the cost of sharing channel only. This exercise delivers several results. First, it is evident that the updating channel plays a very minor role. The simulated sharing rates when one allows only for updating channel are very close to those in no policy group. This is true even in the case of the offer fact check treatment, a priori designed to work through its impact on veracity estimates. Second, it is the salience channel that induces the opposite-sign effects on sharing of false and true tweets. When one allows for only the salience channel, the simulated sharing of false news declines relative to no policy, while increasing for true news. This is the case to a varying degree for all treatments, but it is particularly strong for the priming fake news circulation treatment, as it has strong opposite-direction effects on sharing of true and false news. The cost of sharing channel affects all tweets irrespective of their characteristics in a similar way: adding frictions reduces sharing. The figure confirms the findings from the upper nest regressions: the cost of sharing effect is much more sizable for the offer fact-check and ask to assess tweets treatments than for the extra click and priming fake news circulation treatments. As the cost of sharing is particularly large in the case of the offer fact-check and ask to assess tweets treatments, it is the cost of sharing channel that causes these treatments to reduce the sharing of true news.

We also run simulations, in which we shut down one channel at a time and allow the other two channels to be at play. Figure 6 presents the results. The comparison of columns (1) (ATE) to column (2), where the updating effect is shut down completely shows that essentially all of the effects of each of the treatments is explained by the combination of the salience and cost mechanisms. To put the numbers on this overall conclusion, we use Shapley value decomposition (Shapley, 1953) of the total Average Treatment Effects of each treatment into the effects of each of the three channels (as detailed in Section D of the Online Appendix). The results are presented in Table 5 and illustrated in Figure 7. In this calculation, the sum of the contributions of the three channels by construction is equal to the average treatment effect. We find that the contribution of the updating channel is one order of magnitude smaller than of the other two channels.

### 5.3 Simulating the impact of digital literacy training

In the experiment, we have considered policies designed to reduce the circulation of fake news by directly influencing the sharing process. An important alternative policy is digital literacy training, which aims to modify attitudes and behaviors by helping people to accurately assess the veracity of news. Assessing the impact of such a long-term policy within a short-term experiment is infeasible. However, we can use our model to simulate the effects of a counterfactual policy.

Within the model, it is natural to think that digital literacy training corresponds to an increase in the share of informed individuals, denoted as  $q$ . Such a change affects equilibrium sharing through two channels. The first channel is the direct effect on the senders' expected veracity estimates. We refer to this channel as the "sender's knowledge channel." It captures the fact that, with digital literacy, a higher proportion of senders are fully informed about the veracity of news. As a result, they are less likely to share false news and more likely to share true news. We simulate this effect by randomly selecting a certain share of uninformed participants in our experiment and adjust their veracity estimates to 100% for true tweets and 0% for false tweets. 4% of participants are fully informed in our experiment. We force the share of informed senders to increase by  $(q - 0.04)$  in this counterfactual exercise.

There is also an indirect channel, which we refer to as the "receivers' reaction channel." This channel results from the change in the expected equilibrium reactions of the receiver to information shared by the sender. Specifically, a change in the parameter  $q$  increases the likelihood that the receivers are informed and update negatively on the sender's type if they observe sharing when the content is false. Consequently, a change in  $q$  requires us to solve for the equilibrium. We describe the procedure we follow to do this in Online Appendix E. In short, we assume that digital literacy training does not affect the fundamental preferences of sharer, which means that we can use the values of  $\mu_r$ ,  $\mu_p$ , and  $\mu_s$  estimated in Section 5.1.2. We can then solve for the remaining parameters which allows us to calculate the new values of  $\alpha_1$ ,  $\alpha_2$  and  $\alpha_3$  by treatment. With these recalculated values, we can simulate sharing using the formula for the indirect utility function.

Table 6 presents the results for the counterfactual values of  $q$  equal to 0.06, 0.08, 0.12, 0.16, and 0.20, i.e., the counterfactual increase in the share of informed by 50%, 100%, 300%, and 400% as a result of digital literacy training implemented at different scales. Panel A focuses on sharing of false news and Panel B – on true news. The first column presents the results on aggregate sharing. We find that digital literacy decreases sharing of false news and increases sharing of true news. The magnitude of the effect on true news is much larger than on false news. This is to be expected as an increase in  $q$  has an effect on sharing both in the lower and in the upper nest. In the lower nest, conditional on sharing, it increases the probability of sharing the true news and decreases the probability of sharing false news. This effect should be symmetric. This, then, increases the inclusive utility and, as a consequence, the probability of sharing in the upper nest increases for both true and false news. The magnitude

of the effect is considerable: an increase in  $q$  by 50% from 4% to 6% of informed individuals among social media users leads to a 6.6-percentage-point decline in sharing of false news and a 13.1-percentage-point-increase in sharing of true news.

The effect of increasing  $q$  on sharing is monotonic but nonlinear; it is concave when taking the change in sharing in absolute value as the outcome. Considering the mechanism helps us understand why this is the case. We decompose the total effect into the two effects that go through the two channels, namely the sender’s knowledge and the receivers’ reaction channels, using Shapley value decomposition. The results are presented in the last two columns of the table and illustrated in Figure 8. We find that the total effect is primarily driven by the indirect receivers’ reaction channel. It is not because senders are more informed that they share fewer false news, instead, they do it mostly because they expect receivers to be better informed, making receivers harder to convince and more likely to be critical in their judgments. This is particularly striking for relatively small increases in  $q$ , as it appears that the indirect effect increases at a slower pace than the direct one. This is partly due to the fact that the reputation motive for sharing disappears when  $q$  is large.<sup>25</sup>

Does digital literacy training render short-term policies irrelevant? In Online Appendix Figure A4, we present the effects of the treatments separately for several counterfactual values of  $q$ . As illustrated in the figure, the short-term policies continue to exert a sizable and statistically significant effect for different values of  $q$ . We find that the overall relative ranking of the effectiveness of different policies remains unchanged. This implies that short-term policies, such as priming fake news circulation and fact-checking, have an impact even in the presence of the long-term policy of digital literacy training. The reason for this lies in the mechanisms through which the two types of policies operate: digital literacy training changes the expected receiver’s reactions to shared content, while short-term policies make this reaction more important for the sender, as they increase the weight of reputation in the sender’s utility function through the salience mechanism.

## 6 Policy Implications

Our results have clear policy implications. First and foremost, we show that the most efficient policy is priming fake news circulation advocated in particular by Pennycook and Rand (2022). This policy curtails the circulation of false news, while increasing the circulation of the true news. Our analysis explains why this is the case: this policy has a strong salience effect while its impact on the cost of sharing is limited.

In our experiment, the overall effect of priming treatment on aggregate sharing is insignificant. This is due to the fact that the participants were exposed to a balanced pool of news, one half false and the other half true. In the real world, the effect on overall circulation of news should depend on the average share of the true news available in the pool of potentially sharable news. If, on average, individuals are more exposed to true news, our results imply

---

<sup>25</sup>Recall that the reputation payoff is proportional to  $(E[\psi_i = I|s_{ij} = 1] - q)$ . When  $q$  is large, the receiver is almost certain that the sender is informed and the updating effect is small.

that the priming intervention should increase the total circulation of news. Overall, this policy is the least damaging to the social media platforms' business model among those that we have considered.

The second key policy implication is that fact-checking works mostly through changing the salience of reputational risks of spreading misinformation and not through the receivers' updating their beliefs about veracity. On the one hand, this shows why it is a relatively ineffective policy in our setting. Fact-checking is not designed to work through a behavioral channel, and it is rather costly. On the other hand, this opens a positive perspective for how one could potentially improve fact-checking. For example, one can combine (i) algorithmic fact-checking, which does not require substantial resources, and (ii) nudges, which not only mention to users that a piece of news was detected as potentially suspicious by an algorithm but also reminds people that there are fake news circulating on social media overall. Such a fact-checking-cum-nudge intervention should have a desired effect on sharing through the salience channel, particularly when individuals are good at distinguishing the blatantly false news from the rest. More generally, our results imply that one should use interventions that are the most effective in nudging users toward valuing sharing true rather than false content without imposing high costs of sharing.

It is important to question whether the results we obtain are externally valid or are determined by the setup of our experiment. Our simulation exercises can be used to assess the external validity of our policy implications. For example, one could argue that, because of the experimental setting, the priming treatment may lead participants to update their beliefs about veracity of content, we presented to them. Indeed, the participants who are told "remember that there are a lot of false news circulating on social media," might infer that the experimenter is exposing them to some false information. If this message was to be shown on social media rather than within an experiment, such inferences may not be made. However, our results show that, even if the updating channel were completely shut down in reality, this treatment would still be very effective because this channel does not affect sharing much anyway.

In our experiment, fact-checking is offered to all participants in fact-checking treatment, even those without a priori intentions to share any false news. This approach may exaggerate the perceived cost of the fact-checking treatment for these senders. Even if the actual cost of fact-checking is lower for users who only share true news on social media (and, therefore, do not encounter fact-checking for false news), the priming treatment remains optimal. This is because the salience channel is weaker in the fact-checking treatment compared to priming. Furthermore, as mentioned earlier, there is additional substantial cost associated with fact-checking related to the production of fact-checking information that we ignore in our experiment.

In the ask to assess tweets treatment, participants were incentivized to report their estimates of veracity and partisanship accurately. In real life, if a policy that asks people to think about the quality of the content they are about to share is implemented, there will be no incentive payments. As a consequence, the effect of this policy that we identify should be considered an upper bound.



The other mechanisms behind treatment effects are unlikely to be driven by our experimental setting. Considering its low implementation cost, priming should remain the most effective policy, even though the treatments do not perfectly match real-world policies. The main open question regarding this policy is whether the salience effect would survive repeated exposure to nudges if regularly deployed at large scale. Behavioral scientists suggest that, on the one hand, repeated exposure to similar nudges may lead to “habituation,” potentially reducing its impact over time. On the other hand, repeated nudging could also strengthen desired associations and foster enduring change, such as new habit formation (e.g., [Robitaille et al., 2020](#); [Sasaki et al., 2021](#)). The scalability of interventions is beyond the scope of this paper.<sup>26</sup>

Finally, our results also suggest that digital literacy training is an effective policy for combating the circulation of false news on social media—setting aside the potentially sizable costs of implementing this policy, which we do not consider in our analysis. Digital literacy training primarily operates through the “indirect channel,” i.e., by affecting senders’ expectations regarding receivers’ reactions to the content they share. As the short-term policies in our experiment increase the salience of the reputation mechanism, emphasizing the importance of receivers’ reactions to shared content, the two types of policies complement each other. What appears to have the biggest beneficial effect on curtailing false and amplifying true news is a combination of digital literacy training with a priming fake news circulation policy: educating citizens to be discerning and nudging them to apply their skills.

## 7 Conclusions

We have developed a unified framework for assessing policy interventions that aim at reducing the spread of false information on social media platforms. These interventions include raising awareness about the risks of misinformation, implementing fact-checking procedures, requiring confirmation before sharing, and encouraging users to think critically about the content they are about to share. We conducted a randomized experiment during the 2022 midterm U.S. elections to evaluate the effectiveness of these interventions in curbing the dissemination of false news and boosting dissemination of accurate political news on Twitter. Our findings indicate that the most effective policy, in terms of shifting the balance of shared news toward accurate content, is to prime users about the potential reputational costs of sharing misinformation.

To understand the mechanisms behind the effects of different interventions, we build and structurally estimate a model of sharing political content on social media. We explicitly model three motives for sharing news on social media: maintaining reputation for being a credible source, political persuasion of the audience, and signaling own partisanship. We estimate the equilibrium relationship using the experimental data and recover parameters of the utility function of the sharers. We find that reputational concerns and political persuasion motives are empirically important whereas the partisan signaling is dominated by partisan persuasion.

---

<sup>26</sup>[Bernheim et al. \(2018\)](#) highlight a potential hidden psychological cost of some nudges, using disturbing pictures on cigarette packs as an example. While we refrain from making general welfare claims, it is likely that such psychological costs are small in our context.

Furthermore, our structural estimation allows pinning down the exact channels through which different interventions affect sharing. We find that our treatments have very little effect on the users’ beliefs regarding content characteristics, such as veracity and partisanship of the messages. However, the treatments do raise the salience of reputational concerns, decreasing sharing of false news and increasing sharing of true news. The treatments also increase the overall cost of sharing, which decreases sharing of both true and false news. The optimal policy—the nudge with a “health warning” reminding that there is a lot of false news circulating on social media—is the one that strongly increases salience of reputation, while imposing a low additional cost of sharing.

Overall, our results suggest that reminding people of the consequences of their actions and trusting them to do the right thing is more effective than teaching them about the quality of any specific content.

## References

- Acemoglu, Daron, Asuman Ozdaglar, and Ali ParandehGheibi**, “Spread of (mis)information in social networks,” *Games and economic behavior*, Nov 2010, *70* (2), 194–227.
- , – , and **James Siderius**, “A Model of Online Misinformation,” Working Paper 28884, National Bureau of Economic Research June 2021.
- Allcott, Hunt and Matthew Gentzkow**, “Social Media and Fake News in the 2016 Election,” *Journal of Economic Perspectives*, 2017, *31* (2), 211–36.
- , **Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow**, “The Welfare Effects of Social Media,” *American Economic Review*, March 2020, *110* (3), 629–76.
- , **Matthew Gentzkow, and Chuan Yu**, “Trends in the diffusion of misinformation on social media,” *Research & Politics*, 2019, *6* (2), 2053168019848554.
- , – , and **Lena Song**, “Digital Addiction,” *American Economic Review*, July 2022, *112* (7), 2424–63.
- Anderson, Simon P., Andre de Palma, and Jacques-Francois Thisse**, *Discrete choice theory of product differentiation*, The MIT Press, 1992.
- Angelucci, Charles and Andrea Prat**, “Is Journalistic Truth Dead? Measuring How Informed Voters Are about Political News,” Working Paper 6132-20, MIT Sloan 2023.
- Arechar, Antonio A., Jennifer Allen, Adam J. Berinsky, Rocky Cole, Ziv Epstein, Kiran Garimella, Andrew Gully, Jackson G. Lu, Robert M. Ross, Michael N. Stagnaro, Yunhao Zhang, Gordon Pennycook, and David G Rand**, “Understanding and combatting misinformation across 16 countries on six continents.,” *Nature Human Behaviour*, Jun 2023.
- Barrera, Oscar, Sergei Guriev, Emeric Henry, and Ekaterina Zhuravskaya**, “Facts, alternative facts, and fact checking in times of post-truth politics,” *Journal of Public Economics*, 2020, *182*, 104–123.
- Benabou, Roland and Jean Tirole**, “Laws and Norms,” NBER Working Paper 17579, NBER 2011.
- Bernheim, B. Douglas, Stefano DellaVigna, and David Laibson**, *Handbook of Behavioral Economics - Foundations and Applications 1.*, Hachette UK, 2018.
- Besley, Timothy and Andrea Prat**, “Handcuffs for the grabbing hand? media capture and government accountability,” *American Economic Review*, May 2006, *96* (3), 720–736.
- Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer**, “Salience and Consumer Choice,” *Journal of Political Economy*, 2013, *121* (5), 803–843.

