Social Media and Democracy

Gradwohl, Ronen and Heller, Yuval and Hillman, Arye

Ariel University, Bar-Ilan University

30 June 2022

Social Media and Democracy

Draft: Comments are welcome

Ronen Gradwohl[*]   Yuval Heller[†]   Arye Hillman[‡]

June 30, 2022

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 through social media to manipulation through traditional media.

Keywords: Bayesian persuasion; Political agenda; Information manipulation; Condorcet Jury Theorem; Biased signals. **JEL codes:** D72, D82, P16

---

[*]Department of Economics and Business Administration, Ariel University. roneng@ariel.ac.il.

[†]Department of Economics, Bar-Ilan University & University of California San Diego. yuval.heller@biu.ac.il.

[‡]Department of Economics, Bar-Ilan University. arye.hillman@biu.ac.il.

1   Introduction

In early theories of voting (Hotelling, 1929; Downs, 1957), voters are fully informed and vote based on their given preferences. In subsequent expositions, however, voters do not know which policies or competing political candidates merit their support. They are then susceptible to influence by political advertising (Hillman and Ursprung, 1988; Grossman and Helpman, 2001; Prat, 2002) and by competing endorsements of policies and candidates (Grossman and Helpman, 1999; Lichter, 2017) through mass media.

Voters can also be influenced by social media. Rather than competition between media institutions, network externalities create natural monopolies for users of social media. In addition, rather than being exposed to the same, public information, users of social media face private persuasion through algorithms that select and order the information available to them. Facebook's News Feed algorithm, for example, determines personally, for each user, the posts are visible and the order in which they appear (Facebook, 2022). Users of social media can also be influenced by censorship or by delays in access to information. Twitter, for example, places warning labels on tweets and sometimes blocks information dissemination by locking accounts (Twitter, 2022b,a).

In this paper, we study the ability of a monopoly social media platform with a political agenda to influence users' views through such selective transmission of information. In our model voters may have outside information and preferences that are independent of the social media platform's agenda; however, voters are not perfectly informed, and so do not know with certainty which candidate or policy to support. To what extent can the social media platform exploit this incomplete information?

Our beginning is Condorcet's jury theorem (1785), which shows the merit of majority voting. In the setting of the theorem, voters face a choice between two alternatives, A and B. One alternative is objectively better for all voters, but voters do not know a priori which is the preferred alternative. Each voter has limited information, expressed in probabilities, indicating which alternative is better. The jury theorem states that the better alternative is more likely to be chosen when the decision is made by majority voting than when a single individual makes the decision. As the number of voters increases, the likelihood of the decision under majority

voting being correct approaches certainty. The theorem therefore vindicates democracy.

We ask, in an extension of the theorem: Does a group of voters who obtain information through a social media platform that has a political agenda also necessarily successfully make the correct decision by majority voting? Or, can information manipulation by social media undermine the merit of majority voting?

We undertake our study and derive our results within the Bayesian persuasion framework (see Kamenica, 2019, for a survey). The media platform commits to how it will manipulate information, a commitment that in practice is implemented through the algorithms used by media platforms to select and order the information that individuals obtain. Individuals are aware of the manipulation and make correct inferences using Bayes' rule based on the information available. We show that information manipulation can lead majority voting to result in the choice of the alternative sought by the social media platform rather than the alternative that is preferable for voters. Contrary to Condorcet's jury theorem, the consequence of information manipulation can be that a single individual is more likely to choose the socially preferred alternative than a group that decides by majority voting, even when individuals are aware of and account for the platform's information manipulation. Democracy is thus undermined.

In the remaining parts of the introduction we demonstrate our main results using examples, and then briefly describe our general model and results. Section 2 discusses the related literature. In Section 3 we formally present our model, and then, in Section 4, we describe our results. Various extensions are presented in Section 5, followed by a conclusion in Section 6.

1.1 Illustrative Example

Before we proceed to the formal model we provide an illustrative example. We begin with Condorcet's jury theorem, proceed to the basic logic of Bayesian persuasion, and then show how social-media persuasion can overturn the conclusion of the theorem.

Condorcet's Jury Theorem   There are two alternatives, A and B, one of which is "better" or socially preferred. Voters are a priori unsure about which of A and B is better, but they do have some limited prior information supporting one of the alternatives. Suppose that, independently for each voter, this prior information has accuracy 55%—namely, there is a 55% chance of it

being correct. Specifically, if a voter has information indicating that one of the alternatives (say, A) is better, then this voter believes there is a 55% chance that A is, in fact, better (and a 45% chance that B is better).

Thus, if each voter chooses the alternative that is more-likely to be better based on this prior information, then she has a 55% chance of choosing the socially preferred alternative. Rather than a single voter deciding between A and B, suppose that the decision is made by majority voting. Condorcet's jury theorem states that, with sufficiently many voters, if each voter votes for the alternative that is more-likely to be better based on her own information, then the correct alternative is chosen with near certainty (due to the Law of Large Numbers).

Bayesian Persuasion   We now introduce an information designer (the social media platform) that can send voters additional information in the form of a state-dependent message. The designer is interested in ensuring that alternative A is chosen, regardless of which alternative is socially better. Suppose the designer can send one of two messages—either "A is better" or "B is better". The designer has to choose a rule that determines which message to send as a function of the true socially-better alternative. Importantly, we assume (à la Kamenica and Gentzkow, 2011) that the voters know the designer's rule. The simple rule of always sending "A is better" is useless because the voters are aware that this signal is non-informative. A designer who wants to influence voters to vote for A would instead have to be more subtle.

Consider the following messaging scheme: if A is better, the designer always sends the message "A is better"; if B is better, the designer sends the message "B is better" with probability 30%, and the "incorrect" message "A is better" with the remaining 70% probability. What can voters infer from the designer's message? Observe first that if a voter obtains message "B is better" from the designer, then this voter can be certain that B is, in fact, better.

By contrast, if a voter obtains message "A is better", then the voter cannot be certain about which alternative is better, since this message is sent both when A is actually better but also sometimes when B is better. The exact inference that a voter can make upon obtaining message "A is better" depends on this voter's prior information—whether it supports A or B—and is calculated by Bayes' rule. A simple calculation shows that a voter whose prior information supported A (i.e., she has a prior belief that A is better with probability 55%), updates her

posterior belief that A is better to 64%:

$$\mathrm{P}\left[A|\text{"A is better"}\right] = \frac{\mathrm{P}\left[A \cap \text{"A is better"}\right]}{\mathrm{P}\left[\text{"A is better"}\right]} = \frac{100\% \times 55\%}{100\% \times 55\% + 70\% \times 45\%} \approx 64\%.$$

Similarly, a voter whose prior supported B updates her posterior belief that A is better to 54%:

$$\mathrm{P}\left[A|\text{"A is better"}\right] = \frac{\mathrm{P}\left[A \cap \text{"A is better"}\right]}{\mathrm{P}\left[\text{"A is better"}\right]} = \frac{100\% \times 45\%}{100\% \times 45\% + 70\% \times 55\%} \approx 54\%.$$

Persuasion Overturns Condorcet   We now consider the effect of the designer's message on voting outcomes, and show that the consequence is to overturn the necessity of the conclusion of Condorcet's jury theorem. Consider first a single voter who makes the decision based on which alternative is more likely to be better. Recall from the previous section that if this voter obtained message "B is better" from the designer, then the voter is certain that B is better, and will choose alternative B. If the voter obtained message "A is better", however, then her belief about the likelihood of A being better is either 64% or 54%, depending on her initial information. Importantly, however, *in both cases* A is more likely to be better, and so in both cases the voter will choose A upon obtaining this message.

