Social Media and Democracy

Ronen Gradwohl, Yuval Heller, and Arye Hillman

Abstract

We study the ability of a social media platform with a political agenda to influence voting outcomes. Our benchmark is Condorcet's jury theorem, which states that the likelihood of a correct decision under majority voting increases with the number of voters. We show how information manipulation by a social media platform can overturn the jury theorem, thereby undermining democracy. We also show that sometimes the platform can do so only by providing information that is biased in the *opposite direction* of its preferred outcome. Finally, we compare manipulation of voting outcomes by social media platforms to manipulation by traditional media channels.

1 Introduction

Information influences voters. Traditionally, relevant information is provided publicly to the electorate by mass media through political advertising and endorsements. Increasingly, it is also delivered privately to individual voters by social media. Rather than being exposed to the same, public information, users of social media face private persuasion through algorithms that select and order the information they receive. For example, Facebook's News Feed algorithm determines personally for each user which posts are visible and the order in which they appear [13]. Twitter also uses machine learning algorithms to select and order tweets, and additionally adds warning labels and sometimes blocks information dissemination by locking accounts [36, 35]. Because of the vast reach and impact of social media, the public discourse surrounding political biases in such machine learning algorithms has intensified. These biases, as well as their effects on Facebook and Twitter users, have also been identified empirically (see, e.g., [27, 21]).

In this paper, we study how and to what extent a social media platform can influence voters through such private persuasion. Our beginning is Condorcet's jury theorem [10], a theorem that demonstrates the merit of democracy by showing that a group of imperfectly informed voters is more likely to choose a socially preferred alternative by majority voting than when one voter makes the decision.[1] In our study, imperfectly informed voters obtain additional information through messages from a social media platform. The platform has a designer who commits to a messaging scheme, a commitment that mirrors the algorithms used in practice to select and order the information that individuals obtain. On receiving the additional information from the platform, voters use Bayes' rule to update their beliefs about the likelihood that a candidate or policy is preferable.

An important feature of our study is that the designer can credibly commit to how messages are generated. If the designer commits to always revealing true information about which alternative is better for society, voters can take the information at face value. However, the designer may want a particular outcome for ideological reasons, without regard for which alternative is better for society. The designer may thus benefit from committing to a biased

[1] More specifically, in the setting of the theorem, voters face a choice between two alternatives, one of which is objectively better for all voters. Voters do not know a priori with certainty which is the better alternative, and each voter's limited information is expressed in probabilities about which alternative is better. The theorem states that the socially preferred alternative is more likely to be chosen by majority voting than when a single individual makes the decision, and, as the number of voters increases, the likelihood of the correct decision under majority voting increases and approaches certainty.

scheme. Voters are aware that the scheme is biased, but are individually better off taking the additional information into account when deciding how to vote.

Another feature of our study is that voters know that their single vote has a negligible chance of being decisive, and vote sincerely in favor of the alternative they believe is more likely to be better. This distinguishes our theory from a substantial literature in which voters vote strategically in anticipation that their personal vote will be pivotal in determining the voting outcome.[2] Sincere voting, which is ostensibly expressive, allows us to characterize the direct effects of persuasion on voting outcomes, and to identify the properties of optimal ways for a designer using social media to influence voters.

Our first conclusion is that the designer can often design the scheme so that majority voting results in the choice of the alternative preferred by the social media platform rather than the one preferred by voters. This implies that Condorcet's jury theorem is overturned—elections decided by majority voting are less likely to result in the socially preferred outcome, compared to a single voter making the decision alone. Furthermore, the designer can do this with a very simple scheme—one that, for each voter, yields a conditionally independent message from the same distribution. We next turn to study the properties of optimal messaging schemes, with a focus on their biases. Perhaps surprisingly, we show that the best way for the designer to tilt a majority outcome towards his ideologically preferred alternative may be to commit to a scheme that is biased in the *opposite direction* of this alternative.[3]

Section 2 below surveys the related literature and Section 3 describes the formal model. These are followed by our results in Section 4 and extensions in Section 5. Finally, expository examples are contained in the full version of this paper, as are all proofs [18].

2 Related Literature

The literature studying social media and transmission of information to voters through the lens of Bayesian persuasion and information design has focused on studying the extent to which a designer can manipulate others by selective disclosure of information [23]. In the terminology of this literature, we study private Bayesian persuasion with multiple informed receivers. We also restrict our designer to sending conditionally independent signals from the same signal structure. Within the literature on Bayesian persuasion, the paper most-closely related to ours is [25] (see also [26]). In this paper, the author studies a Bayesian persuasion problem with a single, informed receiver. The model is different from ours, and assumes that the state space is a full-support interval. However, [25] does present a school-employer example that is similar to a single-voter voting game. Our paper may be viewed as extending that example to a model with many receivers, and analyzing the effects on voting. We show that, unlike the case of a single voter, when there are many voters the designer can often completely overturn the outcome of the vote.

A main assumption in Bayesian persuasion is that the designer can commit to the information structure. [28] study the credibility of such commitment, and define a credible information structure as one from which the designer has no incentive to deviate given the receivers' reactions and given that the frequency of messages must not change. Our designer's signal structure is credible in this sense: If the structure is optimal, then the designer has no incentive to deviate because the designer is already getting the best outcome. If the structure is not optimal, then no other signal structure can yield higher utility.

[2] We survey this literature in Section 2.

[3] [21] show that Twitter is biased towards the right, as it amplifies messages from the mainstream political right more than the mainstream political left. In the context of our setting, this might be consistent with a desire to tilt public opinion towards the *left*.

A number of papers study Bayesian persuasion with receivers who subsequently vote in an election. The bulk of this research focuses on strategic voters—voters who condition their behavior on being pivotal or decisive in the election—and the conclusions rely critically on the designer's ability to exploit this behavior (see next paragraph for details).[4] By contrast, we suppose voters vote according to their own information for the outcome that they believe is better ("sincere" voting), and do not regard themselves as possibly decisive or pivotal. Rather, we suppose that voters enjoy an expressive benefit from voting for the correct alternative.[5] This allows us to characterize the direct effects of persuasion on voting outcomes, and to identify the properties of optimal persuasion.