- , – , and – , “Salience,” *Annual Review of Economics*, 2022, 14 (1), 521–544.
- Braghieri, Luca, Ro’ee Levy, and Alexey Makarin**, “Social Media and Mental Health,” *American Economic Review*, November 2022, 112 (11), 3660–93.
- Buggle, Johannes, Thierry Mayer, Seyhun Orcan Sakalli, and Mathias Thoenig**, “The Refugee’s Dilemma: Evidence from Jewish Migration out of Nazi Germany\*,” *The Quarterly Journal of Economics*, 2023, 138 (2), 1273–1345.
- Bursztyn, Leonardo, Georgy Egorov, Ingar Haaland, Aakaash Rao, and Christopher Roth**, “Justifying Dissent,” *The quarterly journal of economics*, Jun 2023, 138 (3), 1403–1451.
- , – , **Ruben Enikolopov, and Maria Petrova**, “Social Media and Xenophobia: Evidence from Russia,” *CEPR Discussion Paper 14877*, 2020.
- Cameron, A. Colin and Pravin K. Trivedi**, *Microeconometrics: methods and applications*, Cambridge University Press, 2005.
- Conger, Kate**, “A Look Inside Twitter’s Move To Flag Trump,” *New York Times*, 2020, May 30, A1. <https://www.nytimes.com/2020/05/30/technology/twitter-trump-dorsey.html> (accessed on Jan 14, 2021).
- and **Davey Alba**, “Trump Posts On Twitter Are Labeled For Falseness,” *New York Times*, 2020, May 27, B1. <https://www.nytimes.com/2020/05/26/technology/twitter-trump-mail-in-ballots.html> (accessed on Jan 14, 2021).
- Danz, David, Lise Vesterlund, and Alistair J. Wilson**, “Belief Elicitation and Behavioral Incentive Compatibility,” *American Economic Review*, September 2022, 112 (9), 2851–83.
- Ershov, Daniel and Juan S. Morales**, “Sharing News Left and Right: Frictions and Misinformation on Twitter,” Technical Report 2023. Mimeo, UCL School of Management.
- EU**, “Digital Services Act: Application of the Risk Management Framework to Russian disinformation campaigns,” Technical Report 2023. Permanent link: <https://op.europa.eu/en/publication-detail/-/publication/c1d645d0-42f5-11ee-a8b8-01aa75ed71a1>.
- Fazio, Lisa**, “Pausing to consider why a headline is true or false can help reduce the sharing of false news,” *The Harvard Kennedy School Misinformation Review*, 2020.
- Finkel, Eli J., Christopher A. Bail, Mina Cikara, Peter H. Ditto, Shanto Iyengar, Samara Klar, Lilliana Mason, Mary C. McGrath, Brendan Nyhan, David G. Rand, Linda J. Skitka, Joshua A. Tucker, Jay J. Van Bavel, Cynthia S. Wang, and James N. Druckman**, “Political sectarianism in America,” *Science*, 2020, 370 (6516), 533–536.
- Godel, William, Zeve Sanderson, Kevin Aslett, Jonathan Nagler, Richard Bonneau, Nathaniel Persily, and Joshua A. Tucker**, “Moderating with the Mob: Evaluating the Efficacy of Real-Time Crowdsourced Fact-Checking,” *Journal of Online Trust and Safety*, Oct. 2021, 1 (1).
- Guay, Brian, Adam Berinsky, Gordon Pennycook, and David Rand**, “How to think about whether misinformation interventions work,” *Nature Human Behavior*, 2023.
- Guess, Andrew M, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar**, “A digital media literacy intervention increases discernment between mainstream and false news in the United States and India.,” *Proceedings of the National Academy of Sciences of the United States of America*, Jun 2020.
- Guess, Andrew M., Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, Sandra González-Bailón, Edward Kennedy, Young Mie Kim, David Lazer, Devra Moehler, Brendan Nyhan, Carlos Velasco Rivera, Jaime Settle, Daniel Robert Thomas, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Beixian Xiong, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker**, “How do social media feed al-



- Effect: Attaching Warnings to a Subset of FakeNews Headlines Increases Perceived Accuracy of Headlines Without Warnings,” *Management Science*, 2020, *66*, 4921–5484.
- , – , – , and – , “The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings.,” *Management Science*, 2022, *66* (66(11)).
- and **David Rand**, “Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation,” *Nature Communications*, 2022, *13* (2333).
- Persily, Nathaniel and Joshua A. Tucker**, *Social Media and Democracy*, Cambridge University Press, 2020.
- and – , eds, *Social media and democracy: the state of the field, prospects for reform*, Cambridge University Press, Aug 2020.
- Pillai, Raunak M. and Lisa K. Fazio**, “Explaining why headlines are true or false reduces intentions to share false information,” *Collabra: Psychology*, Sep 2023, *9* (1).
- Robitaille, Nicole, Julian House, and Nina Mazar**, “Effectiveness of planning prompts on organizations’ likelihood to file their overdue taxes: A multi-wave field experiment,” *Management Science*, Nov 2020.
- Sasaki, Shusaku, Hirofumi Kurokawa, and Fumio Ohtake**, “Effective but fragile? Responses to repeated nudge-based messages for preventing the spread of COVID-19 infection.,” *Japanese economic review (Oxford, England)*, Jun 2021, p. 1–38.
- Shapley, Lloyd S.**, “A Value for n-Person Games,” *Contributions to the Theory of Games*, 1953, *2*.
- Skafle, Ingjerd, Anders Nordahl-Hansen, Daniel S Quintana, Rolf Wynn, and Elia Gabarron**, “Misinformation About COVID-19 Vaccines on Social Media: Rapid Review.,” *Journal of Medical Internet Research*, Aug 2022, *24* (8), e37367.
- Srinivasan, Karthik**, “Paying Attention,” Technical Report 2023. mimeo, University of Chicago.
- Thaler, Richard H. and Cass R. Sunstein**, *Nudge*, New Haven, CT and London: Yale University Press, 2008.
- Train, Kenneth E.**, *Discrete Choice Methods with Simulation*, Cambridge University Press, 2009.
- Tufekci, Zeynep**, “How social media took us from Tahrir Square to Donald Trump,” Technical Report, MIT Technology Review 2018.
- Vosoughi, Soroush, Deb Roy, and Sinan Aral**, “The spread of true and false information online,” *Science*, 2018, *359*, 1146–1151.
- Wojcik, Stefan and Adam Hughes**, “Sizing Up Twitter Users,” *Pew Center*, 2019.
- Yaqub, Waheeb, Otari Kakhidze, Morgan Brockman, Nasir Memon, and Sameer Patil**, “Effects of Credibility Indicators on Social Media News Sharing Intent,” *CHI ’20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.
- Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov**, “Political Effects of the Internet and Social Media,” *Annual Review of Economics*, 2020, *12*, 415–438.




# Figure 1: Tweets presented to the participants

## (a) Tweets with true political information

**2022 Political News**  
@NewsElection22

The Biden administration has opened the application process for Americans seeking student debt relief.



11:20 PM · Oct 30, 2022

**2022 Political News**  
@NewsElection22

Florida schools ordered to provide bathrooms separated by biological sex.



11:19 PM · Oct 30, 2022

## (b) Tweets with false political information

**2022 Political News**  
@NewsElection22

Biden is adding more IRS agents to investigate your taxes than we have detectives investigating every crime in the country.



11:19 PM · Oct 30, 2022

**2022 Political News**  
@NewsElection22

Supreme Court just voted to ban condoms.

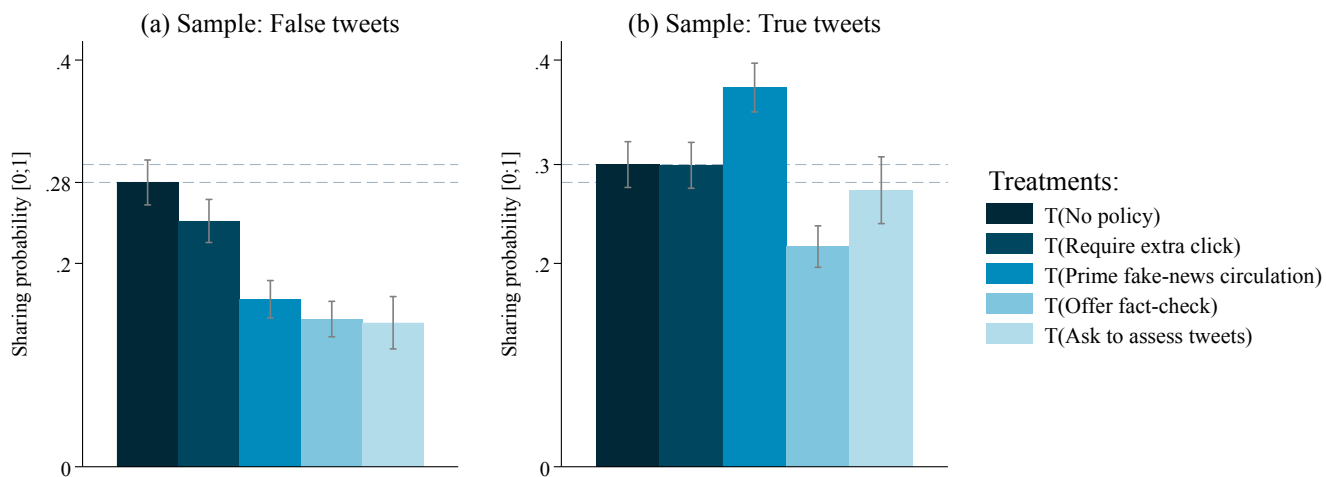


11:19 PM · Oct 30, 2022

**Note:** Tweets on the left are about economic issues and on the right – on cultural issues.

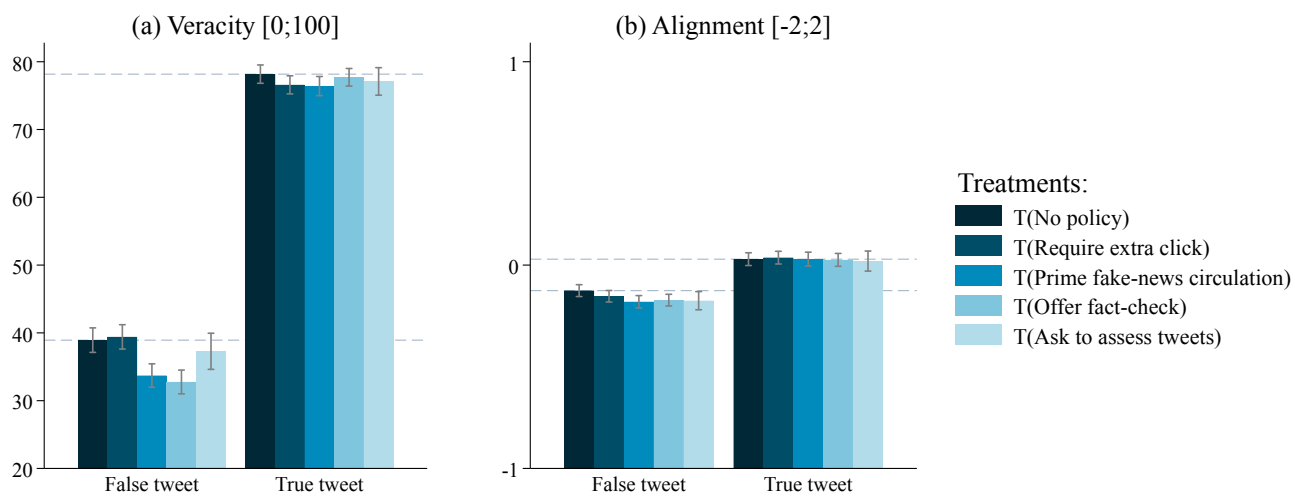


**Figure 2:** Average Treatment Effects on Sharing for False and True Tweets



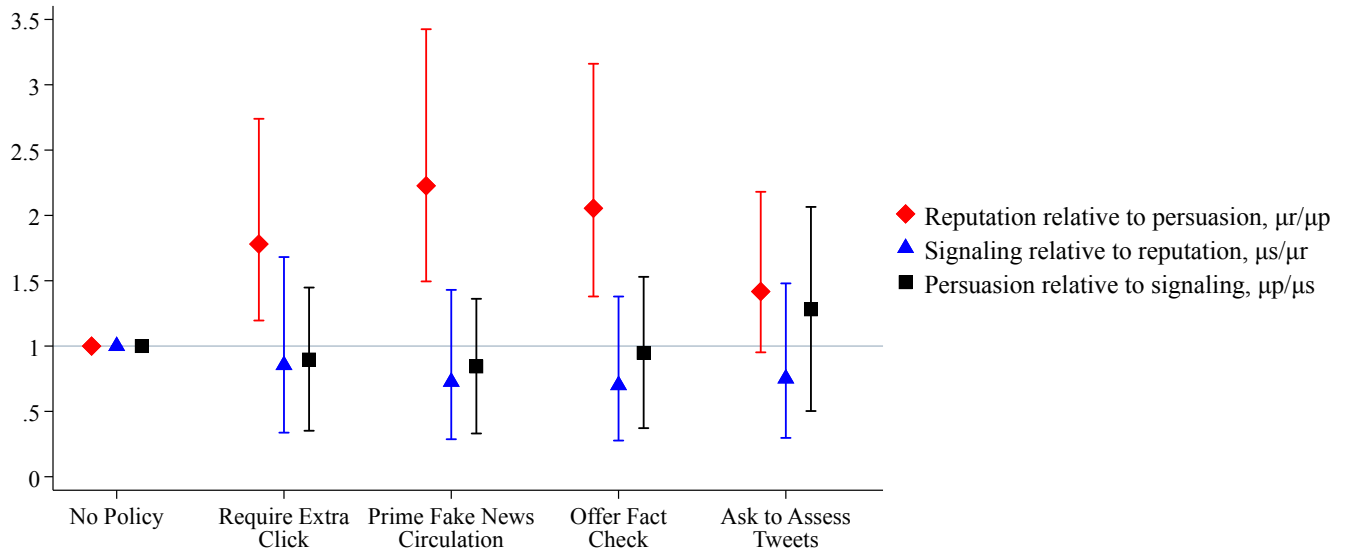
**Note:** Raw-data summary with no controls.