How likely is this voter to choose the alternative that is actually better? Recall that if A is actually better, then the designer always sends message "A is better", and the voter makes the correct choice and votes for A. However, if B is actually better, then the designer sends message "B is better" only with probability 30%. If this happens, then the voter is certain that B is better and makes the correct choice, B. However, with the remaining 70% probability, the designer sends message "A is better", after which the voter chooses the incorrect alternative, A. Thus, if B is actually better, the voter makes the correct choice with probability 30% and the incorrect choice with probability 70%.

Rather than a single voter deciding between A and B, suppose that the decision is made by majority voting in a large population. If A is actually better then, as above, every voter obtains message "A is better" and votes for A. If B is actually better, however, then on average 30% of voters obtain message "B is better" and vote for B, whereas on average 70% of voters obtain message "A is better" and vote for A. In this case, by the Law of Large Numbers, if there are sufficiently many voters then with very high probability a majority votes for A, although

A is the *incorrect* alternative. Thus, if B is better, majority voting never leads to the correct alternative being chosen. Note that this is worse than a single voter making the decision, as then there was a 30% chance that B is chosen when it is better.

In this example we have illustrated how additional information provided by a self-interested designer can overturn the conclusion of Condorcet's jury theorem. When there is no designer, a large group deciding by majority rule is more likely to make the correct decision than a single voter deciding alone. With additional information from a designer, however, the conclusion is reversed: A single voter makes the correct decision with probability 30% if B is better, but majority voting never leads to the correct decision with when B is better (and in both cases the correct decision is made when *A* is better).

Negatively Biased Signals and the Limits of Persuasion   We now expand on the example above in order to answer two questions. First, what are the properties of the messaging scheme used by the designer to overturn an election? In particular, is it necessarily biased towards the outcome desired by the designer? And second, can the designer always overturn the election, or are there limits to persuasion?

We begin with the first question. Observe that in the messaging scheme described above the designer sends the message "A is better" more often than the message "B is better". This scheme is thus *biased* towards A. In the example above, this particular scheme is not unique, and there are other messaging schemes—some of which are biased towards B rather than A—that also allow the designer to overturn the election. We now describe such a scheme.

Suppose that, if A is better, the designer sends the message "A is better" with probability 53%, and the message "B is better" with probability 47%. If B is better, the designer sends the message "B is better" with probability 57% and the message "A is better" with probability 43%. Notice that the designer is more likely to lie in favor of B—to send message "B is better" when A is actually better (47%)—than to lie in favor of A—to send message "A is better" when B is actually better (43%). This messaging scheme is thus biased towards B.

What is the effect of this scheme? Consider first a voter whose prior information supported A, so that this voter initially believes A to be better with probability 55%. If this voter now obtains message "B is better" from the designer, the voter updates the probability of A being

correct by conditioning on the obtained message. The new belief is:

$$\text{P}\left[A|\text{"B is better"}\right] = \frac{\text{P}\left[A \cap \text{"B is better"}\right]}{\text{P}\left[\text{"B is better"}\right]} = \frac{47\% \times 55\%}{47\% \times 55\% + 57\% \times 45\%} \approx 50.2\%.$$

Thus, a voter who initially believed A to be better with probability 55%, updates this belief to 50.2% upon obtaining message "B is better" from the designer. Critically, this updated belief is above 50%, and so the voter still believes that A is more likely to be the better alternative. Similarly, if this same voter obtained message "A is better" instead of "B is better", the updated belief would be even higher (around 60%).

Consider now a voter whose prior information supported B, so that this voter initially believes B to be better with probability 55% (and A to be better with probability 45%). If this voter now obtains message "A is better" from the designer, the voter updates the probability of A being correct by conditioning on the obtained message. The new belief is:

$$\text{P}\left[A|\text{"A is better"}\right] = \frac{\text{P}\left[A \cap \text{"A is better"}\right]}{\text{P}\left[\text{"A is better"}\right]} = \frac{53\% \times 45\%}{53\% \times 45\% + 43\% \times 55\%} \approx 50.2\%.$$

Thus, a voter who initially believed A to be better with probability 45%, updates this belief to 50.2% upon obtaining message "A is better" from the designer. Critically, this updated belief is also above 50%, and so the voter now believes A is more likely to be the better alternative.

In fact, the only voters who vote for B are those whose prior information supported B, and who then additionally obtained message "B is better" from the designer. All other voters end up with a belief indicating that A is more likely to be better, and who then vote for A.

The final step is to determine the fraction of voters for A. If A is the better alternative, then only 45% of voters' initial information supports B, and of those, only 47% obtain message "B is better". Thus, on average only $45\% \times 47\% = 21\%$ of voters vote for B, and the rest for A. With high probability, then, the majority vote for A. Similarly, if B is the better alternative, then 55% of voters' initial information supports B, and of those, 57% obtain message "B is better". Thus, on average $55\% \times 57\% = 31\%$ of voters vote for B. Again, this implies that with high probability, the majority vote for A. Thus, regardless of which of A or B is actually better, the designer's messaging scheme leads to alternative A always being elected.

The example just described shows that the designer can overturn the election with a messaging scheme biased towards B—that is, towards the *opposite* alternative of the one preferred by the designer. In the previous section we showed that the election can also be overturned with a scheme biased towards A. So which will the designer use?

In Section 4 we characterize the biases of the schemes that can be used to overturn the election, and how they are related to the accuracy of voters' prior information. In particular, in Corollary 1 we show that, if this accuracy is between $1/2$ and $2/3$, then the designer can use both classes of schemes to overturn the election. However, we also show that if the accuracy is between $2/3$ and $\sqrt{2}/2$, then *only* schemes that are biased towards B can overturn the election. That is, in this case, the only way for the designer to overturn the election is to design a messaging scheme that is more likely to give incorrect information supporting B.

Finally, Corollary 1 also points to the limits of persuasion. The corollary states that if the voters' prior information has accuracy *greater than* $\sqrt{2}/2$ then there is no messaging scheme that allows the designer to overturn the election. The intuition is the following. In order to overturn the election, the designer needs to convince enough voters whose initial information supported B to switch to A. But the more accurate the voters' prior information, the more difficult it is to convince them to switch. At the extreme, if voters' prior information is *perfectly* accurate, and so they know with certainty which of the alternatives is actually better, then clearly no additional information from the designer could ever shift any of their beliefs. In Corollary 1 we show that even with less extreme accuracies, if the accuracies are above $\sqrt{2}/2$ then there is no messaging scheme that allows the designer to convince a majority of the voters to vote for A.