[6] study strategic voting in the context of Condorcet's jury theorem; they describe strategic voters conditioning their behavior on being pivotal in an election, and show that it may no longer be personally optimal to vote sincerely. [15] show that under general voting rules, such strategic voting may lead to a failure of information aggregation, but [14] show that, nonetheless, the jury theorem holds: the probability that the outcome of the vote under majority voting is correct approaches certainty as the number of voters increases. [19] start with the general strategic voting framework of [14], and add an information designer. They show that, under strategic voting, the designer can provide additional information in a way that leads to any preferred outcome with high probability. [34] studies a symmetric two-voter game, and shows that, in such a setting, it is never optimal for the designer to commit to a conditionally independent signal structure. More generally, [3] study a general framework for private persuasion, and characterize optimal solutions under submodularity assumptions on the designer's utility. In our setting, the utility function is neither supermodular nor submodular, and so [3]'s results do not apply. [24] focus on persuading uninformed voters through correlated signal structures that induce sincere rather than strategic voting in equilibrium. Finally, [9] study optimal manipulation of strategic costly voting and [7] study optimal manipulation of a heterogeneous group of receivers, where the sender can design the information presented to each group member.

Another branch of the literature on Bayesian persuasion has studied public persuasion, where the designer commits to a signal structure and all voters observe the same realization, as opposed to private persuasion through social media where each voter obtains a personally distinct information realization. Public Bayesian persuasion is an appropriate model for traditional media such as newspapers and television. Private persuasion is appropriate for social media, which can use personal information garnered from individuals' browsing history and communications in order to design selective information directed at an individual. The literature includes [2], who study public persuasion of heterogeneous voters. [37] compares public signal structures to private, conditionally independent signal structures, and shows that the former are more informative (and hence worse for the designer) than the latter. [17] also study public persuasion, with a focus on the interplay between the bias of the signal structure and voters' polarization. While less related to our work in terms of focus and results, the model of [17] is similar to ours in that there is a continuum of voters and each voter votes sincerely.

A sizable literature has studied bias in traditional and social media, both theoretically and empirically (see [16, 30, 32] for a variety of surveys). Much of this literature focuses on the informational effects of media bias, as well as its welfare implications. To the best of our knowledge, our insight that the information designer's optimal signal structure is biased

[4] See, for example, [29] and [31] and the references therein for the empirical literature on strategic voting. Although much of the literature detects some amount of strategic voting, in most instances the large majority of voters are not strategic.

[5] A paradox of voting is present [12] if voters have a negligible probability of being pivotal; what reason does the voter have to vote? A benefit imputed to voting is expressive utility [20]. The act of voting is used to express an identity or sense of belonging. In this context, voting is rational.

away from the intended outcome is novel.[6]

Lastly, in other related literature, [11] study bias in a social media platform when those it targets also have exogenous information, and share this information via a social network. They study the effects of the network's connectivity when voters suffer from correlation neglect on the optimal bias of the social media platform. [33] study the effects of public persuasion and cheap talk on turnout in a model with costly voting. Relatedly, [27] performs a field experiment to study the effects of social media (specifically, Facebook) on user's political leanings, and [21] study biases in the amplification of political messages on Twitter.

3 Model

There are two policies (or candidates) denoted by A and B, and a continuum of voters who must decide between them. There are two equally likely states of the world, $\Theta = \{\theta_A, \theta_B\}$, such that the better policy is A in state θ_A and B in state θ_B. Each voter receives a symmetric binary signal about the state of the world. Signal realizations are either a or b, and a signal of accuracy q is one where

$$\mathrm{P}\left[a|\theta_A\right] = \mathrm{P}\left[b|\theta_B\right] = q.$$

We allow heterogeneity in the levels of signal accuracy. Specifically, we assume that a λ-share (where $\lambda \in (0,1)$) of the population has low signal accuracy $q_\ell \in [0.5, 1)$, and the rest have high signal accuracy $q_h \in [q_\ell, 1]$. We say that the population is *homogeneous* if $q_\ell = q_h$, and that it is *heterogeneous* otherwise.[7]

In addition to the voters, there is an information designer. The designer can send the voters additional informative signals that depend on the state of the world. Unlike the voters, however, the designer wants to maximize the probability that policy A is chosen, regardless of the state. In particular, the designer's utility is 1 whenever A is chosen, and 0 otherwise. In order to influence the voters, the designer chooses a binary signal structure s—a probabilistic function from the state to a pair $\{\mathfrak{a}, \mathfrak{b}\}$—and sends each voter a conditionally independent realization of s.

We note that a more general signal structure of the designer could potentially make use of correlated messages and different distributions for different voters.[8] However, we restrict ourselves to simple signal structures, in which messages are conditionally independent and drawn from the same distribution of binary signals. We will show that even such simple signal structures can suffice for the designer to overturn the outcome of majority voting.

We interpret this signal as having been sent through a social media platform that is controlled by the information designer. As in the Bayesian persuasion framework [23], we assume that the designer commits to s prior to learning the true state of the world. In addition, and without loss of generality, we assume that the signal structure s is such that $\mathrm{P}\left[\theta_A|\mathfrak{a}\right] \geq \mathrm{P}\left[\theta_A|\mathfrak{b}\right]$. If this were not the case, one could simply switch the meanings of the two signal realizations.

Remark 3.1. The assumption that the designer's signal structure is binary is without loss of generality, given that the set of actions per voter is binary. We prove this formally in the full version of the paper [18].

[6] We are aware of only one Bayesian persuasion model with negatively biased signal structures, [8], in which such signal structures are used by duopolistic sellers in order to soften price competition.

[7] In Section 5 we allow more general distributions over signal accuracy levels and extend some of our results to this setup.

[8] Most of the previous literature on persuading voters does, in fact, utilize carefully correlated messages. See Section 2 for details.