**Figure 3:** Average Treatment Effects on Estimates of Veracity and Partisan Alignment



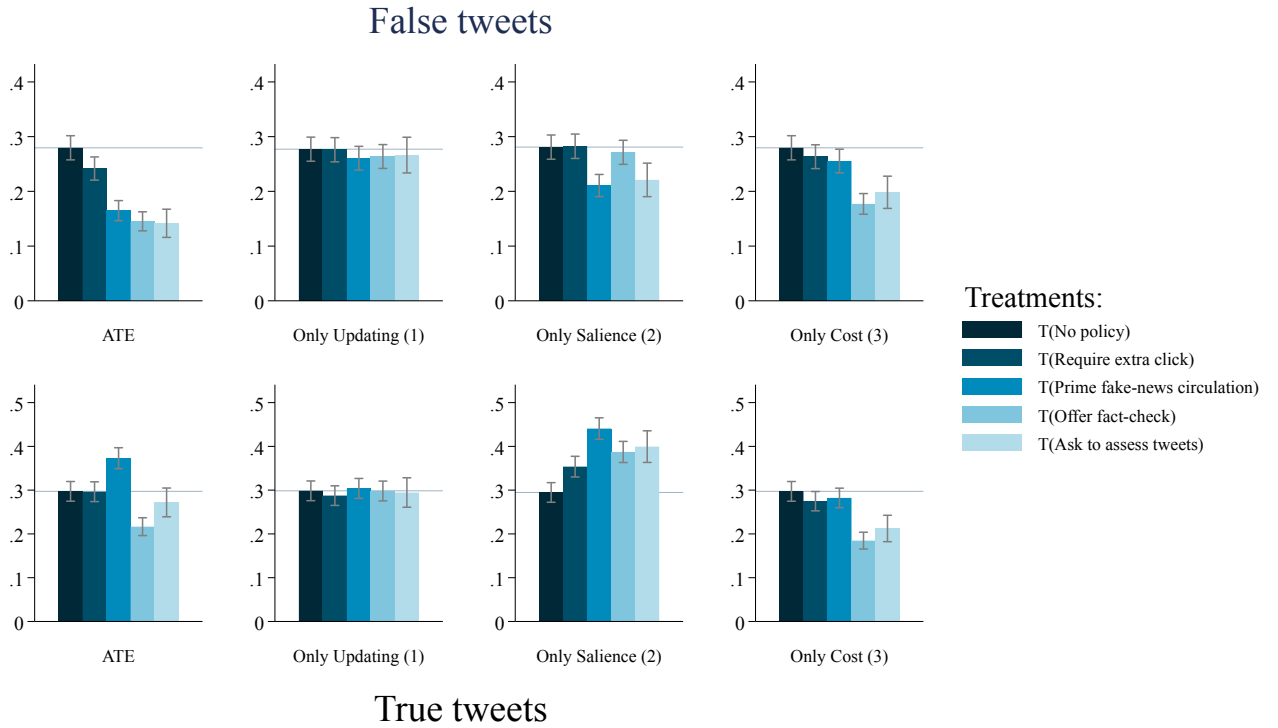
**Note:** Raw-data summary with no controls.

**Figure 4:** The Impact of the Treatments on the Relative Importance of Different Motives in the Sender's Utility Function Relative to the No Policy Treatment Group



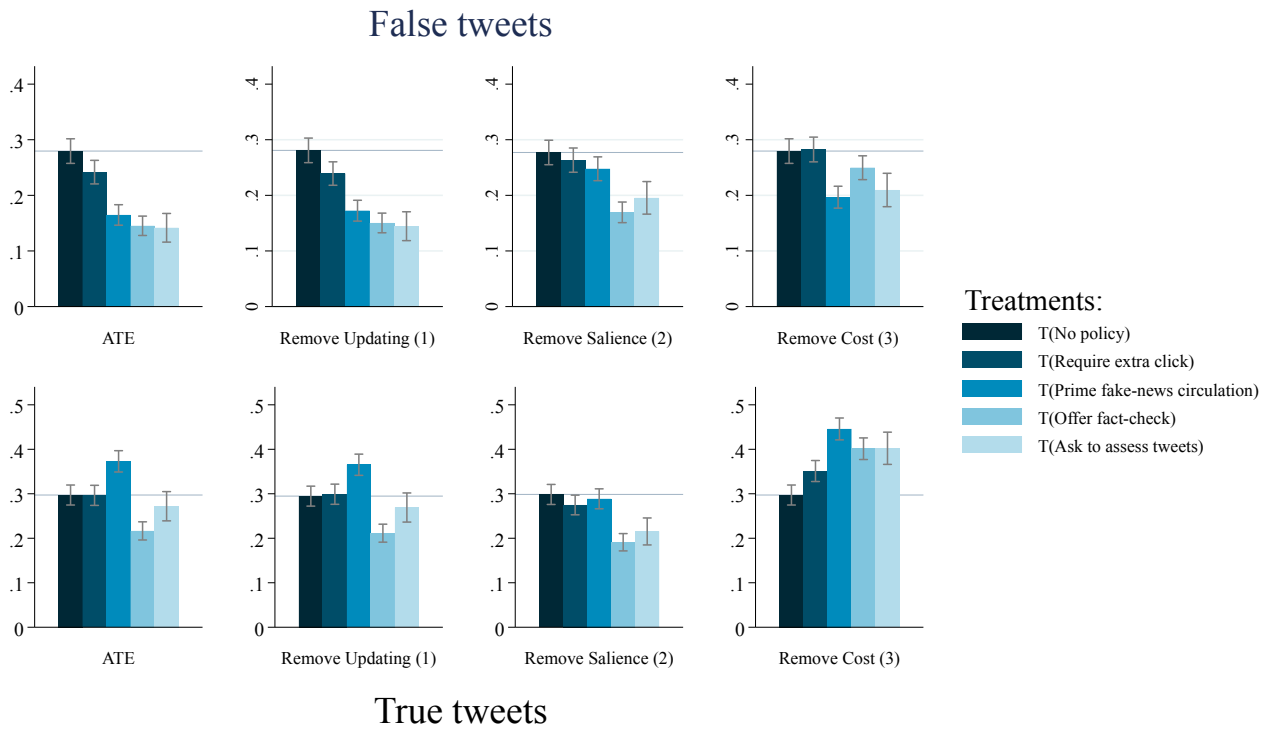
**Note:** The figure presents the result of a numeric solution for the parameters of the sender's utility function. The 90% confidence intervals are presented.

**Figure 5:** Counterfactual exercise: allowing one channel at a time



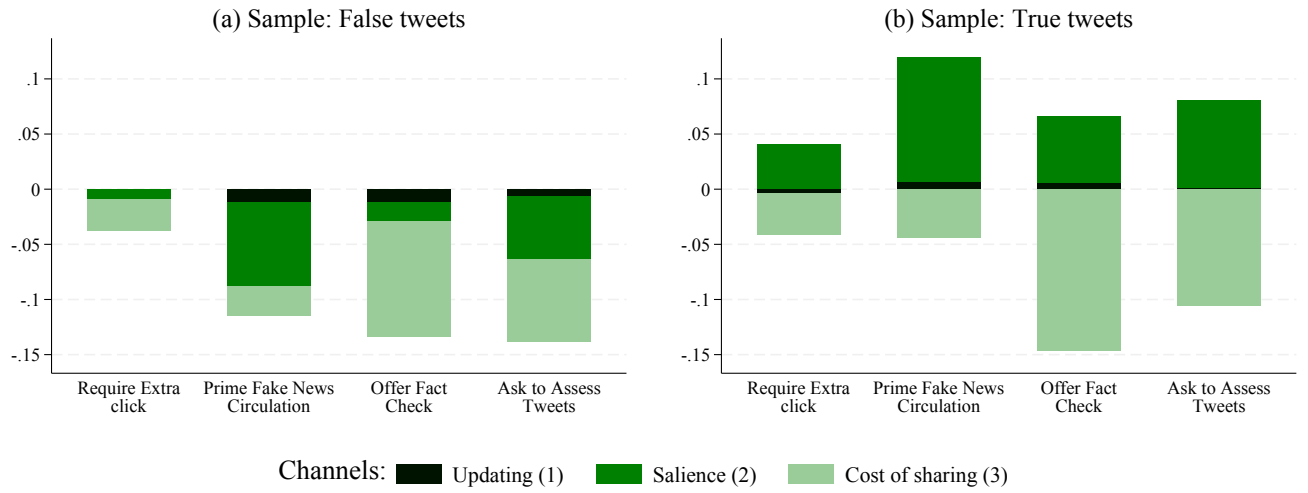
**Note:** The figure shows simulation results for false tweets (in the top row) and true tweets (in the bottom row), presenting counterfactual results keeping one channel at a time: Updating Channel, Salience Channel, and Cost of Sharing Channel. Average Treatment Effects are presented in the first column, for comparison.

**Figure 6:** Counterfactual exercise: removing one channel at a time

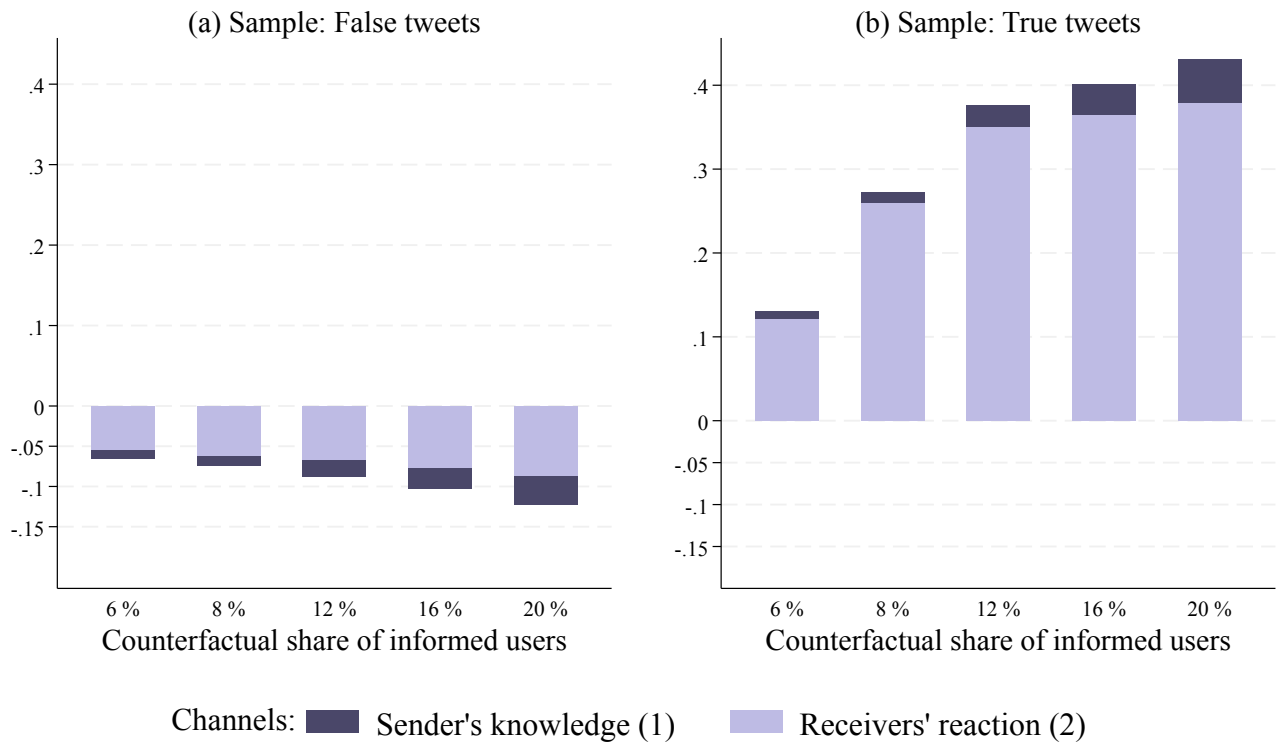


**Note:** The figure shows simulation results for false tweets (in the top row) and true tweets (in the bottom row), presenting counterfactual results removing one channel at a time: Updating Channel, Salience Channel, and Cost of Sharing Channel. Average Treatment Effects are presented in the first column, for comparison.

**Figure 7:** Shapley value decomposition of the average treatment effects into the three channels



**Figure 8:** Shapley value decomposition of the effect of the change in  $q$  into two channels



**Table 1:** Reduced Form: Average Treatment Effects (ATEs) on the Sharing Decision

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable:	Share one of two false tweets			Share one of two true tweets			Share any tweet	
T(Require extra click)	-0.038* (0.022)	-0.036* (0.022)	-0.034 (0.021)	-0.001 (0.023)	0.005 (0.023)	0.001 (0.022)	-0.039 (0.025)	-0.033 (0.024)
T(Prime fake-news circulation)	-0.115*** (0.021)	-0.115*** (0.020)	-0.117*** (0.020)	0.076*** (0.024)	0.081*** (0.023)	0.082*** (0.023)	-0.039 (0.025)	-0.035 (0.024)
T(Offer fact-check)	-0.134*** (0.020)	-0.136*** (0.020)	-0.133*** (0.020)	-0.081*** (0.022)	-0.078*** (0.022)	-0.073*** (0.022)	-0.215*** (0.025)	-0.206*** (0.023)
T(Ask to assess tweets)	-0.138*** (0.024)	-0.141*** (0.024)	-0.137*** (0.024)	-0.025 (0.029)	-0.017 (0.028)	-0.017 (0.028)	-0.163*** (0.032)	-0.154*** (0.030)
Observations	3,501	3,501	3,501	3,501	3,501	3,501	3,501	3,501
R <sup>2</sup>	0.019	0.041	0.075	0.013	0.040	0.083	0.028	0.139
Mean Dep. Var. in No Policy	0.280	0.280	0.280	0.297	0.297	0.297	0.577	0.577
SD Dep. Var.	0.401	0.401	0.401	0.455	0.455	0.455	0.500	0.500
Socio-Economic Controls		✓	✓		✓	✓		✓
Political & Twitter-use Controls			✓			✓		✓

**Note:** The table presents the results of the OLS estimation of the Average Treatment Effects on the sharing decision. The unit of observation is a respondent. Robust standard errors are in parentheses. No policy treatment group is the comparison group.

\* p<0.1, \*\* p<0.05, \*\*\* p<0.01.

**Table 2:** Structural Estimation: Updating Channel: ATEs on Veracity and Partisan Alignment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable:	Veracity				Partisan Alignment			
Sample, Tweets:	False	True	All		False	True	All	
T(Require extra click)	0.006 (0.012)	-0.016 (0.010)	-0.005 (0.008)	0.005 (0.012)	-0.029 (0.020)	0.008 (0.014)	-0.010 (0.013)	-0.028 (0.020)
T(Prime fake-news circulation)	-0.053*** (0.011)	-0.017* (0.010)	-0.035*** (0.007)	-0.052*** (0.012)	-0.058*** (0.021)	-0.000 (0.014)	-0.029** (0.013)	-0.056*** (0.021)
T(Offer fact-check)	-0.062*** (0.012)	-0.005 (0.010)	-0.034*** (0.008)	-0.062*** (0.012)	-0.047** (0.021)	-0.003 (0.014)	-0.025* (0.013)	-0.047** (0.020)
T(Ask to assess tweets)	-0.018 (0.015)	-0.010 (0.012)	-0.014 (0.010)	-0.017 (0.015)	-0.052** (0.024)	-0.009 (0.017)	-0.031* (0.016)	-0.051** (0.024)
True tweet $\times$ T(Extra click)				-0.021 (0.016)				0.035 (0.022)
True tweet $\times$ T(Prime fake news)				0.035** (0.015)				0.055** (0.023)
True tweet $\times$ T(Offer fact-check)				0.057*** (0.016)				0.043* (0.023)
True tweet $\times$ T(Assess tweets)				0.006 (0.019)				0.040 (0.028)
Observations	7,002	7,002	14,004	14,004	7,002	7,002	14,004	14,004
R <sup>2</sup>	0.273	0.136	0.439	0.441	0.010	0.110	0.084	0.085
Mean Dep. Var. in No Policy	.389	.782	.468	.468	-.126	.029	-.039	-.039
SD Dep. Var.	0.360	0.273	0.379	0.379	0.592	0.652	0.630	0.630
Socio-Economic Controls	✓	✓	✓	✓	✓	✓	✓	✓

**Note:** The table presents the results of the OLS estimation of the Average Treatment Effects on estimates of veracity and alignment. The unit of observation is a pair: respondent  $\times$  tweet. Standard errors clustered at the level of respondent are in parentheses. The unit of observation is respondent  $\times$  tweet. T(.) stands for a dummy for the treatment group indicated in parentheses. No policy treatment group is the comparison group.