1.2 A General Model

Our general model, of which the above example is a special case, shows how in more encompassing circumstances the designer is able to sway the vote towards his desired outcome. As in the example, in some cases, the preferred outcome of the designer is chosen with probability one. In other cases, the designer is unable to exert any influence at all. The ability of the designer to influence outcomes depends both on the extent to which users are exogenously informed and on the heterogeneity in the accuracy levels of users' exogenous information.

We begin with a setting in which some voters are uninformed, whereas others are exoge-

nously informed with the same accuracy. We show that, as long as this accuracy is not too high, the social media platform can completely sway the outcome of the vote. Interestingly, we also show that, in this case, the designer can sway the outcome of the vote with an unbiased signal. We also show that, in some cases, the designer can *only* sway the outcome with a signal biased towards B—that is, one that sends message "B is better" more often than the message "A is better". Finally, we show that when users' exogenous signal accuracy is high, the designer can no longer affect the outcome.

Next, we extend the model to a scenario in which there are two kinds of voters corresponding to two levels of exogenous-signal accuracy. Here we show that if both accuracies are sufficiently low, the designer can determine the outcome of the vote with probability one; if both accuracies are sufficiently high, the designer cannot exert any influence; and if accuracies are in an intermediate range, then whether or not the designer has influence depends on both the specific accuracies of information and the fractions of users of each kind. We show that, in this last case, there is a nonmonotonicity wherein the designer can determine the outcome if most users are of one kind and if most users are of the other kind, but not when there is a substantial fraction of each kind. There is also a nonmonotonicity wherein the designer can determine the outcome of the vote when the two accuracies are sufficiently close or sufficiently far, but not in between. Finally, we provide conditions under which the designer can control the outcome of the vote only with a signal that is biased towards A and only with a signal that is biased towards B.

2   Related Literature

A literature on 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 holder of information can manipulate others by selective disclosure of information (Kamenica and Gentzkow, 2011). In the terminology of this literature, we study a setting of private Bayesian persuasion with multiple, informed receivers. We also restrict our designer to conditionally independent signals.

A result that Condorcet's jury theorem can be overturned has been derived based on costly

acquisition of information (Mukhopadhaya, 2003; Koriyama and Szentes, 2009). We retain the assumption of the theorem that individuals incur no cost in acquiring information. On the contrary, we describe information as imposed upon individuals through the social media platform without individuals actually seeking information.[1]

A substantial literature conjoins an information designer with strategic voting, with conclusions relying critically on the interplay between the two (see next paragraph for details). In contrast, we view voters as personally optimally voting for the outcome that they believe is better or as voting "sincerely". This allows us to focus on and cleanly characterize the direct effects of persuasion on voting outcomes. Strategic voting, in contrast, requires the assumption that voters believe that their vote can be pivotal (or decisive) in determining the outcome of a voting.[2] A concession to realism is that voters in usual two-party elections for political representatives know that the likelihood of their one vote being decisive is essentially zero. We assume that voters vote sincerely based on their information and do not regard themselves as possibly decisive or pivotal.[3] We suppose that voters enjoy an expressive benefit to voting for the correct alternative, and that all voters vote with no expectation that their vote will be pivotal in determining an election outcome.

Austen-Smith and Banks (1996) 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. Feddersen and Pesendorfer (1998) show that under general voting rules, such strategic voting may lead to a failure of information aggregation, but Feddersen and Pesendorfer (1997) 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. Heese and Lauermann (2021) start with the general strategic voting framework of Feddersen and Pesendorfer (1997), and add an information designer. They show that, under strategic voting, the designer can provide

---

[1]For scenarios with costly voting, see Sun et al. (2019), who study the effects of public persuasion and cheap talk on voter turnout and Chan et al. (2019), who study the impact of costly voting on uninformed voters.

[2]See, for example, Pons and Tricaud (2018) and Spenkuch (2018) 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.

[3]A paradox of voting is present (Downs, 1957) 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 (Hillman, 2010). The act of voting is used to express an identity or sense of belonging. In this context, voting is rational.

additional information in a way that leads to any preferred outcome with high probability. Taneva (2019) studies a symmetric two-voter game; her main result is that, in this setting, it is never optimal for the designer to commit to a conditionally independent signal. More generally, Arieli and Babichenko (2019) study a general framework for private persuasion, and characterize optimal solutions under submodularity assumptions on designer's utility. In our setting the utility function is neither supermodular nor submodular, so these results do not apply. Finally, Kerman et al. (2020) focus on persuading uninformed voters through correlated signals that induce sincere rather than strategic voting in equilibrium.

Other literature has studied public persuasion, where the designer commits to a signal 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' attributes and communications to design selective information directed at an individual. The literature includes Alonso and Câmara (2016), who study public persuasion of heterogeneous voters. Wang (2013) compares public signals to private, conditionally independent signals, and shows that the former is more informative (and hence worse for the designer) than the latter. Gitmez and Molavi (2022) also study public persuasion, with a focus on the interplay between the bias of the signal and voters' polarization. While less related to our work in terms of focus and results, the model of Gitmez and Molavi (2022) is similar to ours in that there is a continuum of voters and each votes sincerely.

A sizable literature has studied bias in traditional and social media, both theoretically and empirically (see Gentzkow et al., 2015; Puglisi and Snyder Jr, 2015; Strömberg, 2015, 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 manipulator's optimal bias is in the opposite direction of the intended outcome is novel.[4]

Last, in other related literature, Denter et al. (2021) study media bias when those it targets also have exogenous information, and share this information via a social network. They study

---

[4]We are aware of only one Bayesian persuasion model with negatively biased signals, Berman et al. (2021), in which such signals are used by duopolistic sellers in order to soften price competition.

the effects of the network's connectivity when voters suffer from correlation neglect on the optimal bias of media. Sun et al. (2019) study the effects of public persuasion and cheap talk on turnout in a model with costly voting. Relatedly, Levy (2021) perform a field experiment to study the effects of social media (specifically, Facebook) on user's political leanings.

3   Model

There are two policies (or candidates) denoted 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 signal accuracies. Specifically, we assume that a λ-fraction (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.[5]

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 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 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 (Kamenica and Gentzkow, 2011), we assume that the designer commits to s prior to learning the true state of the world. In addition, and without loss of generality, assume that the signal 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 meaning of the two

---

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

signal realizations.

Remark 3.1. The assumption that the designer's signal is binary is without loss of generality, given that the set of actions per voter is binary. This follows from familiar ideas from information design, but is stated and proved formally in Lemma 2 of Appendix B.