Define the *bias* of the designer's signal structure as the weighted difference in the sizes of the "lies" in the two directions, namely,

$$\mathrm{bias}(s) = 2 \cdot (\mathrm{P}\left[s(\theta_B) = \mathfrak{a}\right] \cdot \mathrm{P}\left[\theta_B\right] - \mathrm{P}\left[s(\theta_A) = \mathfrak{b}\right] \cdot \mathrm{P}\left[\theta_A\right]) = \mathrm{P}\left[s(\theta_B) = \mathfrak{a}\right] - \mathrm{P}\left[s(\theta_A) = \mathfrak{b}\right],$$

where the expression after the first equality presents the definition of a general prior, and the expression after the latter equality relies on having a uniform prior. Thus, under the uniform prior, symmetric signal structures in which $\mathrm{P}\left[\mathfrak{a}|\theta_A\right] = \mathrm{P}\left[\mathfrak{b}|\theta_B\right]$ are *unbiased*—that is, they have bias 0. Positive biases correspond to signal structures that are more likely to yield the signal $\mathfrak{a}$, and negative biases to those that are more likely to yield the signal $\mathfrak{b}$. Observe that for the canonical example of Bayesian persuasion, in which a prosecutor seeks to convince a judge that a defendant is guilty (which corresponds to state θ_A), the optimal signal structure is one with positive bias (22). We note that this definition of bias is an instance of the definition of [16] for media bias, under the assumption that the fully informative signal structure in which $\mathrm{P}\left[s(\theta_A) = \mathfrak{a}\right] = \mathrm{P}\left[s(\theta_B) = \mathfrak{b}\right] = 1$ has bias 0.

A strategy σ_i of voter i is a function from signals (both the original signal and the designer's signal) to a distribution over two actions—a vote for A or a vote for B. The outcome of the vote is then determined by the majority, with ties decided in favor of A.[9] We assume that voting is *sincere*: each voter votes according to her own information, which includes the prior and the two signals. Sincere voting can either directly reflect the voters' preferences (i.e., each voter obtains utility 1 if she votes for the better policy), or it can correspond to voters who care for the selected outcome (i.e., each voter obtains utility 1 if the better policy is chosen), yet when voting they do not take into account the strategic effects of pivotality.[10]

Thus, each voter votes for A if her posterior on state θ_A is above 1/2 and votes for B if her posterior is below 1/2. As is standard in the Bayesian persuasion literature, we assume that voters vote for A, the outcome preferred by the designer, when their posterior is exactly 1/2; however, the exact tie-breaking rule is unimportant for our results.

Observe that if the designer sends an uninformative signal and either $q_\ell > 0.5$ or $\lambda < 0.5 < q_h$, then A is chosen in state θ_A and B in state θ_B. (In the complementary case where either $q_h = 0.5$ or $q_\ell = 0.5 < \lambda$, outcome A is chosen in both states of the world, due to our tie-breaking rule.) Furthermore, due to the population being a continuum and signals being private and conditionally independent, the outcome of the vote is deterministic, conditional on the state of the world. This means that it is either the case that the designer *cannot manipulate the outcome* (in which case A is still chosen only in state θ_A, regardless of the designer's signal structure), or that she can *perfectly manipulate* the outcome (in which case A is chosen in both states). In the latter case, we say that a designer's signal structure s is *optimal* if it induces the choice of policy A in both states. Note that there might be multiple optimal signal structures.

4 Results

4.1 Preliminaries

Suppose that a voter's belief that the state is θ_A is equal to 1/2. Then any additional signal structure $s : \Theta \mapsto \{\mathfrak{a}, \mathfrak{b}\}$ induces a pair of posteriors (α, β), where $\alpha = \mathrm{P}\left[\theta_A|\mathfrak{a}\right]$ and $\beta = \mathrm{P}\left[\theta_A|\mathfrak{b}\right]$. The signal structure is uninformative if $\alpha = \beta = 0.5$. If the signal structure is

[9] Tie-breaking in favor of A is assumed for simplicity only—the results are nearly identical for other tie-breaking rules; see Remark 4.1.

[10] The probability of being pivotal is decreasing in the population size (for finite populations), and it is equal to zero in our continuum-population model.

informative we can assume without loss of generality that $0 \leq \beta < \frac{1}{2} < \alpha \leq 1$, and in this case there exists a (unique) signal structure that induces these posteriors. Thus, we can identify any informative signal structure with the pair of posterior distributions it induces when starting with prior[11] $\frac{1}{2}$. For any $\beta < \frac{1}{2} < \alpha$, the exact mapping can be derived as follows: First, since the expected prior is equal to the posterior, we have

$$\alpha p_{\mathfrak{a}} + \beta(1 - p_{\mathfrak{a}}) = \frac{1}{2} \quad \Rightarrow \quad p_{\mathfrak{a}} = \frac{\frac{1}{2} - \beta}{\alpha - \beta}, \quad p_{\mathfrak{b}} = \frac{\alpha - \frac{1}{2}}{\alpha - \beta},$$

where $p_{\mathfrak{a}} = \mathrm{P}[\mathfrak{a}]$ and $p_{\mathfrak{b}} = \mathrm{P}[\mathfrak{b}] = 1 - p_a$. Next, we have that

$$\mathrm{P}[s(\theta_A) = \mathfrak{a}] = \mathrm{P}[\mathfrak{a}|\theta_A] = \frac{\mathrm{P}[\theta_A|\mathfrak{a}]\, p_{\mathfrak{a}}}{\mathrm{P}[\theta_A]} = \frac{\alpha p_{\mathfrak{a}}}{1/2} = \frac{\alpha - 2\alpha\beta}{\alpha - \beta} \quad \Rightarrow \quad \mathrm{P}[\mathfrak{b}|\theta_A] = \frac{2\alpha\beta - \beta}{\alpha - \beta}.$$

Similarly,

$$\mathrm{P}[s(\theta_B) = \mathfrak{a}] = \mathrm{P}[\mathfrak{a}|\theta_B] = \frac{\mathrm{P}[\theta_B|\mathfrak{a}]\, p_{\mathfrak{a}}}{1/2} = \frac{1 - 2\beta - \alpha + 2\alpha\beta}{\alpha - \beta} \quad \Rightarrow \quad \mathrm{P}[\mathfrak{b}|\theta_B] = \frac{2\alpha + \beta - 1 - 2\alpha\beta}{\alpha - \beta}.$$

Furthermore, this analysis implies that the bias of the signal structure (α, β) is:

$$\mathsf{bias}(\alpha, \beta) = \mathrm{P}[s(\theta_B) = \mathfrak{a}] - \mathrm{P}[s(\theta_A) = \mathfrak{b}] = \frac{1 - \beta - \alpha}{\alpha - \beta}. \tag{1}$$

Next, we have the following lemma, which allows us to focus in our analysis on a small set of $6 = 2 \times 3$ binary signal structures, such that, if the outcome of majority voting is manipulable, one of these signal structures must be optimal (and any other optimal signal structure has weakly higher values of α and β).