\* p<0.1, \*\* p<0.05, \*\*\* p<0.01.



**Table 3:** Structural Estimation: Salience Channel

	(1)	(2)	(3)	(4)	(5)	(6)
Conditional Logit Choice Model, Coefficients presented						
Dependent Variable:	Share a particular tweet, sample of participants who share something					
Sample, Treatments:	No Policy	Require Extra Click	Prime Fake News	Offer Fact-Check	Ask to Assess Tweets	All
Veracity ( $\alpha_1$ )	1.041*** (0.250)	1.721*** (0.263)	2.140*** (0.320)	1.996*** (0.331)	1.443*** (0.469)	1.574*** (0.131)
Alignment ( $\alpha_2$ )	0.271** (0.121)	0.243* (0.148)	0.154 (0.174)	0.329** (0.162)	1.111*** (0.341)	0.293*** (0.070)
Veracity $\times$ Alignment ( $\alpha_3$ )	0.932*** (0.342)	1.045*** (0.385)	1.099** (0.500)	1.000*** (0.376)	0.826 (0.581)	1.002*** (0.186)
Observations	1,832	1,692	1,684	1,136	584	6,928
Mean Dep. Var.	0.25	0.25	0.25	0.25	0.25	0.25
Socio-Economic Controls	✓	✓	✓	✓	✓	✓
Tweet fixed effects	✓	✓	✓	✓	✓	✓

**Note:** The table presents the results of the estimation of the McFadden's Conditional Logit Choice Model of the Lower Nest. The unit of observation is a pair: respondent  $\times$  tweet. Standard errors are clustered at the level of respondent. The sample includes only those respondents, who shared something.

\*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

**Table 4:** Structural Estimation: Cost Channel

	(1)	(2)	(3)	(4)
Logit Model				
Dependent Variable:	Share any tweet, sample of all participants			
Specification:	(A)	(B)	(A)	(B)
	Coefficients		Marginal effects	
T(Require extra click) ( $W_{in}/\lambda_b$ )	-0.115 (0.104)	-0.107 (0.104)	-0.026 (0.023)	-0.024 (0.023)
T(Prime fake-news circulation) ( $W_{in}/\lambda_b$ )	-0.116 (0.105)	-0.111 (0.105)	-0.026 (0.024)	-0.025 (0.024)
T(Offer fact-check) ( $W_{in}/\lambda_b$ )	-0.880*** (0.108)	-0.874*** (0.108)	-0.198*** (0.023)	-0.197*** (0.023)
T(Ask to assess tweets) ( $W_{in}/\lambda_b$ )	-0.653*** (0.135)	-0.648*** (0.135)	-0.147*** (0.030)	-0.146*** (0.030)
Inclusive utility ( $\lambda_w/\lambda_b$ )	1.036*** (0.144)	1.399*** (0.286)	0.233*** (0.031)	0.231*** (0.031)
... $\times$ T(Require extra click) ( $\lambda_w/\lambda_b \times T1$ )		-0.708* (0.404)		
... $\times$ T(Prime fake-news circulation) ( $\lambda_w/\lambda_b \times T2$ )		-0.471 (0.397)		
... $\times$ T(Offer fact-check) ( $\lambda_w/\lambda_b \times T3$ )		-0.320 (0.399)		
... $\times$ T(Ask to assess tweets) ( $\lambda_w/\lambda_b \times T4$ )		-0.314 (0.509)		
Observations	3,501	3,501	3,501	3,501
Mean Dep. Var.	0.495	0.495	0.495	0.495
Socio-Economic Controls	✓	✓	✓	✓

**Note:** The table presents the results of estimation of a logit model of the upper nest. In specification (A) (columns (1) and (3)) the coefficient on inclusive utility is forced to be the same for all treatments. In specification (B) (columns (2) and (4)) the coefficient on inclusive utility can vary across treatments. The unit of observation is a respondent. Robust standard errors.

\*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

**Table 5:** Shapley Decomposition of ATE into the Three Channels

Treatment	ATE	Three channels:		
		(1) Updating	(2) Salience	(3) Cost
Panel A: False Tweets				
T(Require extra click)	-0.0379	0.000	-0.0093	-0.0286
T(Prime fake news circulation)	-0.1148	-0.0122	-0.0756	-0.0271
T(Offer fact-check)	-0.1344	-0.0118	-0.0175	-0.1051
T(Ask to assess tweets)	-0.138	-0.007	-0.0564	-0.0746
Panel B: True Tweets				
T(Require extra click)	-0.0008	-0.0043	0.0408	-0.0372
T(Prime fake news circulation)	0.0757	0.0068	0.1132	-0.0443
T(Offer fact-check)	-0.0807	0.0057	0.0599	-0.1463
T(Ask to assess tweets)	-0.0253	0.0012	0.0792	-0.1056

**Note:** Shapley decomposition of the Average Treatment Effects into the three channels. Panel A presents the results for false tweets, Panel B – for true tweets. In each row, the effects from the three channels sum up to the ATE.

**Table 6:** The Simulated Effect of the Digital Literacy Training

Value of $q$	Total effect on sharing	Two channels:	
		(1) Sender's knowledge	(2) Receivers' reaction
Panel A: False Tweets			
$q = 0.06$	-0.0655	-0.0101	-0.0554
$q = 0.08$	-0.0743	-0.0113	-0.063
$q = 0.12$	-0.0882	-0.0202	-0.068
$q = 0.16$	-0.1033	-0.0252	-0.0781
$q = 0.20$	-0.1222	-0.0353	-0.0869
Panel B: True Tweets			
$q = 0.06$	0.131	0.0088	0.1222
$q = 0.08$	0.272	0.0126	0.2594
$q = 0.12$	0.3753	0.0252	0.3501
$q = 0.16$	0.4018	0.0365	0.3652
$q = 0.20$	0.4307	0.0516	0.3791

**Note:** Decomposition of the effect of an increase in  $q$  on equilibrium sharing in the no policy group into the two channels: receiver reaction and sender knowledge. Panel A presents the results for false tweets, Panel B – for true tweet.

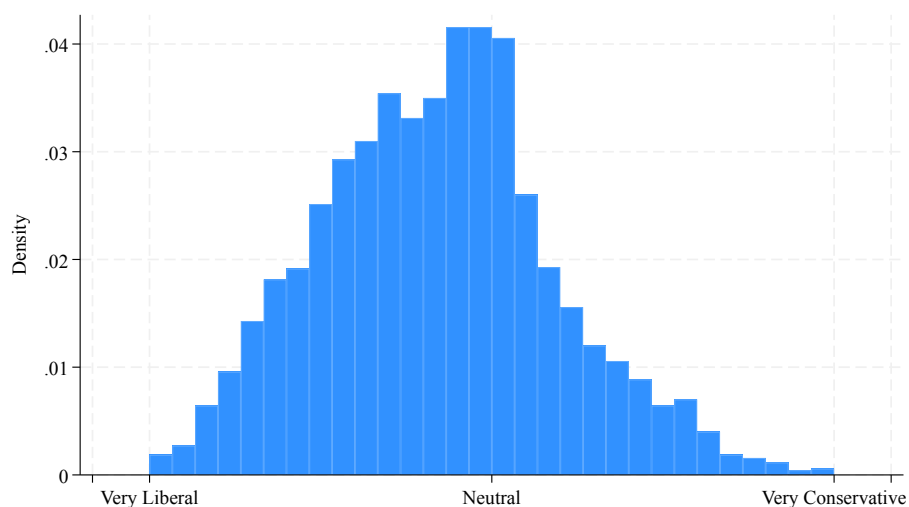
# Online Appendix

## A Online Appendix Tables and Figures

Figure A1: Twitter Account “2022 Political News”

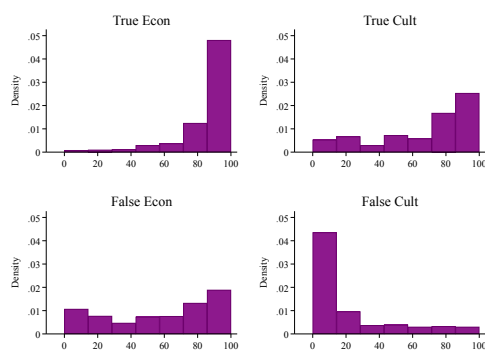
The image shows a screenshot of a Twitter profile for the account "2022 Political News" (@NewsElection22). The profile header includes the account name, a bio stating "Top US Political News" and "Joined September 2022", and follower/following counts. The main content area displays a tweet from Fox News (@FoxNews) with a video thumbnail of John Ratcliffe speaking. The right-hand sidebar contains a search bar, a "You might like" section with recommendations for The Guardian, Breitbart News, and 1inch Network, and a "Trends for you" section listing trending topics like #Taliban and #immobilier.

**Figure A2:** Distribution of the score of political orientation across respondents

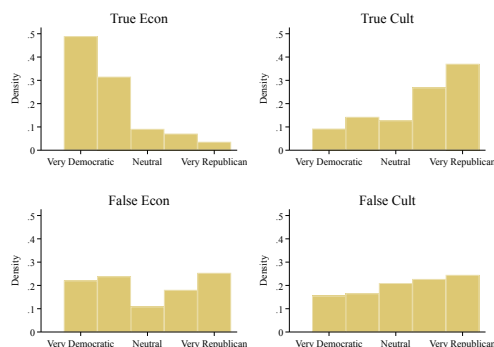


**Figure A3:** Distribution of veracity and partisanship estimates, no policy group

**(a) Veracity estimates**



**(b) Partisanship estimates**



**Note:** The figure presents the distributions of veracity and partisanship evaluations by tweet in no policy treatment group.

True Econ: Tweet about student debt relief;

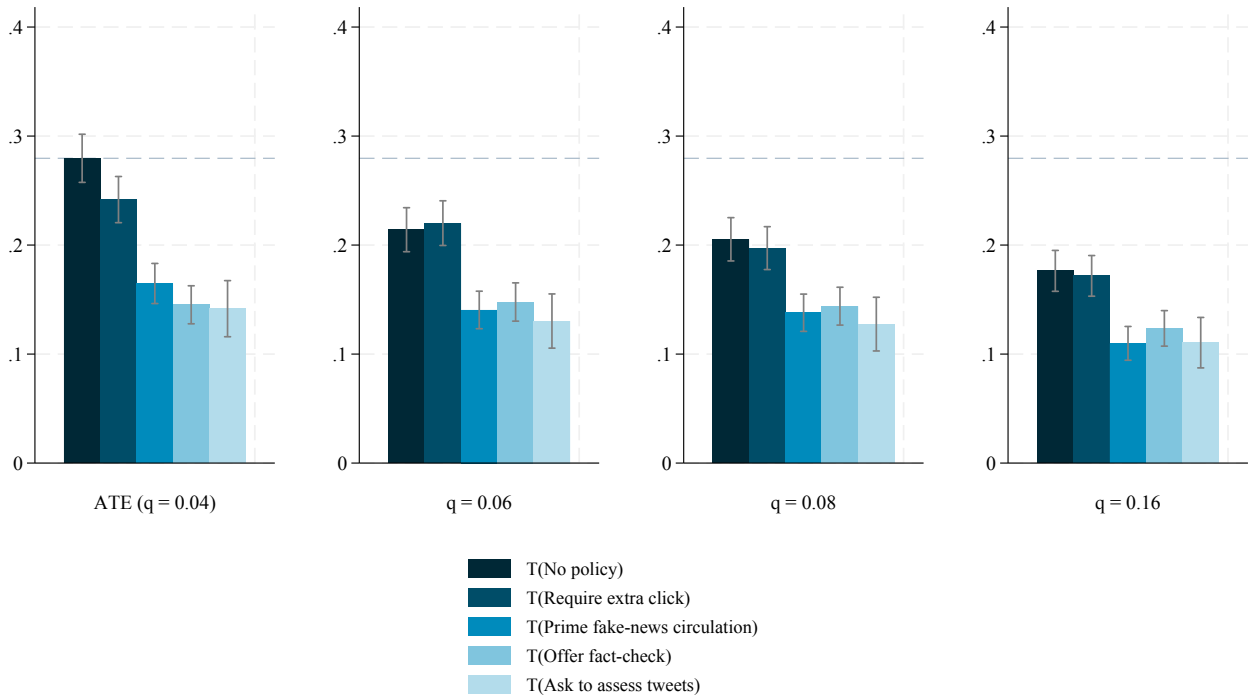
True Cult: Tweet about bathrooms separated by sex;

False Econ: Tweet about IRS agents;

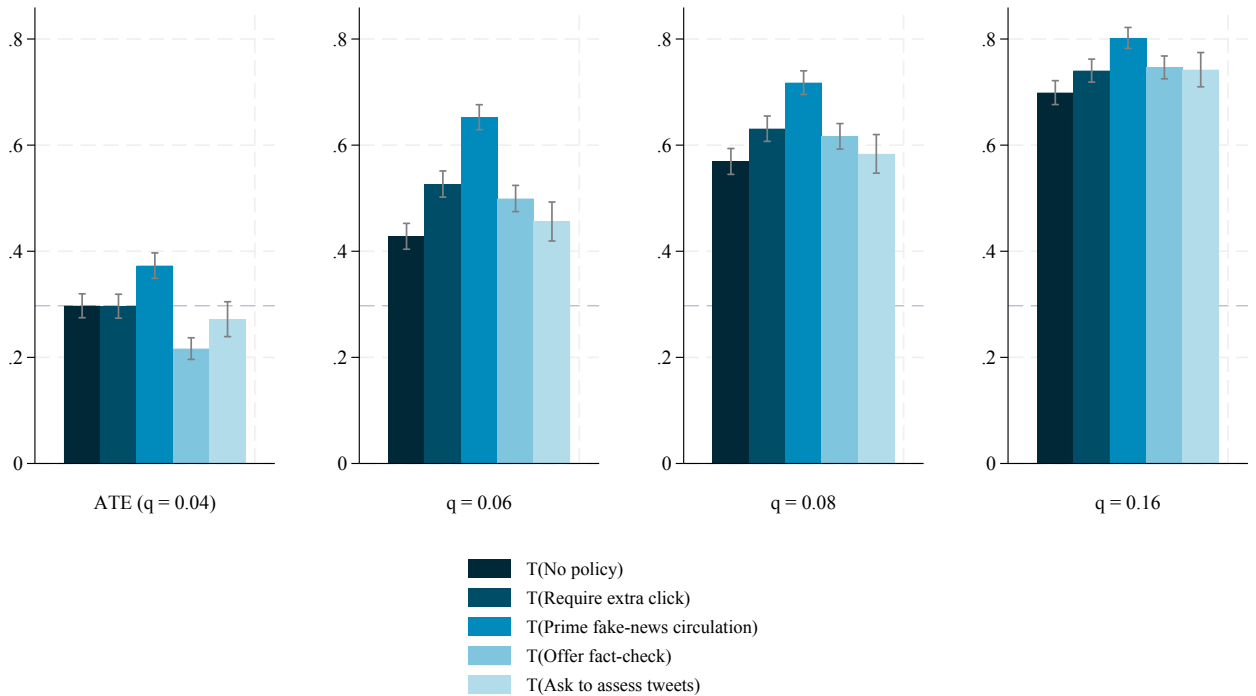
False Cult: Tweet about ban on condoms.