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

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

where the expression after the first equality presents the definition for a general prior, and the expression after the latter equality relies on having a uniform prior. Thus, under the uniform prior, symmetric signals in which $\text{P}\left[\mathfrak{a}|\theta_A\right] = \text{P}\left[\mathfrak{b}|\theta_B\right]$ are *unbiased*—that is, they have bias 0. Positive biases correspond to signals 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 is one with positive bias (Kamenica, 2019). We note that this definition of bias is an instance of the definition of Gentzkow et al. (2015) for media bias, under the assumption that the fully informative signal in which $\text{P}\left[s(\theta_A) = \mathfrak{a}\right] = \text{P}\left[s(\theta_B) = \mathfrak{b}\right] = 1$ 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.[6] We assume that voting is *sincere*: each voter votes according to their own information, which includes the prior and the two signals. Sincere voting can either directly reflect the voters' preferences (namely, 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.[7]

---

[6]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.

[7]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.

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 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 complimentary 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 election's outcome* (in which case A is still chosen only in state θ_A, regardless of the designer's signal), or that he can *perfectly manipulate* the election's outcome, such that A is chosen in both states. In the latter case, we say that a designer's signal s is *optimal* if it induces the choice of policy A in both states. Note that there might be multiple optimal signals.

4 Results

4.1 Preliminaries

Suppose a voter's belief that the state is θ_A is equal to $1/2$. Then any additional signal $s : \Theta \mapsto \{\mathfrak{a}, \mathfrak{b}\}$ induces a pair of posteriors (α, β), where $\alpha = \mathrm{P}[\theta_A | \mathfrak{a}]$ and $\beta = \mathrm{P}[\theta_A | \mathfrak{b}]$. The signal is uninformative if $\alpha = \beta = 0.5$. If the signal is 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 that induces these posteriors. Thus, we can identify any informative signal with the pair of posterior distributions it induces when starting with prior $\frac{1}{2}$.[8] For any $\alpha < \frac{1}{2} < \beta$, the exact mapping can be derived as follows: First, since the expected prior is equal to the posterior, we have

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

---

[8]Representing signals as distributions over posteriors is standard in Bayesian persuasion (Kamenica, 2019). Representing them as posteriors starting with a prior of $\frac{1}{2}$ is common in the literature on social learning (see, e.g., Acemoglu et al., 2011; Arieli et al., 2020, 2021).

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

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

Similarly,

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

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

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

Next, we have the following lemma, which allows us to focus in our analysis on a small set of 2×3 signals, such that, if the election is manipulable, one of these signals must be optimal (and any other optimal signal has weakly higher values of α and β).

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

The intuition for Lemma 1 is that for any signal, 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. See Appendix A for the proof. Table 1 summarizes the key properties of these 2×3 optimal signals (which are derived from straightforward calculations). Note that the share of voters who vote for policy A in state θ_A is always larger than in state θ_B, and, thus, a signal is optimal if and only if it induces at least half of the agents 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, B-favorable tie-breaking rules (namely, each agent 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 signals will be perturbed by a small $\varepsilon > 0$. Formally, if the designer can manipulate the elections, then there would be an optimal signal 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

Table 1: Key Properties of the 2×3 Signals 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 $\mathfrak{a}$	$\left(\lambda q_\ell + (1-\lambda)q_h\right)\frac{q_\ell(2q_h-1)}{q_h+q_\ell-1}$ (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)

sufficiently small $\varepsilon > 0$.

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

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 agents obtain each a private signal with accuracy q_h. An interesting special case, discussed towards the end of the 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.[9]

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

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

---

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

Proof. In either state, uninformed voters always vote for A, and informed voters vote for A if they receive signal a. In state θ_A, the latter make up a $q_h > 1/2$ share, and so overall a majority of voters always votes for A. In state θ_B, voters for A consist of the λ-fraction of uninformed voters and the $(1 - \lambda) \cdot (1 - q_h)$-fraction of informed voters who receive signal b. Thus, at least half of the voters vote for policy A in state θ_B if and only if $\lambda + (1 - \lambda) \cdot (1 - q_h) \geq \frac{1}{2} \Leftrightarrow \lambda \geq 1 - \frac{1}{2q_h} \Leftrightarrow q_h \leq \frac{1}{2(1-\lambda)}$. $\qquad\square$

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 election's outcome iff $q_h \leq \overline{q}(\lambda)$, and that manipulation (when possible) can be implemented by an unbiased signal. Moreover, for values of precision in the interval $(\frac{2}{3}, \overline{q}(\lambda)$ (values above the horizontal dotted (black) line in Figure 4.2), all optimal signals have non-positive biases.

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

1. If $q_h > \overline{q}(\lambda)$ then the designer cannot manipulate the election's result.

2. If $q_h \leq \overline{q}(\lambda)$ then the designer can manipulate the election's outcome, and can do so by using the unbiased signal $(q_h, 1 - q_h)$.

3. If $q_h \in (\frac{2}{3}, \overline{q}(\lambda)]$ then *all* optimal signals have non-positive biases.

An interesting feature of the optimal signals 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
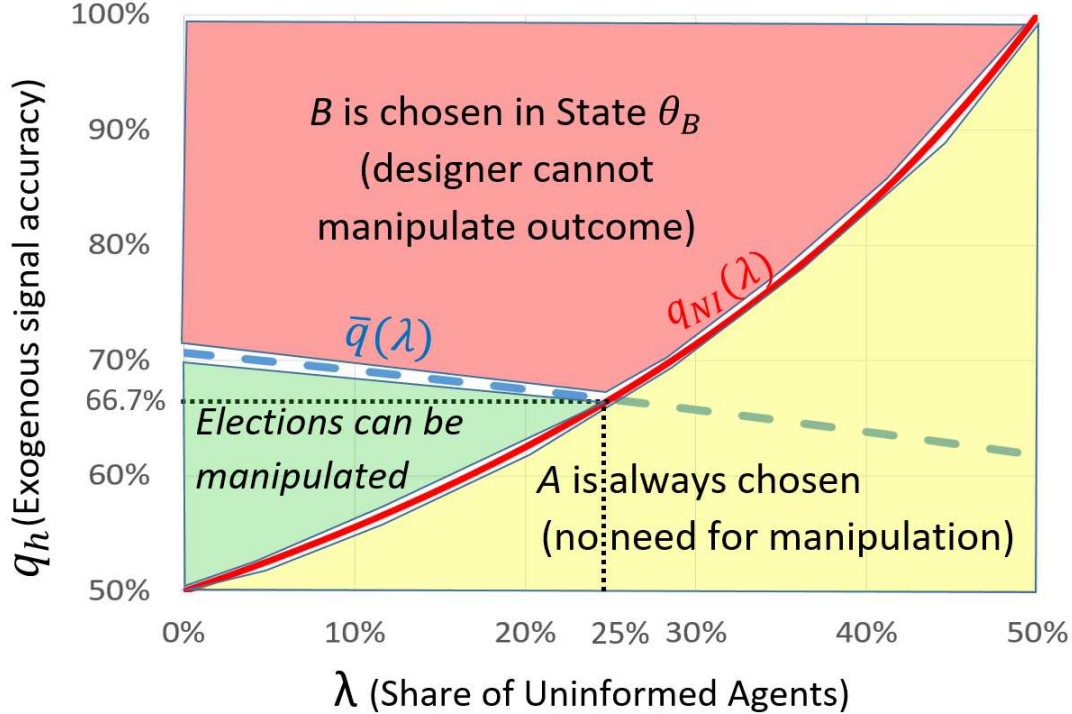
Figure 1: Illustration of Proposition 1 ($q_\ell = 0.5$)

standard application of Bayesian persuasion (Kamenica and Gentzkow, 2011), in which the bias is positive (i.e., in favor of A, the designer-optimal outcome). The intuition underlying the contrast between the standard application and our setup is the following. 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 also vote for A, while at the same time not causing the former to switch. The way to do this is to choose a signal 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 has non-positive bias. We now prove Proposition 1.