Lemma 1. Suppose that signal structure (α', β') is optimal. Then the signal structure (α, β) is also optimal, where $\alpha = \max\{q \in \{q_\ell, q_h\} : q \leq \alpha'\}$ and $\beta = \max\{q \in \{0, 1 - q_h, 1 - q_\ell\} : q \leq \beta'\}$.

The intuition for Lemma 1 is that for any signal structure, slightly decreasing either α or β increases the probability of the A-favorable realization $\mathfrak{a}$, and that if $\alpha \notin \{q_\ell, q_h\}$ and $\beta \notin \{0, 1 - q_h, 1 - q_\ell\}$, then this slight decrease has no effect on the behavior of any voter conditional on her signals. Table 1 summarizes the key properties of these 6 optimal binary signal structures (which are derived from straightforward calculations). Note that the share of voters who vote for policy A is always larger in state θ_A than in state θ_B, and, thus, a signal structure is optimal if and only if it induces at least half of the voters to vote for policy A in state θ_B.

Remark 4.1. The tie-breaking rule has no significant impact on our results. In particular, if one assumes the opposite, namely, a B-favorable tie-breaking rule (i.e., each voter votes for B when indifferent, and B is chosen if supported by exactly half of the voters), then the only minor effect on our results will be that the optimal signal structures will be perturbed by a small $\varepsilon > 0$. outcome of majority voting, then there will be an optimal signal structure of the form $(\frac{1}{2} + \varepsilon, \beta + \varepsilon)$ or $(\alpha + \varepsilon, \beta + \varepsilon)$, with $\alpha \in \{q_\ell, q_h\}$ and $\beta \in \{0, 1 - q_h, 1 - q_\ell\}$, for a sufficiently small $\varepsilon > 0$.

[11] Representing signal structures as distributions over posteriors is standard in Bayesian persuasion [22]. Representing them as posteriors starting with a prior of $\frac{1}{2}$ is common in the literature on social learning (see, e.g.,[1, 4, 5]).

Table 1: Key Properties of the 6 Possibly Optimal Signal Structures of Lemma 1

(α,β)	$\mathrm{P}\left[s(\theta_B)=\mathfrak{b}\right]$	Bias	How voters vote	Share of B voters in θ_B
$(q_\ell,0)$	$2-\frac{1}{q_\ell}$	$\frac{1-q_\ell}{q_\ell}>0$	q_h: Vote A iff a AND $\mathfrak{a}$ q_ℓ: Follow designer's signal	$1-\frac{(1-q_\ell)(1-q_h+\lambda q_h)}{q_\ell}$ (decreases in λ)
$(q_h,0)$	$2-\frac{1}{q_h}$	$\frac{1-q_h}{q_h}>0$	Follow designer's signal	$\frac{2q_h-1}{q_h}$ (independent of λ)
$(q_\ell,1-q_\ell)$	q_ℓ	0	q_h: Follow original signal q_ℓ: Vote A iff a OR $\mathfrak{a}$	$q_h-\lambda\left(q_h-q_\ell^2\right)$ (decreases in λ)
$(q_\ell,1-q_h)$	$\frac{q_h(2q_\ell-1)}{q_h+q_\ell-1}$	$\frac{q_h-q_\ell}{q_\ell+q_h-1}>0$	q_h: Follow original signal q_ℓ: Follow designer's signal	$\lambda\frac{q_h(2q_\ell-1)}{q_h+q_\ell-1}+(1-\lambda)\,q_h$ (decreases in λ)
$(q_h,1-q_\ell)$	$\frac{q_\ell(2q_h-1)}{q_h+q_\ell-1}$	$-\frac{q_h-q_\ell}{q_\ell+q_h-1}<0$	Vote A iff a OR $\mathfrak{a}$	$(\lambda q_\ell+(1-\lambda)\,q_h)\,\frac{q_\ell(2q_h-1)}{q_h+q_\ell-1}$ (decreases in λ)
$(q_h,1-q_h)$	q_h	0	q_h: Vote A iff a OR $\mathfrak{a}$ q_ℓ: Follow designer's signal	$q_h^2+\lambda q_h\,(1-q_h)$ (increases in λ)

4.2 Partially Uninformed Populations ($q_\ell=\frac{1}{2}$)

In this section, we focus on the case in which the low-accuracy signal is uninformative (i.e., $q_\ell=\frac{1}{2}$). That is, we assume that a share λ of the population is uninformed, and the remaining voters each obtain a private signal with accuracy q_h. An interesting special case, discussed toward the end of this section, is that of homogeneous populations, in which all voters have the same signal accuracy (formally, $\lambda=0$).

Our first observation is that if the signal accuracy q_h is below some threshold, denoted by $q_{NI}(\lambda)$, then policy A is always chosen, even without any manipulation by the designer.[12]

Claim 1. Policy A is chosen in both states with an uninformative designer's signal structure iff

$$q_h\leq q_{NI}(\lambda)\equiv\frac{1}{2(1-\lambda)}.$$

It is straightforward to verify that $q_{NI}(\lambda)$ is increasing in λ, that $q_{NI}(25\%)=\frac{2}{3}$, and that $q_{NI}(50\%)=1$. This is illustrated by the solid (red) curve in Figure 4.2. In what follows we focus on the more interesting case in which q_h is above this threshold—namely, the case in which policy B is chosen in state θ_B unless the designer manipulates the outcome.

In order to state our result, we define $\overline{q}:[0,1)\Rightarrow[0.5,1]$ as follows:

$$\overline{q}(\lambda)=\frac{-\lambda+\sqrt{\lambda^2+2-2\lambda}}{2-2\lambda}.$$

It is straightforward to verify the following, illustrated by the dashed (blue) line in Figure 4.2:

Fact 1. $\overline{q}(\lambda)$ is decreasing in λ, $\overline{q}(0)=\frac{\sqrt{2}}{2}$, and $\overline{q}(25\%)=\frac{2}{3}=q_{NI}(25\%)$.