**Figure A4: Counterfactual treatment effects by  $q$**

**(a) False Tweets:**



**(b) True Tweets:**



**Note:** The figure presents counterfactual estimates of combining the short-term policies emulated by the treatments with the long-term policy of digital literacy training implemented at different scales.

**Table A1:** Balance test

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Dummies for the Treatments, relative to T(No policy)										
	T(Extra click)		T(Prime fake news)		T(Offer fact-check)		T(Assess tweets)				
	coeff.	s.e.	coeff.	s.e.	coeff.	s.e.	coeff.	s.e.	R-sq.	Mean	SD
Dependent variable:											
Age	0.821	(0.811)	0.302	(0.814)	0.031	(0.820)	-0.150	(0.1051)	0.0004	44.64	16.305
Male dummy	0.020	(0.025)	0.037	(0.025)	0.033	(0.025)	0.080**	(0.032)	0.002	.46	.498
Region: Northeast	-0.026	(0.020)	-0.019	(0.020)	0.029	(0.021)	-0.031	(0.025)	0.003	.2	.4
Region: Midwest	0.001	(0.021)	0.017	(0.021)	0.000	(0.021)	0.017	(0.027)	0.0003	.23	.421
Region: West	-0.001	(0.020)	-0.022	(0.019)	-0.031	(0.019)	-0.028	(0.024)	0.0013	.17	.377
Race: White	0.003	(0.023)	-0.019	(0.023)	0.011	(0.023)	-0.012	(0.029)	0.0006	.71	.453
Race: African American	0.002	(0.018)	0.016	(0.018)	-0.012	(0.018)	-0.004	(0.023)	0.0007	.15	.359
Race: Asian	-0.007	(0.008)	0.002	(0.009)	0.003	(0.009)	-0.001	(0.011)	0.0005	.03	.166
Married	-0.034	(0.025)	-0.023	(0.025)	-0.027	(0.025)	-0.035	(0.031)	0.0007	.41	.492
Single	0.011	(0.024)	-0.009	(0.024)	0.022	(0.024)	0.027	(0.030)	0.0007	.34	.473
Divorced/separated	0.013	(0.017)	0.021	(0.018)	-0.015	(0.017)	0.028	(0.023)	0.0018	.14	.345
Working full-time	-0.012	(0.025)	-0.012	(0.025)	-0.031	(0.025)	-0.090***	(0.032)	0.0026	.47	.499
Working part-time	-0.018	(0.017)	-0.031*	(0.017)	0.011	(0.018)	-0.015	(0.021)	0.0021	.13	.336
Unemployed	0.007	(0.013)	0.006	(0.013)	0.003	(0.013)	0.016	(0.017)	0.0003	.07	.263
Retired	0.007	(0.018)	0.011	(0.018)	0.006	(0.018)	0.031	(0.024)	0.0006	.16	.365
Care of home/family	-0.012	(0.011)	0.005	(0.012)	0.020	(0.013)	0.024	(0.017)	0.0027	.06	.243
Log earnings	-0.011	(0.043)	-0.033	(0.042)	-0.011	(0.044)	-0.089	(0.056)	0.0009	10.79	.852
Community college degree +	-0.028	(0.025)	-0.040	(0.025)	0.007	(0.025)	-0.040	(0.032)	0.0016	.49	.5
Democrat	0.041*	(0.025)	-0.002	(0.025)	-0.018	(0.025)	-0.019	(0.032)	0.002	.43	.495
Republican	-0.010	(0.022)	0.021	(0.022)	0.012	(0.022)	-0.020	(0.027)	0.001	.25	.435
Independent	-0.027	(0.022)	0.010	(0.022)	0.007	(0.022)	0.025	(0.029)	0.0013	.26	.436
Political Orientation Score	-0.088	(0.502)	-0.040	(0.526)	0.441	(0.519)	-0.496	(0.648)	0.0007	-5.28	10.298
Voted for Clinton in 2016	0.003	(0.025)	-0.039	(0.025)	-0.036	(0.025)	-0.039	(0.031)	0.0016	.39	.488
Voted for Trump in 2016	-0.023	(0.024)	0.012	(0.024)	-0.013	(0.024)	-0.033	(0.030)	0.001	.33	.468
Voted for Biden in 2020	0.031	(0.025)	0.001	(0.025)	-0.011	(0.025)	-0.001	(0.032)	0.0009	.51	.5
Voted for Trump in 2020	-0.050**	(0.023)	-0.011	(0.024)	-0.018	(0.024)	-0.031	(0.030)	0.0015	.31	.463
Social Desirability Score	-0.171	(0.124)	-0.210*	(0.120)	-0.267**	(0.124)	-0.357**	(0.152)	0.0021	6.78	2.45
Get news form radio	-0.035	(0.025)	-0.036	(0.025)	-0.022	(0.025)	-0.070**	(0.031)	0.0017	.38	.484
Get news from TV	0.030	(0.024)	0.005	(0.024)	-0.020	(0.024)	-0.012	(0.031)	0.0014	.66	.474
Get news from newspapers	0.037*	(0.022)	0.017	(0.022)	0.024	(0.022)	-0.025	(0.026)	0.0018	.25	.431
Get news from internet	-0.034	(0.022)	-0.003	(0.021)	-0.018	(0.022)	0.008	(0.027)	0.0011	.75	.431
Do not follow news	0.006	(0.012)	-0.009	(0.012)	0.016	(0.013)	-0.011	(0.015)	0.0016	.06	.244
Twitter: Days used last week	-0.078	(0.126)	-0.194	(0.124)	-0.175	(0.125)	-0.218	(0.157)	0.0011	4.87	2.469
Twitter: Time Spent	0.005	(0.044)	-0.017	(0.044)	-0.060	(0.044)	-0.037	(0.054)	0.0008	1.76	.88
Twitter: Followers	0.051	(0.059)	0.038	(0.059)	0.016	(0.059)	-0.112	(0.068)	0.0015	2	1.161
Twitter: Respondent follows	0.050	(0.057)	-0.017	(0.055)	-0.015	(0.056)	-0.048	(0.070)	0.0007	1.95	1.112
Issue importance: Abortion	-0.090	(0.126)	0.026	(0.126)	-0.092	(0.125)	-0.005	(0.160)	0.0004	4.48	2.495
Issue importance: Gun control	0.040	(0.099)	0.074	(0.099)	0.053	(0.099)	0.276**	(0.126)	0.0015	4.36	1.971
Issue importance: Crime	0.192*	(0.105)	0.039	(0.104)	-0.014	(0.102)	-0.017	(0.130)	0.0015	2.62	2.074
Issue importance: Healthcare	-0.001	(0.097)	-0.137	(0.096)	-0.057	(0.098)	-0.016	(0.118)	0.0008	4.87	1.916
Issue importance: Inflation	-0.082	(0.111)	-0.080	(0.112)	-0.129	(0.112)	-0.165	(0.141)	0.0006	4.75	2.222
Issue importance: Immigration	-0.080	(0.108)	0.040	(0.108)	0.018	(0.109)	-0.105	(0.141)	0.0006	5.49	2.182
Issue importance: War	-0.044	(0.099)	-0.079	(0.096)	0.026	(0.1)	-0.219*	(0.127)	0.0013	4.03	1.945
Issue importance: Education	0.065	(0.111)	0.116	(0.111)	0.194*	(0.108)	0.250*	(0.137)	0.0015	5.4	2.162

**Note:** The table presents the results of the balance tests. We regress each of the pre-treatment characteristics on the four dummies indicating each treatment group keeping the no policy as the comparison group. Each row presents the results of a separate regression. The number of observations is 3,501. Robust standard errors are in parentheses.

\* p<0.1, \*\* p<0.05, \*\*\* p<0.01.



**Table A2:** Omnibus test of randomization quality

Sample: Dependent var – dummy for:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	The following treatments in comparison to group T(No Policy):							
	T(Extra click)		T(Prime fake news)		T(Offer fact-check)		T(Assess tweets)	
	coeff.	se	coeff.	se	coeff.	se	coeff.	se
Age	0.001	(0.001)	0.000	(0.001)	0.000	(0.001)	-0.001	(0.001)
Male dummy	0.014	(0.028)	0.047*	(0.028)	0.046	(0.028)	0.094***	(0.030)
Region: Northeast	-0.047	(0.036)	-0.038	(0.036)	0.017	(0.035)	-0.066*	(0.038)
Region: Midwest	-0.011	(0.034)	0.012	(0.034)	-0.008	(0.034)	-0.014	(0.038)
Region: West	-0.019	(0.037)	-0.046	(0.038)	-0.055	(0.038)	-0.073*	(0.040)
Race: White	-0.001	(0.044)	-0.014	(0.044)	0.008	(0.045)	-0.014	(0.048)
Race: African American	-0.021	(0.053)	0.001	(0.052)	-0.022	(0.054)	-0.028	(0.057)
Race: Asian	-0.063	(0.087)	0.025	(0.085)	0.015	(0.082)	0.019	(0.092)
Married	-0.051	(0.044)	-0.046	(0.044)	-0.078*	(0.044)	0.018	(0.048)
Single	-0.007	(0.047)	-0.046	(0.047)	-0.029	(0.047)	0.026	(0.051)
Divorced/separated	-0.012	(0.052)	0.011	(0.051)	-0.069	(0.053)	0.084	(0.057)
Working full-time	-0.060	(0.047)	-0.052	(0.048)	0.015	(0.050)	-0.095*	(0.055)
Working part-time	-0.084	(0.054)	-0.115**	(0.055)	0.053	(0.056)	-0.067	(0.061)
Unemployed	-0.045	(0.063)	-0.033	(0.063)	0.029	(0.066)	-0.023	(0.071)
Retired	-0.061	(0.057)	-0.026	(0.058)	0.043	(0.060)	0.009	(0.064)
Care of home/family	-0.095	(0.070)	-0.011	(0.068)	0.132*	(0.068)	0.071	(0.076)
Log earnings	0.018	(0.019)	0.002	(0.019)	0.001	(0.018)	-0.004	(0.021)
Community college degree + Democrat	-0.039	(0.029)	-0.039	(0.028)	0.023	(0.029)	0.011	(0.030)
Independent	-0.034	(0.037)	0.017	(0.037)	-0.005	(0.037)	0.011	(0.039)
Political Orientation Score	0.001	(0.002)	0.000	(0.002)	0.002	(0.002)	-0.001	(0.002)
Voted for Trump in 2016	0.017	(0.044)	0.043	(0.044)	-0.018	(0.044)	-0.032	(0.047)
Voted for Trump in 2020	-0.086*	(0.046)	-0.025	(0.047)	-0.030	(0.046)	-0.013	(0.049)
Get news form radio	-0.049*	(0.028)	-0.043	(0.028)	-0.014	(0.028)	-0.045	(0.030)
Get news from TV	0.027	(0.031)	-0.002	(0.031)	-0.013	(0.031)	0.005	(0.033)
Get news from newspapers	0.063*	(0.032)	0.036	(0.032)	0.054*	(0.032)	-0.004	(0.035)
Get news from internet	-0.045	(0.034)	0.000	(0.035)	0.003	(0.035)	0.009	(0.038)
Do not follow news	-0.018	(0.065)	-0.094	(0.067)	0.028	(0.063)	-0.087	(0.071)
Twitter: Days used last week	-0.007	(0.006)	-0.014**	(0.006)	-0.008	(0.007)	-0.010	(0.007)
Twitter: Time Spent	0.001	(0.018)	0.001	(0.019)	-0.019	(0.019)	0.003	(0.020)
Twitter: Followers	0.004	(0.016)	0.028*	(0.016)	0.018	(0.016)	-0.011	(0.018)
Twitter: Respondent follows	0.017	(0.016)	-0.010	(0.016)	0.000	(0.017)	0.011	(0.018)
Issue importance: Abortion	-0.004	(0.006)	-0.002	(0.006)	-0.008	(0.006)	-0.002	(0.006)
Issue importance: Healthcare	-0.005	(0.007)	-0.011	(0.007)	-0.005	(0.007)	-0.002	(0.008)
Issue importance: Immigration	-0.008	(0.007)	0.002	(0.007)	0.002	(0.007)	-0.011	(0.008)
Observations	1,561		1,549		1,557		1,131	
R-squared	0.022		0.0210		0.0180		0.0360	
p-value for joint significance	0.438		0.4690		0.7510		0.2250	

**Note:** The table presents the results of the omnibus test of randomization quality. A dummy for each treatment is regressed on pre-treatment characteristics in the four samples that include participants of one treatment at a time as well as the participants of the no policy group. Every two columns present the results of a separate regression. Robust standard errors are in parentheses.

\* p<0.1, \*\* p<0.05, \*\*\* p<0.01.

**Table A3:** Summary statistics, cross-section of participants

Panel A: Full sample						
	Mean	Median	SD	Min	Max	Obs.
Share: True Econ	0.233	0	0.422	0	1	3,501
Share: True Cult	0.061	0	0.239	0	1	3,501
Share: False Econ	0.118	0	0.322	0	1	3,501
Share: False Cult	0.084	0	0.277	0	1	3,501
Veracity estimate: True Econ	86.986	99	20.046	0	100	3,501
Veracity estimate: True Cult	67.43	75	30.032	0	100	3,501
Veracity estimate: False Econ	54.266	60	34.735	0	100	3,501
Veracity estimate: False Cult	18.372	2	27.221	0	100	3,501
Partisanship estimate: True Econ	-1.156	-1	1.063	-2	2	3,501
Partisanship estimate: True Cult	0.675	1	1.366	-2	2	3,501
Partisanship estimate: False Econ	0.061	0	1.506	-2	2	3,501
Partisanship estimate: False Cult	0.269	0	1.377	-2	2	3,501
Partisan alignment with True Econ	0.244	0.133	0.607	-2	2	3,501
Partisan alignment with True Cult	-.186	-.067	0.624	-2	2	3,501
Partisan alignment with False Econ	-.126	0	0.616	-2	2	3,501
Partisan alignment with False Cult	-.193	-.033	0.565	-2	2	3,501
Respondent's Political Orientation	-5.28	-5	10.298	-30	30	3,501
Panel B: No policy subsample						
	Mean	Median	SD	Min	Max	Obs.
Share: True Econ	0.23	0	0.421	0	1	794
Share: True Cult	0.067	0	0.25	0	1	794
Share: False Econ	0.166	0	0.373	0	1	794
Share: False Cult	0.113	0	0.317	0	1	794
Veracity estimate: True Econ	87.984	100	19.356	0	100	794
Veracity estimate: True Cult	68.355	76	30.702	0	100	794
Veracity estimate: False Econ	58.718	67	33.754	0	100	794
Veracity estimate: False Cult	19.137	2	27.912	0	100	794
Partisanship estimate: True Econ	-1.151	-1	1.076	-2	2	794
Partisanship estimate: True Cult	0.683	1	1.34	-2	2	794
Partisanship estimate: False Econ	0.008	0	1.522	-2	2	794
Partisanship estimate: False Cult	0.241	0	1.39	-2	2	794
Partisan alignment with True Econ	0.255	0.167	0.594	-2	2	794
Partisan alignment with True Cult	-.196	-.067	0.6	-2	2	794
Partisan alignment with False Econ	-.082	0	0.61	-2	2	794
Partisan alignment with False Cult	-.169	-.033	0.576	-2	2	794
Respondent's Political Orientation	-5.3	-5	10.097	-30	30	794

**Note:**

True Econ: Tweet about student debt relief;  
True Cult: Tweet about bathrooms separated by sex;  
False Econ: Tweet about IRS agents;  
False Cult: Tweet about ban on condoms.