Proof. Claim 1 and the inequality $q_h > q_{NI}(\lambda)$ imply that policy A is not chosen in state θ_B if the signal is uninformative. Due to this we can focus on informative signals. Thus, Lemma 1 implies that if the designer can manipulate the election, then either $(q_h, 0)$ or $(q_h, 1 - q_h)$ are optimal. Signal $(q_h, 0)$ induces all voters to vote according to the designer's signal s (and to ignore the original signal). The probability of signal s being in favor of A in state θ_B is

$\Pr(s(\theta_B) = \mathfrak{a}) = \frac{1-2\beta-\alpha+2\alpha\beta}{\alpha-\beta} = \frac{1-q_h}{q_h}$, which is weakly larger than 50% iff $q_h \leq \frac{2}{3}$.

Signal $(q_h, 1-q_h)$ induces informed (resp., uninformed) agents to vote for A if either of their two signals (resp., the designer's signal) is in favor of A, which has probability of $1-q_h^2$ (resp., $1-q_h$) in state θ_B. Thus, the share of voters who vote for policy A in state θ_B is at least 50% iff $\lambda \cdot (1-q_h) + (1-\lambda) \cdot (1-q_h^2) \geq 50\% \Leftrightarrow q \leq \overline{q}(\lambda)$. Combining these observartions with Fact 1 implies bullets (1) and (2) of the proposition.

The argument for bullet (3) is the following. Due to Lemma 1, if $\frac{2}{3} < q_h \leq \frac{\sqrt{2}}{2}$, then any optimal signal (α, β) must satisfy $\alpha \geq q_h$ and $\beta \geq 1 - q_h$. In this case, the numerator in (1), namely, $1 - (\alpha + \beta)$, is non-positive. Since the denominator is always positive, this implies that the bias is non-positive. $\square$

An interesting special case is a homogeneous population in which all agents have the same signal quality (i.e., $\lambda = 0$). Applying Proposition 1 yields the following characterization of election 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 election's result.

2. If $q_h \leq \frac{\sqrt{2}}{2}$ then the designer can manipulate the election's outcome, and can do so by using the unbiased signal $(q_h, 1-q_h)$.

3. If $q_h \in (\frac{2}{3}, \frac{\sqrt{2}}{2}]$ then *all* optimal signals 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 election; and second, an identification of conditions under which optimal signals have positive or negative biases.

In order to state our first result we define $\underline{\lambda} : (\frac{1}{2}, \frac{\sqrt{2}}{2})] \rightarrow [0,1]$ 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 $(q_h, 1-q_h)$ 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 election.

2. If $q_\ell < \frac{\sqrt{2}}{2} < q_h$, then the designer can manipulate the election iff λ is sufficiently large.

3. If $\frac{2}{3} < q_h \leq \frac{\sqrt{2}}{2}$, then the designer can manipulate the elections iff either (1) $\lambda \leq \underline{\lambda}(q_h)$, or (2) λ is sufficiently large.

4. If $q_h \leq \frac{2}{3}$, then the designer can manipulate the elections for any q_ℓ and λ.

Proof.

1. Corollary 1 implies that for each $q > \frac{\sqrt{2}}{2}$ all signals induce less than 50% of agents with signal accuracy q to vote for policy A in state θ_B. This implies that if $q_\ell > \frac{\sqrt{2}}{2}$, then a majority of the voters support policy B in state θ_B (i.e., the designer cannot manipulate the elections).

2. This is implied by the observation that when $\frac{\sqrt{2}}{2} < q_h$ both signals $(q_h, 0)$ and $(q_h, 1 - q_h)$ cannot manipulate the elections, while the share of agents who support policy A in state θ_B is increasing in λ (and is greater than than 50% for sufficiently high λ) for each of the other optimal signals of Lemma 1 (see Table 1). Signal $(q_h, 0)$ cannot manipulate the elections because the induced share of B supporters in state θ_B is $\frac{2q_h - 1}{q_h}$, which is larger than 50% iff $q_h > \frac{\sqrt{2}}{2}$. Signal $(q_h, 1 - q_h)$ induces a share $q_h^2 + \lambda q_h(1 - q_h)$ of the voters to vote for policy B in state θ_B (see Table 1). This share is smaller than 50% iff $\lambda \leq \underline{\lambda}(q_h)$ because $q_h^2 + \lambda q_h(1 - q_h) \leq 0.5 \Leftrightarrow \lambda \leq \frac{0.5 - q_h^2}{q_h(1 - q_h)} \equiv \underline{\lambda}(q_h)$. As observed in Fact 2, $\underline{\lambda}(\frac{\sqrt{2}}{2}) = 0$, which implies that signal $(q_h, 1 - q_h)$ cannot manipulate the elections if $q_h > \frac{\sqrt{2}}{2}$.

3. Part (1) is implied by the observation that the designer's signal $(q_h, 1 - q_h)$ induces a share $q_h^2 + \lambda q_h(1 - q_h)$ to vote for policy B in state θ_B, and that this share is less than 50% if $\lambda \leq \underline{\lambda}(q_h)$. Part (2) is implied by the fact that signal $(q_h, 0)$ (resp., $(q_h, 1 - q_h)$) does not manipulate the elections when $q_h > \frac{2}{3}$ (resp., $\lambda > \underline{\lambda}(q_h)$), and that the share of supporters of policy A in state θ_B is increasing in λ for each of the other four signals of Lemma 1.

4. Observe (Table 1) that the designer's signal $(q_h, 0)$ induces a share of $\frac{2q_h - 1}{q_h}$ of the agents to vote for B in state θ_B (for all values of $q_\ell \leq q_h$), and that this share is smaller than 50% iff $q_h \leq \frac{2}{3}$. $\qquad\square$

Our results are illustrated in Figure 2. Bullet (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 elections are manipulable if either the share of agents with low accuracy is sufficiently small (namely, $\lambda \leq \underline{\lambda}$) or if it is sufficiently large, but the elections 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 he must choose the same signal distribution to all agents. When λ is either very low or very high, the designer can tune this distribution to the large majority of agents by sending an unbiased signal with the "right" level of accuracy (i.e., a level that induces most voters to vote 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 agents (and is not optimally "tailored" to either group), and this limits his ability to manipulate the elections.