Our first result shows that the designer can manipulate the outcome of majority voting iff $q_h\leq\overline{q}(\lambda)$, and that manipulation (when possible) can be implemented by an unbiased signal structure. Moreover, for values of precision in the interval $\left(\frac{2}{3},\overline{q}(\lambda)\right)$ (values above the horizontal dotted (black) line in Figure 4.2), all optimal signal structures have non-positive biases.

[12] If one changes the tie-breaking rule to favor policy B, then under the same condition on λ, the designer can perfectly manipulate the outcome by sending the signal structure $(\frac{1}{2}+\varepsilon,1-q_h)$ for a sufficiently small $0<\varepsilon\ll1$. This signal structure is almost uninformative in the sense that with an arbitrarily high probability of $1-O(\varepsilon)$ the voters get the realization $\mathfrak{a}$, which only slightly increases the likelihood of state θ_A, and which is sufficient to make the uninformed voters strictly prefer policy A.

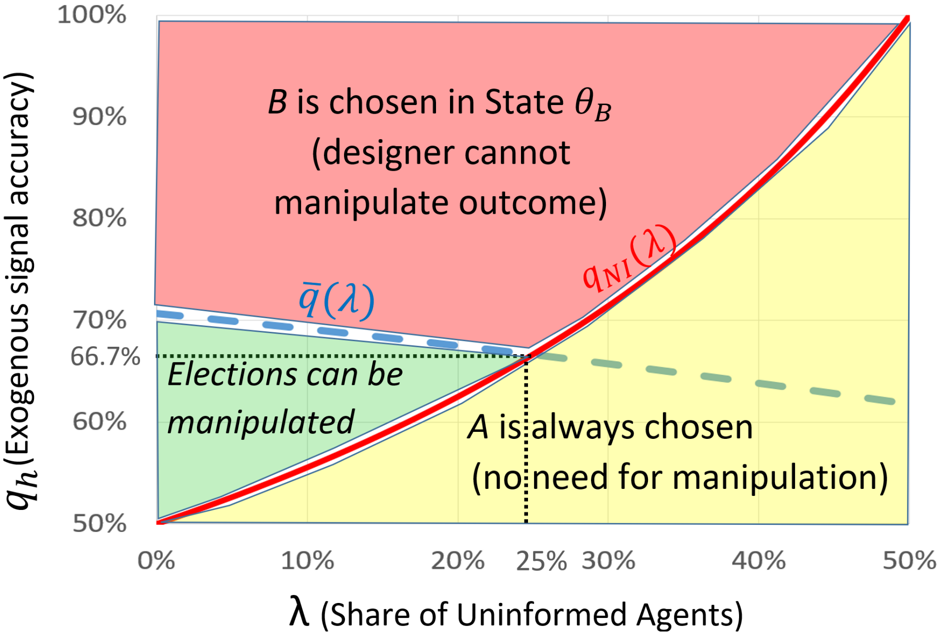

Figure 1: Illustration of Proposition 1 ($q_\ell = 0.5$)

Proposition 1. Suppose that $q_\ell = \frac{1}{2}$ and $q_h > q_{NI}(\lambda)$.

1. If $q_h > \bar{q}(\lambda)$ then the designer cannot manipulate the outcome of majority voting.
2. If $q_h \leq \bar{q}(\lambda)$ then the designer can manipulate the outcome of majority voting, and she can do so by using the unbiased signal structure $(q_h, 1 - q_h)$.
3. If $q_h \in (\frac{2}{3}, \bar{q}(\lambda)]$ then *all* optimal signal structures have non-positive biases.

An interesting feature of the optimal signal structures is that they have non-positive bias (bullets 2 and 3 of Proposition 1). Observe that this is the opposite direction of bias relative to the standard application of Bayesian persuasion [23], in which the bias is positive (i.e., in favor of A, the designer-optimal outcome). The intuition for the contrast between the standard application and our setup is as follows. A voter who is informed and obtains signal a is prepared to vote for A, whereas a voter who is informed and obtains signal b is prepared to vote for B. The designer would like to convince the latter to vote for A, while at the same time not causing the former to switch. The way to do this is to choose a signal structure for which realization $\mathfrak{a}$ is stronger than realization $\mathfrak{b}$, so that voters with realizations $(b, \mathfrak{a})$ switch to A but voters with realizations $(a, \mathfrak{b})$ do *not* switch to B. Of course, since $\mathfrak{a}$ is stronger than $\mathfrak{b}$, it must be the case that $\mathfrak{b}$ is realized with higher probability, or, in other words, that the signal structure has non-positive bias. We now prove Proposition 1. An interesting special case is a homogeneous population in which all voters have the same signal quality (i.e., $\lambda = 0$). Applying Proposition 1 yields the following characterization of vote manipulation in homogeneous populations.

Corollary 1 (Homogeneous populations)**.** Suppose that $\lambda = 0$.

1. If $q_h > \frac{\sqrt{2}}{2}$ then the designer cannot manipulate the outcome of majority voting.
2. If $q_h \leq \frac{\sqrt{2}}{2}$ then the designer can manipulate the outcome of majority voting, and she can do so by using the unbiased signal structure $(q_h, 1 - q_h)$.

3. If $q_h \in (\frac{2}{3}, \frac{\sqrt{2}}{2}]$ then *all* optimal signal structures have non-positive biases.

4.3 General Binary Signals $\left(\frac{1}{2} < q_\ell < q_h,\ \lambda \in [0,1]\right)$

In this section we return to the general binary model in which the low-accuracy signal is informative ($q_\ell > 1/2$). We prove two results: first, a characterization of conditions under which the designer can manipulate the outcome of majority voting; and second, an identification of conditions under which optimal signal structures have positive or negative biases.

In order to state our first result we define $\underline{\lambda} : (\frac{1}{2}, \frac{\sqrt{2}}{2})] \to [0,1]$ as follows: $\underline{\lambda}(q_h) = \frac{0.5 - q_h^2}{q_h(1-q_h)}$. We will show in the proof of Proposition 2 below that $\underline{\lambda}(q_h)$ is the highest share of λ for which the signal structure $(q_h, 1-q_h)$ induces a majority of the voters to vote for A in state θ_B for all values of $q_\ell \leq q_h$. Observe that:

Fact 2. The function $\underline{\lambda}(q_h)$ is decreasing, $\underline{\lambda}(\frac{2}{3}) = 25\%$, and $\underline{\lambda}(\frac{\sqrt{2}}{2}) = 0$.