**Table A4:** Probability to be informed across treatments

	(1)	(2)	(3)	(4)	(5)
	Informed Resp.: Got veracity of all tweets correctly				
T(Require extra click)	0.007 (0.009)	0.005 (0.009)	0.006 (0.009)	0.005 (0.009)	0.005 (0.009)
T(Prime fake-news circulation)	0.016 (0.010)	0.015 (0.010)	0.016 (0.010)	0.016* (0.010)	0.016* (0.010)
T(Offer fact-check)	0.027*** (0.010)	0.027*** (0.010)	0.027*** (0.010)	0.027*** (0.010)	-0.016 (0.010)
T(Ask to assess tweets)	0.001 (0.011)	0.000 (0.011)	0.001 (0.011)	0.001 (0.011)	0.001 (0.011)
Respondent Democrat		0.033*** (0.007)	0.034*** (0.007)	0.037*** (0.007)	0.036*** (0.007)
Respondent Independent		0.005 (0.009)	0.004 (0.009)	0.021** (0.010)	0.022** (0.010)
Respondent saw Fact Check					0.060*** (0.014)
Observations	3,501	3,501	3,501	3,501	3,501
R <sup>2</sup>	0.003	0.009	0.013	0.028	0.032
Mean Dep. Var.	0.041	0.041	0.041	0.041	0.041
SD Dep. Var.	0.199	0.199	0.199	0.199	0.199
Twitter use			✓	✓	✓
Socio-Economic				✓	✓

**Notes.** Unit of analysis is respondent. The table presents the average treatment effects for being informed, i.e., correctly identify true and false tweets with certainty, compared to the group with no policy treatment. Heteroskedasticity-robust standard errors are in parenthesis. \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

**Table A5:** Test for motivated beliefs for veracity and partisan alignment

	(1)	(2)	(3)	(4)
Dependent Variable:	Veracity		Partisan Alignment	
Sample, Treatments:	T(No policy) and T(Assess tweets)			
Decision to share	0.063*** (0.013)	0.063*** (0.013)	0.204*** (0.030)	0.204*** (0.030)
Decision to share $\times$ T(No policy)	0.006 (0.015)	0.006 (0.015)	-0.062* (0.036)	-0.062* (0.036)
T(Ask to assess tweets)	-0.010 (0.009)	0.000 (.)	-0.036** (0.017)	0.000 (.)
Observations	5,735	5,735	5,735	5,735
R <sup>2</sup>	0.623	0.707	0.100	0.226
Mean Dep. Var.	0.465	0.465	-0.046	-0.046
Respondent FEs		✓		✓

**Note:** Robust standard errors are in parentheses.

\*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

## B Proofs

### B.1 Proof of Proposition 1

The utility of individual  $i$  from sharing  $j$ , is given by equation (1). We determine sequentially the persuasion payoff (second term of first line), reputation payoff (third line) and signaling partisanship payoff (second line), linked in equilibrium by conditions we make explicit.

**Persuasion payoff:** Consider a receiver with a partisan prior  $\pi_k$  who receives a message with veracity they estimate to be  $\nu_{kj}$ . Given their estimate of veracity, they update their partisan beliefs. With probability  $1 - \nu_{kj}$  they understand that the message is false, as a result stay with their prior belief as far as the state of the world is concerned. With probability  $\nu_{kj}$  the message is true, and they update based on their perception of the content. If they perceive the tweet  $j$  as favoring Republican (occurs with probability  $\pi_{ij}$ ), they become certain the state is 1.<sup>27</sup> On the contrary, if they perceive the message to be favoring the Democrats (probability  $1 - \pi_{ij}$ ), they take action 0. Therefore, for a given value of  $\pi_k$  and  $\nu_{kj}$ , sender  $i$  expects the action of receiver  $k$  to be:

$$a_k = \nu_{kj}\pi_{ij} + (1 - \nu_{kj})\pi_k.$$

Taking into account the uncertainty that sender  $i$  has on the partisanship of  $k$  (random variable with mean  $\bar{\pi}$ ) and on their evaluation of veracity (i.e., whether they were informed), the expected influence that sender  $i$  can have on the receivers is as follows:

$$E[a_k] = E_{\nu_{kj}} [\nu_{kj}\pi_{ij} + (1 - \nu_{kj})\bar{\pi}] = E_{\nu_{kj}} [\nu_{kj}(\pi_{ij} - \bar{\pi})] + \bar{\pi}. \quad (8)$$

We now determine the receiver's beliefs about veracity. With probability  $q$ , the receiver is informed and learns the true state. The sender, therefore, expects the receiver to have the same belief as their own  $\nu_{ij}$ . Indeed, if the sender is informed, both sender and receiver have the same exact knowledge of the state. If the sender is uninformed, their belief is the prior  $\nu_j^0$ ; as the prior is unbiased, it is equal to expected belief of the informed receiver.

With probability  $1 - q$  the receiver is uninformed. They however update their belief about veracity conditional on the fact that sender  $i$  shared (i.e., event  $s_{ij} = 1$ ) as follows:

$$\begin{aligned} P[\nu_{ij} = 1 | s_{ij} = 1] &= \frac{P[s_{ij} = 1 | \nu_{ij} = 1]P[\nu_{ij} = 1]}{P[s_{ij} = 1]} = \frac{P[s_{ij} = 1 | \nu_{ij} = 1]q\nu_j^0}{P[s_{ij} = 1]}, \\ P[\nu_{ij} = \nu_j^0 | s_{ij} = 1] &= \frac{P[s_{ij} = 1 | \nu_{ij} = \nu_j^0]P[\nu_{ij} = \nu_j^0]}{P[s_{ij} = 1]} = \frac{P[s_{ij} = 1 | \nu_{ij} = \nu_j^0](1 - q)}{P[s_{ij} = 1]}. \end{aligned}$$

Thus the expected value  $\tilde{\nu}_{ij}$  is given by:

$$\tilde{\nu}_j = E[\nu_{ij} | s_{ij} = 1] = \frac{1}{P[s_{ij} = 1]} (q\nu_j^0 P[s_{ij} = 1 | \nu_{ij} = 1] + (1 - q)\nu_j^0 P[s_{ij} = 1 | \nu_{ij} = \nu_j^0]).$$

---

<sup>27</sup>This simplifies the notation, but the model is easily extendable to accommodate intermediate levels of updating.

This belief is the same for all uninformed receivers  $k$ .

The sender's expectation of the average perception of veracity of receiver  $k$  is thus given by:

$$E[\nu_{kj}] = (1 - q)\tilde{\nu}_j + q\nu_{ij}. \quad (9)$$

This expectation (9) is increasing in  $\nu_{ij}$ . Substituting (9) into (8), we obtain:

$$E[a_k|s_{ij} = 1] - \bar{\pi} = ((1 - q)\tilde{\nu}_j + q\nu_{ij})(\pi_{ij} - \bar{\pi}).$$

The persuasion term, therefore, becomes:

$$\begin{aligned} \mu_p(2\pi_i - 1)(E[a_k|s_{ij} = 1] - \bar{\pi}) &= \mu_p(2\pi_i - 1)(\pi_{ij} - \bar{\pi})((1 - q)\tilde{\nu}_j + q\nu_{ij}) \\ &= a_j\pi_{ij}^a + b\nu_{ij}\pi_{ij}^a \end{aligned}$$

where  $a_j = \mu_p 2(1 - q)\tilde{\nu}_j > 0$  and  $b = \mu_p 2q > 0$ ,

We see that the persuasion payoff is increasing in  $\pi_{ij}$  if the sender favors Republicans and decreasing in  $\pi_{ij}$  if the sender favors Democrats. We can define  $\hat{\pi}_1^R$ ,  $\hat{\pi}_0^R$ , and  $\hat{\pi}_U^R$  such that a sender in state  $\sigma \in \{U, 0, 1\}$  who favors Republicans, shares if and only if  $\pi_{ij} \geq \hat{\pi}_\sigma^R$ .

We assume that parameters are such that  $\hat{\pi}_0^R = 1$  in equilibrium. This is consistent with the data: the informed senders (those who correctly determine veracity of all four tweets and are absolutely confident in their assessments) virtually never share false messages.

Symmetrically, we can define  $\hat{\pi}_1^D$  and  $\hat{\pi}_U^D$ , such that an informed sender favoring Democrats who knows the information is true, shares if and only if  $\pi_{ij} \leq \hat{\pi}_1^D$  and an uninformed sender favoring Democrats shares if and only if  $\pi_{ij} \leq \hat{\pi}_U^D$ . We derive these threshold values  $\hat{\pi}_1^R$ ,  $\hat{\pi}_U^R$ ,  $\hat{\pi}_1^D$ ,  $\hat{\pi}_U^D$  below.

**Reputation payoff:** In addition to trying to change the receivers's partisan beliefs, the sender also wants to convince the receivers that they are informed. The receivers form beliefs about the type of the sender based on what they shared and the information they hold about the truthfulness of the content. Receivers can be one of three types: (i) uninformed  $U$ , (ii) informed who know that the state is 1, and (iii) informed and know that the state is 0.

According to the Bayes rule, the receivers' belief that the sender is informed given that the sender shares a message is as follows:

$$P[\psi_i = I|s_{ij} = 1] = \frac{P[s_{ij} = 1|\psi_i = I]P[\psi_i = I]}{P[s_{ij} = 1|\psi_i = I]P[\psi_i = I] + P[s_{ij} = 1|\psi_i = U]P[\psi_i = U]}.$$

By definition,  $P[\psi_i = I] = q$ . Since only the informed senders who know that the state is 1 share and the receivers' belief that the state is 1 is  $\nu_{kj}$ , we have  $P[s_{ij} = 1|\psi_i = I] = \nu_{kj}P[s_{ij} = 1|\psi_i = 1]$ . We, therefore, obtain the following expression:

$$P[\psi_i = I|s_{ij} = 1] = \frac{\nu_{kj}qP[s_{ij} = 1|\psi_i = 1]}{\nu_{kj}qP[s_{ij} = 1|\psi_i = 1] + (1 - q)P[s_{ij} = 1|\psi_i = U]}.$$

The sender's reputation inferred by the receiver can take three different values depending on the receiver's type  $\nu_{kj}$ . We denote these values  $r_0, r_U, r_1$  (where  $r$  stands for reputation and the subscript is the receiver's type). Replacing the value of  $\nu_{kj}$  in the expression above, we obtain:

$$\begin{aligned} r_1 &= \frac{qP[s_{ij} = 1|\psi_i = 1]}{qP[s_{ij} = 1|\psi_i = 1] + (1 - q)P[s_{ij} = 1|\psi_i = U]}, \\ r_U &= \frac{q\nu_j^0 P[s_{ij} = 1|\psi_i = 1]}{q\nu_j^0 P[s_{ij} = 1|\psi_i = 1] + (1 - q)P[s_{ij} = 1|\psi_i = U]}, \\ r_0 &= 0. \end{aligned}$$

The expected reputation of a sender who chooses to share the message  $j$  (i.e.,  $s_{ij} = 1$ ) is, therefore, as follows (weights are according to the prevalence of each type):

$$(1 - q)r_U + q[\nu_{ij}r_1 + (1 - \nu_{ij})r_0] = (1 - q)r_U + q\nu_{ij}r_1.$$

We, therefore, obtain the following expression for the reputation payoff:

$$\mu_r (E[\psi_i = I|s_{ij} = 1] - q) = \text{const} + \mu_r q r_1 \nu_{ij}.$$

where  $\text{const} = \mu_r ((1 - q)r_U - q)$ . As  $\mu_r q r_1 > 0$ , this expression increases in  $\nu_{ij}$ . So, the sender who is informed that the news is truthful expects a higher image payoff than someone who is certain that the news is false.

**Signaling partisanship payoff:** Upon receiving the message, the receiver perceives it either as favoring Republicans or as favoring Democrats. They do not observe  $\pi_{ij}$ , but the realization of the message is informative on  $\pi_{ij}$ . We denote  $\bar{F}$  the posterior beliefs of the receiver about the distribution of  $\pi_{ij}$  of the sender after receiving a message favoring Republicans and  $\underline{F}$  the posterior after receiving a message favoring Democrats.

Using the notation  $\bar{S} = P[R|s_{ij} = 1]$  (belief that the sender is Republican when they share) when the message was perceived as Republican and  $\underline{S}$  when it was perceived as Democrat, we can write the expected payoff from signaling partisanship as:

$$\mu_s (2\pi_i - 1) \left( \pi_{ij} \bar{S} + (1 - \pi_{ij}) \underline{S} - \frac{1}{2} \right).$$

Given the symmetry of the problem, we can focus on equilibria where  $\underline{S} = 1 - \bar{S}$ , so that the signaling payoff can be expressed as:

$$\begin{aligned} &\mu_s (2\pi_i - 1) \left( \left( \frac{1}{2} - \pi_{ij} \right) + \bar{S} (2\pi_{ij} - 1) \right) \\ &= \mu_s \left( -2\pi_{ij}^a + 4\pi_{ij}^a \bar{S} \right). \end{aligned}$$

By applying the Bayes rule, we obtain

$$P[R|s_{ij} = 1] = \frac{\frac{1}{2} [(1-q)(1 - \bar{F}(\hat{\pi}_U^R)) + q\nu_{kj}(1 - \bar{F}(\hat{\pi}_1^R))]}{\frac{1}{2} [(1-q)(1 - \bar{F}(\hat{\pi}_U^R)) + q\nu_{kj}(1 - \bar{F}(\hat{\pi}_1^R))] + \frac{1}{2} [(1-q)\bar{F}(\hat{\pi}_U^D) + q\nu_{kj}\bar{F}(\hat{\pi}_1^D)]}.$$

The expected partisanship is, therefore, given by:

$$\begin{aligned}\bar{S} &= (1-q)s_U + q\nu_{ij}s_1 + q(1 - \nu_{ij})s_0 \\ &= (1-q)s_U + qs_0 + \nu_{ij}q(s_1 - s_0),\end{aligned}$$

with

$$\begin{aligned}s_0 &= \frac{1 - \bar{F}(\hat{\pi}_U^R)}{(1 - \bar{F}(\hat{\pi}_U^R)) + \bar{F}(\hat{\pi}_U^D)}, \\ s_1 &= \frac{1 - \bar{F}(\hat{\pi}_U^R) + q(\bar{F}(\hat{\pi}_U^R) - \bar{F}(\hat{\pi}_1^R))}{(1 - \bar{F}(\hat{\pi}_U^R)) + \bar{F}(\hat{\pi}_U^D) + q(\bar{F}(\hat{\pi}_U^R) - \bar{F}(\hat{\pi}_1^R)) + q(\bar{F}(\hat{\pi}_1^D) - \bar{F}(\hat{\pi}_U^D))}, \\ s_U &= \frac{1 - \bar{F}(\hat{\pi}_U^R) + q(\bar{F}(\hat{\pi}_U^R) - \nu^0\bar{F}(\hat{\pi}_1^R)) - q(1 - \nu^0)}{(1 - \bar{F}(\hat{\pi}_U^R)) + \bar{F}(\hat{\pi}_U^D) + q(\bar{F}(\hat{\pi}_U^R) - \nu^0\bar{F}(\hat{\pi}_1^R)) - q(1 - \nu^0) + q(\nu^0\bar{F}(\hat{\pi}_1^D) - \bar{F}(\hat{\pi}_U^D))}.\end{aligned}$$

The signaling payoff is, therefore, as follows:

$$\mu_s(2\pi_i - 1) \left( E[R|s_{ij} = 1] - \frac{1}{2} \right) = \text{const} + d\pi_{ij}^a + e\nu_{ij}\pi_{ij}^a,$$

where  $d = \mu_s(-2 + 4((1-q)s_U + qs_0))$ ,  $e = \mu_s 4q(s_1 - s_0)$ .