Moreover, one can show that the ability to manipulate the elections is also non-monotonic in q_ℓ. Specifically, for each q_h, the interval of λ-s for which the elections 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 agents. The designer can send a positively biased signal that with high probability sends a "weak-pro-A" realization, which manipulates most low-accuracy agents (and a sufficient number of high-accuracy agents). When q_ℓ is close to q_h, the designer can exploit the relative homogeneity of the population by sending the negatively biased signal $(q_h, 1-q_\ell)$, which induces both types of agents to vote for A if either of their signals is "pro-A". In contrast, when q_ℓ has an intermediate value the designer has neither of these two benefits, and thus his ability to manipulate the elections is limited.

Recall that in the special case of $q_\ell = 0.5$ (partially informed agents), manipulation, when possible, could always be implemented by an unbiased signal (and in some cases only by signals 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
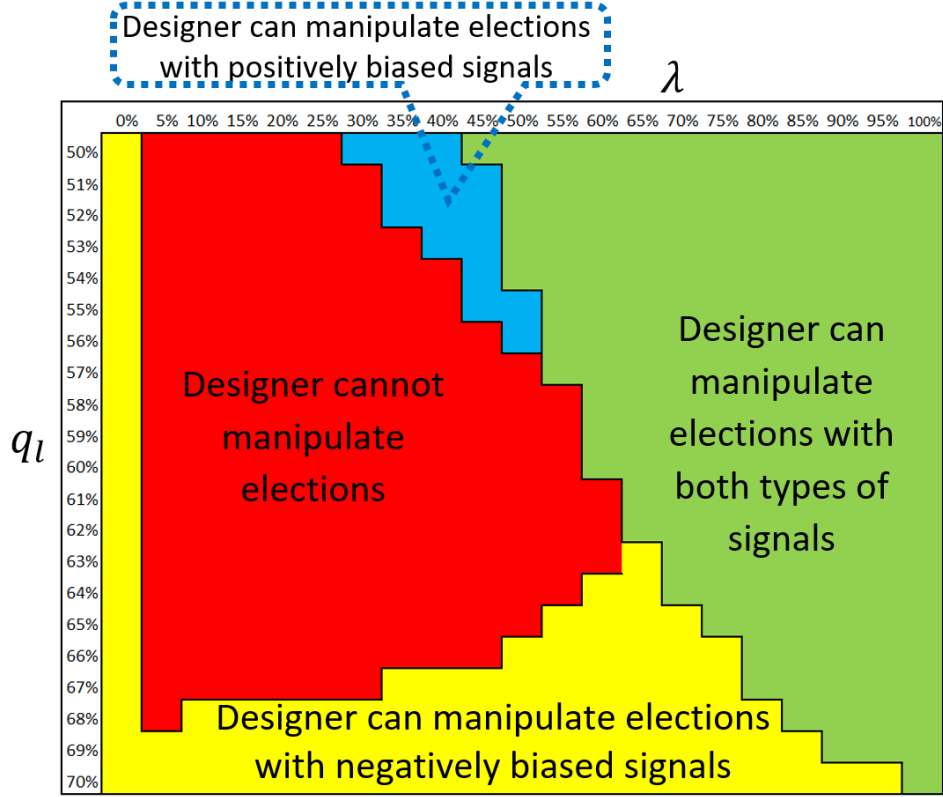
Figure 2: Outcome Manipulation for $q_h = 70\%$ and Various Values of q_ℓ (values between 50% and 70% in steps of 1%) and λ (values between 0 and 100% in steps of 5%)

be implemented by positively biased signals (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 signals. 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 signals must have a positive bias if $q_h > \frac{2}{3}$ and q_ℓ is sufficiently low.

2. All optimal signal must 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 signals with the opposite direction of bias relative to the standard applications of Bayesian persuasion (see, e.g., Kamenica, 2019).The intuition for this is that the optimal negatively biased signal $(q_h, 1 - q_\ell)$ exploits the (modest) heterogeneity of the population: the relatively infrequent signal $\mathfrak{a}$ is sufficiently informative to induce both types of agents to vote A regardless of their exogenous signal, while the more frequently used signal $\mathfrak{b}$ is sufficiently weakly-informative, such that it

does not change the voting of any agent who got exogenous signal in favor of policy A.

Proposition 3.

1. For any $q_h > \frac{2}{3}$ there is a $\bar{q} \in (0.5, q_h)$ such that, if $q_\ell \in (0.5, \bar{q})$, then there is an interior open interval of λ-s for which all optimal signals have positive biases.

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

Proof. Fix $q_h > \frac{2}{3}$. Proposition 2 and Fact 2 imply that the signals $(q_h, 0)$ and $(q_h, 1 - q_h)$ are not optimal if $\lambda > 0.25 - \epsilon$ for a sufficiently small $\epsilon > 0$.

1. The limit of the share of B voters in state θ_B when q_ℓ converges to 0.5 (see Table 1) is equal to $q_h(1 - \lambda)$ for the positively biased signals $(q_\ell, 0)$ and $(q_\ell, 1 - q_h)$, while it is strictly higher for the remaining two signals: it is equal to $q_h(1 - \lambda) + \frac{1}{4}$ for signal $(q_\ell, 1 - q_\ell)$, and it is equal to $q_h(1 - \lambda) + \frac{\lambda}{2}$ for signal $(q_h, 1 - q_\ell)$. This implies that for sufficiently small q_ℓ-s, if λ is slightly above $1 - \frac{1}{2q_h}$, then only the positively biased signals $(q_\ell, 0)$ and $(q_\ell, 1 - q_h)$ are optimal in the set of six signals of Lemma 1. By Lemma 1, this implies that the only potentially optimal signals are $(\alpha', \beta') \in [q_\ell, q_h) \times [0, 1 - q_h)$ or $(\alpha', \beta') \in [q_\ell, q_h) \times [1 - q_h, 1 - q_\ell)$. Note that the former are all positively biased. The latter, however, may be either positively or negatively biased.