Proposition 2.

1. If $\frac{\sqrt{2}}{2} < q_\ell$, then the designer cannot manipulate the outcome of majority voting.
2. If $q_\ell < \frac{\sqrt{2}}{2} < q_h$, then the designer can manipulate the outcome of majority voting iff λ is sufficiently high.
3. If $\frac{2}{3} < q_h \leq \frac{\sqrt{2}}{2}$, then the designer can manipulate the outcome of majority voting iff either (I) $\lambda \leq \underline{\lambda}(q_h)$, or (II) λ is sufficiently large.
4. If $q_h \leq \frac{2}{3}$, then the designer can manipulate the outcome of majority voting for any q_ℓ and λ.

Our results are illustrated in Figure 2. Part (3) of Proposition 2 implies an interesting non-monotonicity in λ: when q_h is in the interval $\left(\frac{2}{3}, \frac{\sqrt{2}}{2}\right)$, then the outcome of majority voting is manipulable if either the share of voters with low accuracy is sufficiently low (namely, $\lambda \leq \underline{\lambda}$) or if it is sufficiently high, but the outcome of majority voting cannot be manipulated if this share is intermediate. The intuition for this non-monotonicity is as follows. Our designer is constrained by the fact that she must choose the same signal structure for all voters. When λ is either very low or very high, the designer can tune this distribution to the large majority of voters by sending an unbiased signal with the "right" level of accuracy (i.e., a level that induces most voters to vote for A if either of their signals is in favor of A). By contrast, when λ is close to 50%, the designer has a trade-off, and must send a "versatile" signal that can handle both groups of voters (and is not "tailored" to either group), and this limits her ability to manipulate the outcome of majority voting.

Moreover, one can show that the ability to manipulate the outcome of majority voting is also non-monotonic in q_ℓ. Specifically, for each q_h, the interval of λ-s for which the outcome of majority voting cannot be manipulated is non-monotone in q_ℓ: the interval is small (and it might even disappear) when q_ℓ is either small (close to 50%) or large (close to q_h), while it is largest when q_ℓ is intermediate. The intuition for the non-monotonicity in q_ℓ is as follows. When q_ℓ is low, it is easy to manipulate the low-accuracy voters. The designer can send a positively biased signal structure that with high probability sends a "weak pro-A" realization, which manipulates most low-accuracy voters (and a sufficient number of high-accuracy voters). When q_ℓ is close to q_h, the designer can exploit the relative homogeneity of the population by sending the negatively biased signal structure $(q_h, 1-q_\ell)$, which induces both types of voters to vote for A if either of their signals is "pro-A." By contrast, when

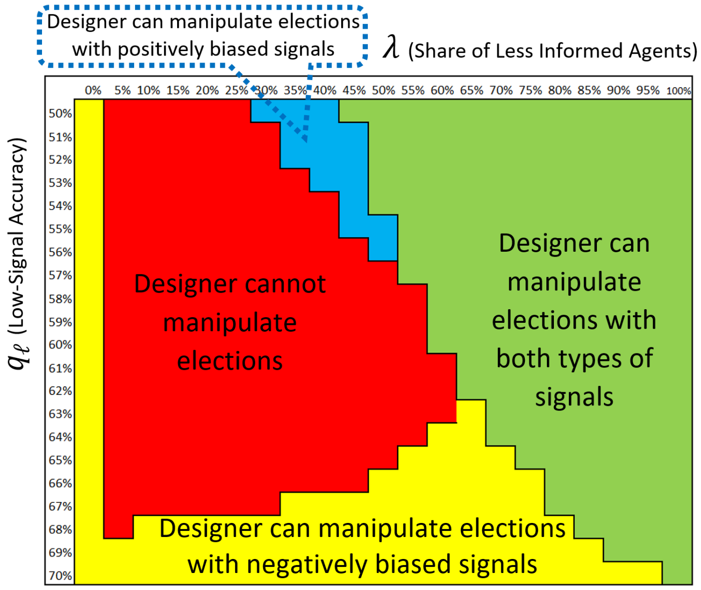

Figure 2: Outcome Manipulation for $q_h = 70\%$ and Various Values of q_ℓ (values between 50% and 70% in steps of 1%) and λ (values between 0 and 100% in steps of 5%)

q_ℓ has an intermediate value the designer has neither of these two benefits, and thus her ability to manipulate the outcome of majority voting is limited.

Recall that in the special case of $q_\ell = 0.5$ (partially informed voters), manipulation, when possible, could always be implemented by an unbiased signal structure (and in some cases only by signal structures with non-positive biases). In what follows, we show that this is no longer the case in the general setup (with $q_\ell > 0.5$). Specifically, we characterize both setups in which manipulation can only be implemented by positively biased signal structures (in which the signal $\mathfrak{a}$ is sent more often than $\mathfrak{b}$), and setups in which manipulation can only be implemented by negatively biased signal structures. Specifically, Proposition 3 below shows that the direction of the bias is uniquely determined (for an interior interval of λ-s) in the following setups:

1. All optimal structures have a positive bias if $q_h > \frac{2}{3}$ and q_ℓ is sufficiently low.

2. All optimal structures have a negative bias if $q_h \in (\frac{2}{3}, \frac{\sqrt{2}}{2})$ and q_ℓ is sufficiently high.

Observe that in the second case, the manipulation is implemented by signal structures with the opposite direction of bias relative to the standard applications of Bayesian persuasion (see, e.g., 22). The intuition for this is that the optimal negatively biased signal structure $(q_h, 1 - q_\ell)$ exploits the (modest) heterogeneity of the population: the relatively infrequent signal $\mathfrak{a}$ is sufficiently informative to induce both types of voters to vote for A regardless of their exogenous signal, while the more frequently used signal $\mathfrak{b}$ is sufficiently weakly informative, so that it does not change the vote of voters who get exogenous signal a.

Proposition 3.

1. For any $q_h > \frac{2}{3}$, there is a $\overline{q} \in (0.5, q_h)$ such that, if $q_\ell \in (0.5, \overline{q})$, then there is an interior open interval of λ-s for which all optimal signal structures have positive biases.
2. For any $q_h \in \left(\frac{2}{3}, \frac{\sqrt{2}}{2}\right)$ there is a $\underline{q} \in (0.5, q_h)$, so that, if $q_\ell \in (\underline{q}, q_h)$, then there is an interior open interval of λ-s where all optimal signal structures have negative biases.