The sign of the last term  $e\nu_{ij}\pi_{ij}^a$  is determined by the sign of  $s_1 - s_0$ . We can easily show that:

$$s_1 < s_0 \Leftrightarrow \frac{\bar{F}(\hat{\pi}_1^D)}{1 - \bar{F}(\hat{\pi}_1^R)} > \frac{\bar{F}(\hat{\pi}_U^D)}{1 - \bar{F}(\hat{\pi}_U^R)}.$$

Note that the inequality  $s_1 < s_0$  can be expressed as:

$$(1 - \bar{F}(\hat{\pi}_U^R)) (\bar{F}(\hat{\pi}_1^D) - \bar{F}(\hat{\pi}_U^D)) > \bar{F}(\hat{\pi}_U^D) (\bar{F}(\hat{\pi}_U^R) - \bar{F}(\hat{\pi}_1^R)).$$

This is equivalent to:

$$\frac{\bar{F}(\hat{\pi}_1^D)}{1 - \bar{F}(\hat{\pi}_1^R)} > \frac{\bar{F}(\hat{\pi}_U^D)}{1 - \bar{F}(\hat{\pi}_U^R)}.$$

Given the symmetry, the right-hand side can be rewritten as:

$$\frac{\bar{F}(1 - \hat{\pi}_1^R)}{1 - \bar{F}(\hat{\pi}_1^R)} > \frac{\bar{F}(1 - \hat{\pi}_U^R)}{1 - \bar{F}(\hat{\pi}_U^R)}. \quad (10)$$

In order to prove that  $s_1 < s_0$ , we, therefore, need to show that  $H(x) = \frac{\bar{F}(1-x)}{1-\bar{F}(x)}$  is a decreasing



function of  $x$ .

Let us calculate its first derivative:

$$H'(x) = \frac{-\bar{f}(1-x)(1-\bar{F}(x)) + \bar{f}(x)\bar{F}(1-x)}{(1-\bar{F}(x))^2}. \quad (11)$$

According to Bayes rule, the posterior probability density equals

$$\bar{f}(\pi_i) = \frac{P[R|\pi_i]f(\pi_i)}{p[R]} = \frac{\pi_i f(\pi_i)}{1/2}.$$

Using this expression in condition (11), we have

$$H'(x) = 2 \frac{-(1-x)f(1-x)(1-\bar{F}(x)) + xf(x)\bar{F}(1-x)}{(1-\bar{F}(x))^2}.$$

As  $f$  is symmetric, we can further show that

$$H'(x) = 2f(x) \frac{-(1-x)(1-\bar{F}(x)) + x\bar{F}(1-x)}{(1-\bar{F}(x))^2}.$$

As  $-(1-x)(1-\bar{F}(x)) + x\bar{F}(1-x)$  is an increasing function which is equal to 0 when  $x = 1$ , we establish that  $H'(x) < 0$ .  $H$  is decreasing. Condition (10) is satisfied and  $s_1 - s_0 < 0$ .

**Equilibrium conditions:** We now characterize the equilibrium conditions for a Republican leaning sender, the conditions for a Democrat leaning naturally follow.

Values  $\hat{\pi}_1^R$  and  $\hat{\pi}_U^R$  are characterized by the indifference conditions below, capturing the fact the informed who knows the information is true is indifferent between sending or not if their partisanship is  $\hat{\pi}_1^R$ . Similarly, uninformed individual is indifferent between sending or not if their partisanship is  $\hat{\pi}_U^R$ . Note that if the sender does not share, receivers stay with their prior  $q$  that the sender is informed and  $\pi_k$  on the state of the world, so we have  $V_n = 0$ .

The following equations characterize the equilibrium:<sup>28</sup>

---

<sup>28</sup> $\hat{\xi}_j$  is a combination of  $\xi_j$  and the constant terms corresponding to the reputation and partisan payoffs.

$$\begin{aligned}
0 &= \hat{\xi}_j + \mu_p(2\pi_i - 1) \left[ ((1-q)\tilde{\nu}_j^R + q)(\hat{\pi}_1^R - \bar{\pi}) \right] + \mu_r [(1-q)r_U + qr_1 - q] \\
&+ \mu_s(2\pi_i - 1) \left( \left( \frac{1}{2} - \hat{\pi}_1^R \right) + ((1-q)s_U + qs_0 + q(s_1 - s_0))(2\hat{\pi}_1^R - 1) \right) \\
0 &= \hat{\xi}_j + \mu_p(2\pi_i - 1) \left[ ((1-q)\tilde{\nu}_j + q\nu_j^0)(\hat{\pi}_U - \bar{\pi}) \right] + \mu_r [(1-q)r_U + q\nu_j^0 r_1 - q] \\
&+ \mu_s(2\pi_i - 1) \left( \left( \frac{1}{2} - \hat{\pi}_U \right) + ((1-q)s_U + qs_0 + q\nu_j^0(s_1 - s_0))(2\hat{\pi}_U - 1) \right) \\
r_1 &= \frac{q(1 - F(\hat{\pi}_1^R))}{q(1 - F(\hat{\pi}_1^R)) + (1-q)(1 - F(\hat{\pi}_U^R))} \\
r_U &= \frac{q\nu_j^0(1 - F(\hat{\pi}_1^R))}{q\nu_j^0(1 - F(\hat{\pi}_1^R)) + (1-q)(1 - F(\hat{\pi}_U^R))} \\
s_0 &= \frac{1 - \bar{F}(\hat{\pi}_U^R)}{(1 - \bar{F}(\hat{\pi}_U^R)) + \bar{F}(\hat{\pi}_U^D)} \\
s_1 &= \frac{1 - \bar{F}(\hat{\pi}_U^R) + q(\bar{F}(\hat{\pi}_U^R) - \bar{F}(\hat{\pi}_1^R))}{(1 - \bar{F}(\hat{\pi}_U^R)) + \bar{F}(\hat{\pi}_U^D) + q(\bar{F}(\hat{\pi}_U^R) - \bar{F}(\hat{\pi}_1^R)) + q(\bar{F}(\hat{\pi}_1^D) - \bar{F}(\hat{\pi}_U^D))} \\
s_U &= \frac{1 - \bar{F}(\hat{\pi}_U^R) + q(\bar{F}(\hat{\pi}_U^R) - \nu^0 \bar{F}(\hat{\pi}_1^R)) - q(1 - \nu^0)}{(1 - \bar{F}(\hat{\pi}_U^R)) + \bar{F}(\hat{\pi}_U^D) + q(\bar{F}(\hat{\pi}_U^R) - \nu^0 \bar{F}(\hat{\pi}_1^R)) - q(1 - \nu^0) + q(\nu^0 \bar{F}(\hat{\pi}_1^D) - \bar{F}(\hat{\pi}_U^D))} \\
\tilde{\nu}_j &= \frac{q\nu_j^0(1 - F(\hat{\pi}_1^R)) + (1-q)\nu_j^0(1 - F(\hat{\pi}_U^R))}{q\nu_j^0(1 - F(\hat{\pi}_1^R)) + (1-q)(1 - F(\hat{\pi}_U^R))}.
\end{aligned}$$

## B.2 Discussion of existence of equilibria

We show the existence of an equilibrium in the case where  $\mu_s = 0$ . The equilibrium is then solution to the system of five non-linear equations described with five unknowns ( $r_1, r_U, \tilde{\nu}_j, \hat{\pi}_1, \hat{\pi}_U$ ) as described below:

$$\begin{aligned}
0 &= -r_1 + \frac{q(1 - F(\hat{\pi}_1))}{q(1 - F(\hat{\pi}_1)) + (1 - q)(1 - F(\hat{\pi}_U))} \\
0 &= -r_U + \frac{q\nu_j^0(1 - F(\hat{\pi}_1))}{q\nu_j^0(1 - F(\hat{\pi}_1)) + (1 - q)(1 - F(\hat{\pi}_U))} \\
0 &= -\tilde{\nu}_j + \frac{q\nu_j^0(1 - F(\hat{\pi}_1)) + (1 - q)\nu_j^0(1 - F(\hat{\pi}_U))}{q\nu_j^0(1 - F(\hat{\pi}_1)) + (1 - q)(1 - F(\hat{\pi}_U))} \\
0 &= \hat{\xi}_j + \mu_p(2\pi_i - 1) [((1 - q)\tilde{\nu}_j + q)(\hat{\pi}_1 - \bar{\pi})] + \mu_r [(1 - q)r_U + qr_1 - q] \\
0 &= \hat{\xi}_j + \mu_p(2\pi_i - 1) [((1 - q)\tilde{\nu}_j + q\nu_j^0)(\hat{\pi}_U - \bar{\pi})] + \mu_r [(1 - q)r_U + q\nu_j^0r_1 - q].
\end{aligned}$$

This system can be written  $H(x) = 0$ . To show the existence of a solution, we will use the Inverse Function theorem in a neighborhood of a particular vector  $x_0$ . We start by calculating the Jacobian of H given by:

$$J = \begin{pmatrix} -1 & 0 & 0 & f_1^1 & f_U^1 \\ 0 & -1 & 0 & f_1^2 & f_U^2 \\ 0 & 0 & -1 & f_1^3 & f_U^3 \\ bq & b(1 - q) & a(1 - q)(\hat{\pi}_1 - \bar{\pi}) & a((1 - q)\tilde{\nu}_j + q) & 0 \\ bq\nu_j^0 & b(1 - q) & a(1 - q)(\hat{\pi}_U - \bar{\pi}) & 0 & a((1 - q)\tilde{\nu}_j + q\nu_j^0) \end{pmatrix}$$

where  $a = \mu_p(2\pi_i - 1)$ ,  $b = \mu_r$  and  $f_i^1$  is the derivative of equation  $i$  with respect to  $\hat{\pi}_1$ .

We now derive conditions under which the Jacobian matrix  $J$  is invertible.  $J$  can be re-expressed as:

$$J = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

with  $A = -I$  and

$$B = \begin{pmatrix} f_1^1 & f_U^1 \\ f_1^2 & f_U^2 \\ f_1^3 & f_U^3 \end{pmatrix}$$

$$C = \begin{pmatrix} bq & b(1-q) & a(1-q)(\hat{\pi}_1 - \bar{\pi}) \\ bq\nu_j^0 & b(1-q) & a(1-q)(\hat{\pi}_U - \bar{\pi}) \end{pmatrix}$$

and

$$D = \begin{pmatrix} a((1-q)\tilde{\nu}_j + q) & 0 \\ 0 & a((1-q)\tilde{\nu}_j + q\nu_j^0) \end{pmatrix}$$

We have  $\det(J) = \det(A) \det(D - CA^{-1}B)$ . Since  $A = -I$ , we first calculate  $C * B$ .

$$C * B = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

with

$$\begin{aligned} a_{11} &= bqf_1^1 + b(1-q)f_1^2 + a(1-q)(\hat{\pi}_1 - \bar{\pi})f_1^3 \\ a_{12} &= bqf_U^1 + b(1-q)f_U^2 + a(1-q)(\hat{\pi}_U - \bar{\pi})f_U^3 \\ a_{21} &= bq\nu_j^0 f_1^1 + b(1-q)f_1^2 + a(1-q)(\hat{\pi}_1 - \bar{\pi})f_1^3 \\ a_{22} &= bq\nu_j^0 f_U^1 + b(1-q)f_U^2 + a(1-q)(\hat{\pi}_U - \bar{\pi})f_U^3. \end{aligned}$$

This implies that:

$$D - CA^{-1}B = \begin{pmatrix} a'_{11} & a'_{12} \\ a'_{21} & a'_{22} \end{pmatrix}$$

with

$$\begin{aligned} a'_{11} &= a_{11} + a((1-q)\tilde{\nu}_j + q) \\ a'_{12} &= a_{12} \\ a'_{21} &= a_{21} \\ a'_{22} &= a_{22} + a((1-q)\tilde{\nu}_j + q\nu_j^0) \end{aligned}$$

Using the fact  $a_{21} = a_{11} + bqf_1^1(\nu_j^0 - 1)$  and  $a_{12} = a_{22} + bqf_U^1(1 - \nu_j^0)$ , we can write the determinant of  $J$  as

$$\begin{aligned} \det(J) &= a'_{11}a'_{22} - a_{21}a_{12} \\ &= a_{11} [a((1-q)\tilde{\nu}_j + q\nu_j^0) - bqf_U^1(1 - \nu_j^0)] \\ &\quad + a_{22} [a((1-q)\tilde{\nu}_j + q) + bqf_1^1(1 - \nu_j^0)] \\ &\quad + a^2 [(1-q)\tilde{\nu}_j + q] [(1-q)\tilde{\nu}_j + q\nu_j^0] \\ &\quad - b^2 [qf_1^1(\nu_j^0 - 1)] [qf_U^1(1 - \nu_j^0)]. \end{aligned}$$

Consider a vector  $x^0 = (r_1, r_U, \tilde{\nu}_j, \hat{\pi}_1, \hat{\pi}_U)$  with  $F(\hat{\pi}_U) = 1$ . We have:

$$f_1^1(x^0) = \frac{-q(1-q)f(\hat{\pi}_1)(1-F(\hat{\pi}_U))}{(q(1-F(\hat{\pi}_1)) + (1-q)(1-F(\hat{\pi}_U)))^2} = 0$$

$$f_1^U(x^0) = \frac{q(1-q)f(\hat{\pi}_U)(1-F(\hat{\pi}_1))}{(q(1-F(\hat{\pi}_1)) + (1-q)(1-F(\hat{\pi}_U)))^2} > 0$$

Similarly, we can show  $f_1^2(x^0) = f_1^3(x^0) = 0$  and  $f_U^2(x^0) > 0$ ,  $f_1^3(x^0) > 0$ .