   We now show that only the positively biased signals $(\alpha', \beta') \in [q_\ell, q_h) \times [1 - q_h, 1 - q_\ell)$ are potentially optimal. Fix some $(\alpha', \beta') \in [q_\ell, q_h) \times [1 - q_h, 1 - q_\ell)$, and define q'_ℓ and q'_h so that $(\alpha', \beta') = (q'_\ell, 1 - q'_h)$. Observe that this signal is positively biased if $q'_\ell < q'_h$, and negatively biased if the inequality is flipped. Under this signal, the fraction of voters for B in state θ_B is

   $$ \mathrm{P}\left[B | \theta_B \right] = \lambda \frac{q'_h(2q'_\ell - 1)}{q'_h + q'_\ell - 1} + (1 - \lambda) q_h. $$

   Now, if the signal is negatively biased, and so $q'_\ell \geq q'_h$, then $2q'_\ell - 1 \geq q'_h + q'_\ell - 1$, and so the first term of $\mathrm{P}\left[B | \theta_B \right]$ is bounded below by $\lambda q'_h \geq \lambda/2$. The total fraction of voters for B is then at least $q_h(1 - \lambda) + \lambda/2$, the same as for signal $(q_h, 1 - q_\ell)$. But just like that signal, if λ is slightly above $1 - \frac{1}{2q_h}$ then a negatively biased signal $(\alpha', \beta') \in [q_\ell, q_h) \times [1 - q_h, 1 - q_\ell)$ is also not optimal.

23

2. Assume $q_h \in (\frac{2}{3}, \frac{\sqrt{2}}{2})$. The limit of the share of B voters in state θ_B when q_ℓ converges to q_h (see Table 1) for each of the other four signals is equal to:

   (a) Signal $(q_\ell, 0)$: The limit is equal to $1 - \frac{(1-q_h)(1-(1-\lambda)q_h)}{q_h}$, which is increasing in q_h and decreasing in λ. When substituting $q_h = \frac{2}{3}$ and $\lambda = 1$, the expression is equal 50%, which implies that it is larger than 50% for each $q_h > \frac{2}{3}$ and each λ.

   (b) Signal $(q_\ell, 1 - q_\ell)$: The limit is equal to $q_h(1 - \lambda(1 - q_h))$, which is increasing in q_h and decreasing in λ. When substituting $q_h = \frac{2}{3}$ and $\lambda = \frac{2}{3}$, the expression is equal $\frac{14}{27} > 50\%$, which implies that the limit is larger than 50% for each $q_h > \frac{2}{3}$ and each $\lambda < \frac{2}{3}$.

   (c) Signal $(q_\ell, 1 - q_h)$: The limit is equal to $q_h > \frac{2}{3}$, which implies that the limit is larger than 50% for each $q_h > \frac{2}{3}$ and each λ.

   (d) Signal $(q_h, 1 - q_\ell)$: The limit is equal to q_h^2, which is smaller than 50% for all values of λ because $q_h < \frac{\sqrt{2}}{2}$.

This implies that when q_ℓ is sufficiently close to q_h and $\lambda \in (0.25, \frac{2}{3})$, then the unique optimal signal in the set of six signals of Lemma 1 is $(q_h, 1 - q_\ell)$, which implies that all optimal signals are negatively biased due to Lemma 1. $\qquad\square$

5 Extensions

5.1 Continuous Signal Accuracies

In this section we extend our model to a general distribution over signal accuracies (and not just binary). 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 supp(f) in $[0.5, 1]$. We note that the closely related extension of the model to a setup with a finite (but greater than 2) number of signal accuracies is analogous (and is omitted for brevity).

Our result here is an extension of Proposition 2 to this general setup, showing that the election is manipulable if all agents have accuracies of at most $\frac{2}{3}$, and that it is non-manipulable if all agents have accuracies of at least $\frac{\sqrt{2}}{2}$. Formally,

Proposition 4.

1. If $\text{supp}(f) \subseteq \left[\frac{\sqrt{2}}{2}, 1\right]$, then the designer cannot manipulate the election.

2. If $\text{supp}(f) \subseteq \left[\frac{1}{2}, \frac{2}{3}\right]$, then the designer can manipulate the election.

Proof.

1. Corollary 1 implies that for each $q > \frac{\sqrt{2}}{2}$ all signals induce less than 50% of agents with signal accuracy q to vote for policy A in state θ_B. This implies that when $\text{supp}(f) \subseteq \left[\frac{\sqrt{2}}{2}, 1\right]$, a majority of the voters support policy B in state θ_B irrespective of the exact distribution of signal accuracies (i.e., the designer cannot manipulate the elections).

2. Let $q_h = \sup_{q \in Q} q$. Observe (Table 1) that the designer's signal $(q_h, 0)$ induces a share of $\frac{2q_h - 1}{q_h}$ of the agents to vote for B in state θ_B (for all values of $q \leq q_h$), and that this share is smaller than 50% iff $q_h \leq \frac{2}{3}$. $\qquad\square$

5.2   Targeted Information

In our main analysis, we required the designer to design one signal for all voters. What if the designer can differentiate between voters based on their signal accuracies, and can thus design different signals for different voters?

Suppose that the designer can distinguish between the two types of voters, q_ℓ and q_h. Furthermore, suppose she can commit to two different signals, 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 is optimal regardless of the type of voter, and so the designer can manipulate the election, whereas if $q_h > q_\ell > \sqrt{2}/2$ then the designer cannot manipulate the election. The proof is the same as that of the proposition.

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

larger. Overall, the designer can manipulate the election 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 with 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 election extends in a straightforward manner to the variant with a continuous distribution of signal accuracies from Section 5.1 above. In this case, the designer can manipulate the election 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 the realization of this exogenous signal? In this case, the designer can provide a strongly targeted signal, in which voters with different signal accuracies *and different realizations* get different signals.

For a given signal accuracy q, a voter with realization a has interim belief q about the probability 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 kinds voters. In order to maximize the probability that such voters vote for A, the designer should supply the former voter with no additional signal, and the latter voter with additional signal $(q, 0)$. The latter signal here is the optimal signal from the standard Bayesian persuasion setting (Kamenica and Gentzkow, 2011), supplied to a voter who has belief $1-q$ and is will to vote for A once the 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 fraction 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 agents have accuracy q, the designer can ma-

nipulate the elections iff $2(1 - q) \geq \frac{1}{2} \Leftrightarrow q \leq \frac{3}{4}$. When there are two signals accuracies, q_ℓ and q_h, the designer can then manipulate the election 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 accuracies, as in Section 5.1 above, the designer can manipulate the election 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, all voters obtain the same realization. Is such a public signal better or worse for the designer?

Observe that with a public signal, the designer can never manipulate the election with probability 1. However, she can always at least slightly increase the probability that voters vote for A (with a positively biased signal as in Kamenica and Gentzkow, 2011). Thus, the answer to whether this is better than private persuasion depends on whether or not the designer can manipulate the election 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.

6 Conclusion

In portraying private beliefs that contradict public information, Kuran (1997) describes a government monopoly on information under autocracy. Monopoly on information also arises under democracy when, because of network externalities, a social media platform caters to designated users seeking to maintain contact with one another. In this paper we showed that manipulation through personalized private information in social media can influence beliefs. This, in turn,

may invalidate the conclusion of Condorcet's jury theorem, a theorem that captures the benefits of majority voting. In particular, information manipulation through private persuasion can result in a majority voting for an information designer's preferred policy rather than for a socially preferred alternative. This compromise of Condorcet's jury theorem undermines the merits of democracy.