5 Extensions

5.1 Continuous Signal Accuracy Levels

In this section we extend our model to a general distribution over signal accuracy levels. Specifically, we consider here a variant of our model in which the exogenous signal's accuracy is continuous, and is distributed according to density f with a support $\text{supp}(f)$ in $[0.5, 1]$. We note that the closely related extension of the model to a setup with a finite but large number of signal accuracy levels is analogous (and therefore omitted for brevity).

Our result here is an extension of Proposition 2 to this general setup, showing that the outcome of majority voting is manipulable if all voters have accuracy levels of at most $\frac{2}{3}$, and that it is non-manipulable if all voters have a signal accuracy at least $\frac{\sqrt{2}}{2}$. Formally,

Claim 2.

1. If $\text{supp}(f) \subseteq \left[\frac{\sqrt{2}}{2}, 1\right]$, the designer cannot manipulate the outcome of majority voting.
2. If $\text{supp}(f) \subseteq \left[\frac{1}{2}, \frac{2}{3}\right]$, the designer can manipulate the outcome of majority voting.

5.2 Targeted Information

In our main analysis, we required the designer to design one signal structure for all voters, and we showed that, even under this restriction, the designer can often manipulate the outcome of majority voting. In this section we show that, if the designer can differentiate between voters based on their signal accuracy levels, then the designer's ability to manipulate is even greater.

Suppose that the designer can distinguish between the two types of voters, q_ℓ and q_h. Furthermore, suppose that she can commit to two different signal structures, s_h and s_ℓ, where the former targets q_h voters and the latter targets q_ℓ voters. Observe first that Proposition 2 applies: if $q_\ell < q_h \leq 2/3$ then the $(q_h, 0)$ signal structure is optimal regardless of the type of voter, and so the designer can manipulate the outcome of majority voting, whereas if $q_h > q_\ell > \sqrt{2}/2$ then the designer cannot manipulate the outcome of majority voting. The proof is the same as that of the proposition.

However, targeted signal structures differ in the intermediate case. Note that for either $i \in \{\ell, h\}$, the signal structure that leads the maximal fraction of q_i voters to vote for A is either $(q_i, 0)$ or $(q_i, 1 - q_i)$, by (the proof of) Lemma 1. Both signal structures lead a majority of voters for A in state θ_A. In state θ_B, however, $(q_i, 0)$ leads to $\mathrm{P}\left[A|\theta_B\right] = \frac{1-q_i}{q_i}$, whereas $(q_i, 1-q_i)$ leads to $\mathrm{P}\left[A|\theta_B\right] = 1-q_i^2$ (see Table 1). For $q_i \in [1/2, \frac{\sqrt{5}-1}{2})$, the fraction $\frac{1-q_i}{q_i}$ is larger, whereas for $q_i \in (\frac{\sqrt{5}-1}{2}, 1)$, the fraction $1 - q_i^2$ is larger. Overall, the designer can manipulate the outcome of majority voting if and only if

$$\lambda \cdot \max\left\{\frac{1-q_\ell}{q_\ell}, 1 - q_\ell^2\right\} + (1-\lambda) \cdot \max\left\{\frac{1-q_h}{q_h}, 1 - q_h^2\right\} \geq \frac{1}{2}.$$

Observe that, in contrast to Proposition 2, there is no non-monotonicity here: the left-hand side of the inequality decreases with q_ℓ and with λ.

This last necessary and sufficient condition on the designer's ability to manipulate the outcome of majority voting extends in a straightforward manner to the variant with a continuous distribution of signal accuracy levels from Section 5.1 above. In this case, the designer can manipulate the outcome of majority voting if and only if

$$\int_{1/2}^{1} \max\left\{\frac{1-q}{q}, 1-q^2\right\} \cdot f(q)dq \geq \frac{1}{2}.$$

5.3 Strongly Targeted Information

What if the designer has even more finely grained information about voters, such that she knows not only each voter's signal accuracy but also its realization? In this case, the designer can provide a strongly targeted signal structure, where voters with different levels of signal accuracy *and different realizations* get signals from different structures.

For a given signal accuracy q, a voter with realization a has interim belief q about the probability that the state is A, whereas a voter with realization b has interim belief $1-q$ about this probability. If the designer knows these realizations, she faces a standard Bayesian persuasion problem relative to each one of these groups of voters. In order to maximize the probability that such voters vote for A, the designer should supply the former group of voters with no additional signal, and the latter group of voters with an additional signal structure $(q, 0)$. This latter signal structure is the optimal signal structure from the standard Bayesian persuasion setting [23], supplied to a voter who has belief $1-q$ and is willing to vote for A once her belief is above $1/2$. Given these signals, the former voter will always vote for A, whereas the latter will vote for A with probability $\frac{1-q}{q}$ in state θ_B. In state θ_B, the total share of voters with accuracy q who vote for A is thus $(1-q)\cdot 1 + q \cdot \frac{1-q}{q} = 2(1-q)$.

For a homogeneous population in which all voters have accuracy q, the designer can manipulate the outcome of majority voting iff $2(1-q) \geq \frac{1}{2} \Leftrightarrow q \leq \frac{3}{4}$. When there are two levels of signal accuracy, q_ℓ and q_h, the designer can then manipulate the outcome of majority voting if and only if $2\lambda \cdot (1-q_\ell) + 2(1-\lambda)\cdot(1-q_h) \geq \frac{1}{2}$.

When there is a continuous distribution of signal accuracy levels[13] the designer can manipulate the outcome of majority voting if and only if $2\int_{1/2}^{1}(1-q)f(q)dq \geq \frac{1}{2}$.

5.4 Social Media vs. Traditional Media

Our analysis and results lead to a straightforward comparison of the effects of social media (through *private* persuasion) and traditional media (through *public* persuasion). For the latter, suppose that instead of each voter obtaining a conditionally independent realization of the designer's signal structure, all voters obtain the same realization. Is such a public signal structure better or worse for the designer?