Overall, we can rewrite:

$$\det(J) = a_{22} [a((1-q)\tilde{\nu}_j + q)] + a^2 [(1-q)\tilde{\nu}_j + q] [(1-q)\tilde{\nu}_j + q\nu_j^0] > 0.$$

Thus, given that  $\det(J) > 0$ ,  $f$  is invertible near  $x_0$ . According to the Inverse Function Theorem, the equation  $f(x) = 0$  has a solution near  $f(x_0)$ .

## C Procedure to recover the parameters of the model

We explain in this section how we recover the parameters of the model for a given set of estimates for  $\alpha_1$ ,  $\alpha_2$  and  $\alpha_3$ . The system of equations that characterizes the equilibrium is as follows.<sup>29</sup>

$$0 = \hat{\xi}_j + \mu_p(2\pi_i - 1) [((1-q)\nu^0 + q)(\hat{\pi}_1 - \bar{\pi})] + \mu_r [(1-q)r_U + qr_1 - q] + \mu_s(2\pi_i - 1) \left( \left( \frac{1}{2} - \hat{\pi}_1 \right) + ((1-q)s_U + qs_0 + q(s_1 - s_0))(2\hat{\pi}_1 - 1) \right) \quad (\text{C.1})$$

$$0 = \hat{\xi}_j + \mu_p(2\pi_i - 1) [((1-q)\nu^0 + q\nu_j^0)(\hat{\pi}_U - \bar{\pi})] + \mu_r [(1-q)r_U + q\nu_j^0 r_1 - q] + \mu_s(2\pi_i - 1) \left( \left( \frac{1}{2} - \hat{\pi}_U \right) + ((1-q)s_U + qs_0 + q\nu_j^0(s_1 - s_0))(2\hat{\pi}_U - 1) \right) \quad (\text{C.2})$$

$$0 = -r_1 + \frac{q(1 - F(\hat{\pi}_1))}{q(1 - F(\hat{\pi}_1)) + (1-q)(1 - F(\hat{\pi}_U))} \quad (\text{C.3})$$

$$0 = -r_U + \frac{q\nu_j^0(1 - F(\hat{\pi}_1))}{q\nu_j^0(1 - F(\hat{\pi}_1)) + (1-q)(1 - F(\hat{\pi}_U))} \quad (\text{C.4})$$

$$0 = -s_0 + \frac{1 - \bar{F}(\hat{\pi}_U)}{(1 - \bar{F}(\hat{\pi}_U)) + \bar{F}(1 - \hat{\pi}_U)} \quad (\text{C.5})$$

$$0 = -s_1 + \frac{1 - \bar{F}(\hat{\pi}_U) + q(\bar{F}(\hat{\pi}_U) - \bar{F}(\hat{\pi}_1))}{(1 - \bar{F}(\hat{\pi}_U)) + \bar{F}(1 - \hat{\pi}_U) + q(\bar{F}(\hat{\pi}_U) - \bar{F}(\hat{\pi}_1)) + q(\bar{F}(1 - \hat{\pi}_1) - \bar{F}(1 - \hat{\pi}_U))} \quad (\text{C.6})$$

$$0 = -s_U + \frac{1 - \bar{F}(\hat{\pi}_U) + q(\bar{F}(\hat{\pi}_U) - \nu^0 \bar{F}(\hat{\pi}_1)) - q(1 - \nu^0)}{(1 - \bar{F}(\hat{\pi}_U)) + \bar{F}(1 - \hat{\pi}_U) + q(\bar{F}(\hat{\pi}_U) - \nu^0 \bar{F}(\hat{\pi}_1)) - q(1 - \nu^0) + q(\nu^0 \bar{F}(1 - \hat{\pi}_1) - \bar{F}(1 - \hat{\pi}_U))} \quad (\text{C.7})$$

$$\alpha_1 = \mu_r q r_1 \quad (\text{C.8})$$

$$\alpha_2 = \mu_p 2(1-q)\nu^0 + \mu_s(-2 + 4((1-q)s_U + qs_0)) \quad (\text{C.9})$$

$$\alpha_3 = \mu_p 2q + \mu_s 4q(s_1 - s_0) \quad (\text{C.10})$$

In addition, there are constraints on the parameters ( $r_\sigma, s_\sigma, \hat{\pi}_1, \hat{\pi}_U$  are contained in the interval  $[0, 1]$ ) and on their relative rankings ( $s_1 < s_U < s_0, \hat{\pi}_1 < \hat{\pi}_U$  and  $r_U < r_1$ ).

There are 10 equations and 13 unknowns (given that we estimate  $\alpha_1, \alpha_2, \alpha_3$ ). We calibrate two parameters using the data from the experiment and one parameter using an external data source. In particular, we set the share of informed individuals  $q$  equal to 0.04, which is the share of the participants in our experiment who evaluated the veracity of all four tweets correctly with certainty. We set the prior belief on the veracity of tweets in general  $\nu_0$  equal to 0.7, which is the average evaluation of veracity in the control group for three out of four our tweets. We exclude the tweet about the ban on condoms (False Cult) because we want to assess the belief on a randomly drawn content on social media and we recognize that this particular tweet is non-representative. Finally, we set the equilibrium value of  $s_U$ , i.e., the belief of an uninformed receiver that the sender is Republican, to be equal to 0.4, which is the share of U.S. adult Twitter users who report that they do not lean towards the Democratic party according to the Pew Research Center.<sup>30</sup>

<sup>29</sup>It is the system of equation derived in the proof of Proposition 1, characterizing a Republican sender. As discussed in the proof, the system is the same for a Democrat sender with  $\pi_\sigma^R = 1 - \pi_\sigma^D$ . We drop the index R. Note also that, to simplify the estimation that already includes 10 nonlinear equations, we assume that  $\tilde{\nu}_j = \nu^0$  thus removing one equation.

<sup>30</sup>See <https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/>, accessed November 27,

We solve the system using an algorithm that minimizes the sum of squared errors made on each equation in the system above. We use the limited-memory BFGS (L-BFGS), implemented using the Python package *scipy*. This method does not explicitly compute the Hessian matrix but uses an estimate of the matrix to steer its search through variable space.<sup>31</sup> Since we have no guidance on the initial values do a grid search based on initial conditions keeping only sets of parameters that satisfy the constraints. We pick the solution that achieves the lowest sum of squared errors.

We present in the Table A6 below the solution of the system using values of  $\alpha_i$  in the control group, recovered from the discrete choice estimation presented in Table 3.

**Table A6:** The solution to the system of equations (C.1) – (C.10)

Parameter	Description	Value
<b>Equilibrium Solution</b>		
$\hat{\pi}_U$	sharing cutoff for sender in state U	0.62
$\hat{\pi}_1$	sharing cutoff for sender in state 1	0.07
$r_U$	sender’s reputation for receiver in state U	0.08
$r_1$	sender’s reputation for receiver in state 1	0.12
$s_0$	sender’s partisan image for receiver in state 0	0.41
$s_1$	sender’s partisan image for receiver in state 1	0.39
$\hat{\xi}_j$	benefit of sharing	-8.40
$\mu_r$	weight on reputation payoff	208.05
$\mu_p$	weight on persuasion payoff	12.96
$\mu_s$	weight on partisan signaling payoff	42.99
<b>Estimated Parameters</b>		
$\alpha_1$	weight on veracity	1.041
$\alpha_2$	weight on partisan alignment	0.271
$\alpha_3$	weight on interaction	0.932
<b>Calibrated Parameters</b>		
$q$	proportion of informed senders	0.04
$\nu_0$	prior belief on veracity	0.7
$s_U$	sender’s partisan image for receiver in state U	0.4

We point out that there could be multiple solutions to the system of equations, but that, when we compute Figure 4 for other solutions identified by the solver (with higher MSE), the figure is remarkably stable.

2023.

<sup>31</sup>The method is well suited for systems with a large number of unknowns.

## D Shapley Value Decomposition

Denote channels of influence by: 1 = Updating; 2 = Saliency; 3 = Cost.

Denote also:

- $T$  - Treatments;
- $V$  - Sharing;
- $V_0$  - Sharing under no policy;
- $V_{123}^T$  - Reduced-form sharing under treatment  $T$ , i.e., when we allow all three channels;
- $V_1^T$  - Simulated sharing under treatment  $T$  when allow channel 1 only;
- $V_2^T$  - Simulated sharing under treatment  $T$  when allow channel 2 only;
- $V_3^T$  - Simulated sharing under treatment  $T$  when allow channel 3 only;
- $V_{12}^T$  - Simulated sharing under treatment  $T$  when allow two channels: 1 and 2;
- $V_{23}^T$  - Simulated sharing under treatment  $T$  when allow two channels: 2 and 3;
- $V_{13}^T$  - Simulated sharing under treatment  $T$  when allow two channels: 1 and 3.

The total average treatment effect of the treatment  $T$ , therefore, is equal to:  $ATE^T = V_{123}^T - V_0$ .

It can be decomposed into the effects of each of the three channels:

$$ATE^T = V_{123}^T - V_0 = \phi_1^T + \phi_2^T + \phi_3^T,$$

where:

$$\phi_1^T = \frac{1}{3}(V_{123}^T - V_{23}^T) + \frac{1}{6}(V_{12}^T - V_2^T) + \frac{1}{6}(V_{13}^T - V_3^T) + \frac{1}{3}(V_1^T - V_0),$$

$$\phi_2^T = \frac{1}{3}(V_{123}^T - V_{13}^T) + \frac{1}{6}(V_{12}^T - V_1^T) + \frac{1}{6}(V_{23}^T - V_3^T) + \frac{1}{3}(V_2^T - V_0),$$

$$\phi_3^T = \frac{1}{3}(V_{123}^T - V_{12}^T) + \frac{1}{6}(V_{13}^T - V_1^T) + \frac{1}{6}(V_{23}^T - V_2^T) + \frac{1}{3}(V_3^T - V_0).$$



## E Simulating digital literacy training

We assume that  $\mu_r$ ,  $\mu_p$  and  $\mu_s$  take the values estimated for the baseline value of  $q$  ( $q = 0.04$ ), according to the methodology described in Section C. For each counterfactual value of  $q$ , we solve the following system of equations (E.1)-(E.7) with unknowns:  $(r_1, r_U, s_0, s_U, s_1, \hat{\pi}_U, \hat{\pi}_1)$ . Note that the first two equations (E.1) and (E.2) are modified versions of (C.1) and (C.2), where we take into account that  $\hat{\xi}_j$  is a function of  $q$ .<sup>32</sup> Specifically, we fix  $\hat{\xi}_j$  at its value for  $q = 0.04$  and include the extra term  $\mu_r r_U(0.04 - q)$ , i.e., the constant term corresponding to the reputation payoff as derived in the proof of Proposition 1.

$$0 = \hat{\xi}_j + \mu_r r_U(0.04 - q) + \mu_p(2\pi_i - 1) [((1 - q)\nu^0 + q)(\hat{\pi}_1 - \bar{\pi})] + \mu_r [(1 - q)r_U + qr_1 - q] + \mu_s(2\pi_i - 1) \left( \left( \frac{1}{2} - \hat{\pi}_1 \right) + ((1 - q)s_U + qs_0 + q(s_1 - s_0))(2\hat{\pi}_1 - 1) \right) \quad (\text{E.1})$$

$$0 = \hat{\xi}_j + \mu_r r_U(0.04 - q) + \mu_p(2\pi_i - 1) [((1 - q)\nu^0 + q\nu_j^0)(\hat{\pi}_U - \bar{\pi})] + \mu_r [(1 - q)r_U + q\nu_j^0 r_1 - q] + \mu_s(2\pi_i - 1) \left( \left( \frac{1}{2} - \hat{\pi}_U \right) + ((1 - q)s_U + qs_0 + q\nu_j^0(s_1 - s_0))(2\hat{\pi}_U - 1) \right) \quad (\text{E.2})$$

$$0 = -r_1 + \frac{q(1 - F(\hat{\pi}_1))}{q(1 - F(\hat{\pi}_1)) + (1 - q)(1 - F(\hat{\pi}_U))} \quad (\text{E.3})$$

$$0 = -r_U + \frac{q\nu_j^0(1 - F(\hat{\pi}_1))}{q\nu_j^0(1 - F(\hat{\pi}_1)) + (1 - q)(1 - F(\hat{\pi}_U))} \quad (\text{E.4})$$

$$0 = -s_0 + \frac{1 - \bar{F}(\hat{\pi}_U)}{(1 - \bar{F}(\hat{\pi}_U)) + \bar{F}(1 - \hat{\pi}_U)} \quad (\text{E.5})$$

$$0 = -s_1 + \frac{1 - \bar{F}(\hat{\pi}_U) + q(\bar{F}(\hat{\pi}_U) - \bar{F}(\hat{\pi}_1))}{(1 - \bar{F}(\hat{\pi}_U)) + \bar{F}(1 - \hat{\pi}_U) + q(\bar{F}(\hat{\pi}_U) - \bar{F}(\hat{\pi}_1)) + q(\bar{F}(1 - \hat{\pi}_1) - \bar{F}(1 - \hat{\pi}_U))} \quad (\text{E.6})$$

$$0 = -s_U + \frac{1 - \bar{F}(\hat{\pi}_U) + q(\bar{F}(\hat{\pi}_U) - \nu^0 \bar{F}(\hat{\pi}_1)) - q(1 - \nu^0)}{(1 - \bar{F}(\hat{\pi}_U)) + \bar{F}(1 - \hat{\pi}_U) + q(\bar{F}(\hat{\pi}_U) - \nu^0 \bar{F}(\hat{\pi}_1)) - q(1 - \nu^0) + q(\nu^0 \bar{F}(1 - \hat{\pi}_1) - \bar{F}(1 - \hat{\pi}_U))} \quad (\text{E.7})$$

We numerically solve this system of equations for different values of  $q$ , using the same algorithm as described above. We use as a starting value the solution for the closest value of  $q$  already estimated.

### Sender-knowledge channel

To capture the fact that digital literacy directly affects the knowledge of senders, we assume the change in  $q$  would imply a change in the fraction of informed individuals, using the same definition as in the baseline: an informed individual is the one who evaluates correctly the veracity of all four tweets. Specifically, for a counterfactual value of  $q$ , we randomly select a proportion  $(q - 0.04)$  respondents (excluding those who appear already perfectly informed in our sample), and change their beliefs about veracity assigning a value of 100 to the two correct tweets and 0 to the two false tweets. Using these updated values, we recompute sharing.

### Receiver-reaction channel

Using the solution of the system of equations above, we can then compute  $\alpha_1$ ,  $\alpha_2$  and  $\alpha_3$  for

---

<sup>32</sup>For  $q = 0.04$ , we did not take into account this relationship and aggregated all the constant terms in the variable  $\hat{\xi}_j$ .

the different values of  $q$  using the system of equations below:

$$\begin{aligned}\alpha_1 &= \mu_r q r_1 \\ \alpha_2 &= \mu_p 2(1-q)\nu^0 + \mu_s (-2 + 4((1-q)s_U + q s_0)) \\ \alpha_3 &= \mu_p 2q + \mu_s 4q(s_1 - s_0).\end{aligned}$$

Finally, using these values of  $\alpha_1$ ,  $\alpha_2$  and  $\alpha_3$ , we also compute the counterfactual sharing. We recompute sharing of the different tweets using the new values of  $\alpha_i$  and subsequently calculate the inclusive utility. We use the upper-nest coefficients estimated for the baseline in Table 4.