Observe that with a public signal structure, the designer can never manipulate the outcome of majority voting with probability one. However, she can always at least slightly increase the probability that voters vote for A (with a positively biased signal structure as in [23]). Thus, the answer to whether this is better than private persuasion depends on whether or not the designer can manipulate the outcome of majority voting in the latter case. If she can, then private persuasion by social media is better. If she cannot, then public persuasion by traditional media is better.

[13] As in Section 5.1 above.

References

[1] Daron Acemoglu, Munther A Dahleh, Ilan Lobel, and Asuman Ozdaglar. Bayesian learning in social networks. *Review of Economic Studies*, 78(4):1201–1236, 2011.

[2] Ricardo Alonso and Odilon Câmara. Persuading voters. *American Economic Review*, 106(11):3590–3605, 2016.

[3] Itai Arieli and Yakov Babichenko. Private bayesian persuasion. *Journal of Economic Theory*, 182:185–217, 2019.

[4] Itai Arieli, Yakov Babichenko, and Rann Smorodinsky. Identifiable information structures. *Games and Economic Behavior*, 120:16–27, 2020.

[5] Itai Arieli, Ronen Gradwohl, and Rann Smorodinsky. Herd design. *American Economic Review: Insights*, forthcoming.

[6] David Austen-Smith and Jeffrey S Banks. Information aggregation, rationality, and the Condorcet jury theorem. *American Political Science Review*, 90(1):34–45, 1996.

[7] Arjada Bardhi and Yingni Guo. Modes of persuasion toward unanimous consent. *Theoretical Economics*, 13(3):1111–1149, 2018.

[8] Ron Berman, Hangcheng Zhao, and Yi Zhu. When (not) to persuade consumers: Persuasive and demarketing information designs. *mimeo*, 2021.

[9] Jimmy Chan, Seher Gupta, Fei Li, and Yun Wang. Pivotal persuasion. *Journal of Economic Theory*, 180:178–202, 2019.

[10] Nicolas De Condorcet. *Essai sur l'Application de l'Analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix.* Paris. Reprinted by Cambridge University Press, 2014, 1785.

[11] Philipp Denter, Martin Dumav, and Boris Ginzburg. Social connectivity, media bias, and correlation neglect. *The Economic Journal*, 131(637):2033–2057, 2021.

[12] Anthony Downs. *An Economic Theory of Democracy.* Harper & Row, New York, 1957.

[13] Facebook. Facebook News Feed: An introduction for content creators. https://www.facebook.com/business/learn/lessons/facebook-news-feed-creators, 2022. [Online; accessed 12-June-2022].

[14] Timothy Feddersen and Wolfgang Pesendorfer. Voting behavior and information aggregation in elections with private information. *Econometrica*, pages 1029–1058, 1997.

[15] Timothy Feddersen and Wolfgang Pesendorfer. Convicting the innocent: The inferiority of unanimous jury verdicts under strategic voting. *American Political Science Review*, 92(1):23–35, 1998.

[16] Matthew Gentzkow, Jesse M Shapiro, and Daniel F Stone. Media bias in the marketplace: Theory. In *Handbook of Media Economics*, volume 1, pages 623–645. Elsevier, Amsterdam, 2015.

[17] A Arda Gitmez and Pooya Molavi. Polarization and media bias. *arXiv preprint arXiv:2203.12698*, 2022.

[18] Ronen Gradwohl, Yuval Heller, and Arye L Hillman. Social media and democracy. *Available at SSRN*, 2022. URL https://ssrn.com/abstract=4149316.

[19] Carl Heese and Stephan Lauermann. Persuasion and information aggregation in elections. Technical report, Working Paper, 2021.

[20] Arye L Hillman. Expressive behavior in economics and politics. *European Journal of Political Economy*, 26(4):403–418, 2010.

[21] Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. Algorithmic amplification of politics on twitter. *Proceedings of the National Academy of Sciences*, 119(1):e2025334119, 2022.

[22] Emir Kamenica. Bayesian persuasion and information design. *Annual Review of Economics*, 11:249–272, 2019.

[23] Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

[24] Toygar Kerman, P Jean-Jacques Herings, and Dominik Karos. Persuading strategic voters. *GSBE Research Memorandum 20/04, Maastricht University*, 2020.

[25] Anton Kolotilin. Optimal information disclosure: A linear programming approach. *Theoretical Economics*, 13(2):607–635, 2018.

[26] Anton Kolotilin, Tymofiy Mylovanov, Andriy Zapechelnyuk, and Ming Li. Persuasion of a privately informed receiver. *Econometrica*, 85(6):1949–1964, 2017.

[27] Ro'ee Levy. Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, 111(3):831–70, 2021.

[28] Xiao Lin and Ce Liu. Credible persuasion. *arXiv preprint arXiv:2205.03495*, 2022.

[29] Vincent Pons and Clémence Tricaud. Expressive voting and its cost: Evidence from runoffs with two or three candidates. *Econometrica*, 86(5):1621–1649, 2018.

[30] Riccardo Puglisi and James M Snyder Jr. Empirical studies of media bias. In *Handbook of Media Economics*, volume 1, pages 647–667. Elsevier, Amsterdam, 2015.

[31] Jörg L Spenkuch. Expressive vs. strategic voters: An empirical assessment. *Journal of Public Economics*, 165:73–81, 2018.

[32] David Strömberg. Media and politics. *Annual Review of Economics*, 7(1):173–205, 2015.

[33] Junze Sun, Arthur Schram, and Randolph Sloof. A theory on media bias and elections. Technical report, Working Paper, 2019.

[34] Ina Taneva. Information design. *American Economic Journal: Microeconomics*, 11(4): 151–85, 2019.

[35] Twitter. Notices on Twitter and what they mean. https://help.twitter.com/en/rules-and-policies/notices-on-twitter, 2022. [Online; accessed 12-June-2022].

[36] Twitter. Help with locked or limited account. https://help.twitter.com/en/managing-your-account/locked-and-limited-accounts, 2022. [Online; accessed 12-June-2022].

[37] Yun Wang. Bayesian persuasion with multiple receivers. *Available at SSRN 2625399*, 2013.

Ronen Gradwohl
Ariel University
Israel
Email: roneng@ariel.ac.il

Yuval Heller
Bar-Ilan University
Israel
Email: yuval.heller@biu.ac.il

Arye Hillman
Bar-Ilan University
Israel
Email: arye.hillman@biu.ac.il