From the perspective of theory, we focused on the direct effect of informational persuasion on voting outcomes by dispensing with the assumption, prominent in the literature on information design, that voters regard themselves as possibly pivotal. Instead, we viewed voting as sincere, and accounted for the paradox of voting by assuming that voters derive direct utility from expressing themselves through voting. Our general model covers a range of circumstances that differ in the extent to which users are exogenously informed and in the amount of heterogeneity in the accuracy levels of users' exogenous information. In all these circumstances we provided tight characterizations of when Condorcet's jury theorem can be overturned.

Our analysis and results provide a comparison between information manipulation by private persuasion through social media and by public persuasion through traditional media. Under public persuasion, the designer can always slightly increase the probability that voters choose the designer's preferred outcome, but can never determine the outcome with probability one. Under private persuasion, in contrast, our results show that the designer either completely determines the outcome, or has no effect at all, depending on the circumstances. In the former case, a designer with a political agenda would thus prefer control of social media to information manipulation through traditional media, whereas in the latter case, the designer's preference would be reversed.

A   Appendix: Proof of Lemma 1

Fix any optimal (α', β'), and note that, under this signal,

$$\mathrm{P}\left[\mathfrak{a}|\theta_B\right] = 1 - \frac{2\alpha' + \beta' - 1 - 2\alpha'\beta'}{\alpha' - \beta'} = \frac{(\alpha' - 1)(2\beta' - 1)}{\alpha' - \beta'}.$$

Observe that this probability is decreasing in α' and β', and so lowering either of these leads to a higher probability of realization $\mathfrak{a}$ in state θ_B.

Suppose now that $\alpha' \notin \{q_\ell, q_h\}$. First, observe that α' cannot lie in the interval $[1/2, q_\ell)$. This is because, if it did, then when a voter obtains exogenous signal b, neither the pair $(b, \mathfrak{a})$ nor the pair $(b, \mathfrak{b})$ will cause the voter to vote A. This means that voters always vote B on signal b. However, in state θ_B the signal b is more likely than a, and so a majority will always vote for B in this state. But this implies that (α', β') is not optimal, a contradiction.

Suppose next that $\alpha' \in (q_\ell, q_h)$. In this case, switching to signal (α, β'), with $\alpha = q_\ell$ does not affect the behavior of any voter: Voters with signals $(a, \mathfrak{b})$ and $(b, \mathfrak{b})$ still make the same inferences, whereas voters with signals $(a, \mathfrak{a})$ still vote for A, q_ℓ voters with signals $(b, \mathfrak{a})$ still vote for A, and q_h voters with signals $(b, \mathfrak{a})$ still vote for B. Similarly, if $\alpha' \in (q_h, 1)$, then switching to signal (α, β'), with $\alpha = q_h$ does not affect the behavior of any voter. In all cases, however, lowering α' to α increases the probability of signal $\mathfrak{a}$, and hence a weakly greater fraction of voters votes for A. This implies that (α, β') is optimal.

Finally, suppose $\beta' \notin \{0, 1 - q_h, 1 - q_\ell\}$. If $\beta' \in (0, 1 - q_h)$ (resp., $\beta' \in (1 - q_h, 1 - q_\ell)$ or $\beta' \in (1 - q_\ell, 1)$) then lowering β' to $\beta = 0$ (resp., $\beta = 1 - q_h$ or $\beta = 1 - q_\ell$) does not affect the behavior of any voter. In all cases, however, lowering β' to β increases the probability of signal $\mathfrak{a}$, and hence a weakly larger share of voters votes for A. This implies that (α, β) is optimal.

B   Appendix: Binary Signals are Optimal

In this section we show that the restriction of the designer's signal to a binary one is without loss of generality. Kamenica and Gentzkow (2011) show that, in the standard Bayesian persuasion setting with a single, uninformed receiver, it is without loss of generality to restrict the designer to a *straightforward* signal—one where each signal realization corresponds to a distinct action, these realizations are interpreted by the receiver as recommendations, and the receiver optimally follows the recommendation.

When there are different types of receivers (in our case, ones with different exogenous-signal accuracies) who have different information (in our case, different realizations of the exogenous signal), the argument of Kamenica and Gentzkow (2011) does not directly apply. This is because

different receivers would optimally take different actions even given the same signal realization, and so these realizations cannot be interpreted as recommendations. For a simple example consider homogeneous voters with signal accuracy $q \in (0.5, 1)$, and where the designer's signal is $(0.5+\varepsilon, 0.5-\varepsilon)$ for small $\varepsilon > 0$. In this case each voter votes according to their own exogenous information, and not according to the sender's "recommended" action $\mathfrak{a}$ or $\mathfrak{b}$.

However, an extended argument does show that binary signals are sufficient. Consider the general model of Section 4.3. Then:

Lemma 2. If there exists an optimal signal, then there exists an optimal binary signal.

Proof. In our model there are four different types of voters, distinguished by their signal accuracies (q_ℓ or q_h) and the realization of their exogenous signal (a or b). We note that our argument works for any finite, arbitrarily large number of voters' types.

Now, suppose there exists an optimal signal. We first argue that there also exists an optimal signal with a finite number of realizations. This follows from the argument of Kamenica and Gentzkow (2011), which implies that there exists an optimal signal in which there are at most 16 realizations, as follows: Each realization of the signal is a recommendation to vote either for A or for B for each of the four different types of voters. There are 2^4 such four-tuples of recommendations, hence 16 signal realizations.

Next, let s be an optimal designer signal with the *minimal* number of signal realizations, and suppose towards a contradiction that s is not binary. Let $n > 2$ be the number of its signal realizations, and denote these realizations by $\text{supp}(s) = \{r_1, \ldots, r_n\}$. By Lemma 3 below, the signal s is equal to a distribution over two signals, s_1 and s_2, with the following four properties: (1) $s \equiv \eta s_1 + (1-\eta)s_2$ with $\eta \in (0,1)$, (2) $|\text{supp}(s_1)|, |\text{supp}(s_2)| < n$, (3) $|\text{supp}(s_1) \cap \text{supp}(s_2)| \leq 1$, and (4) if $r_i \in \text{supp}(s_1) \cap \text{supp}(s_2)$ then $\mathrm{P}\,[\theta_B|s_1 = r_i] = \mathrm{P}\,[\theta_B|s_2 = r_i]$.

Now, s is optimal if and only if the probability it induces a random voter to vote for A in state θ_B is at least $1/2$. Since s is a mixture of s_1 and s_2 (and the unique shared realization induces the same posterior), either s_1 or s_2 must also be optimal. For otherwise, if in state θ_B a random voter has probability less than $1/2$ of voting for A under each of s_1 and s_2, then this must also be the case under s. However, s_1 and s_2 have fewer realizations than s, contradicting the assumption that s is an optimal signal with the minimal number of realizations. □

Lemma 3. Fix a non-binary signal s with $n > 2$ signal realizations, and denote those realizations by $\text{supp}(s) = \{r_1, \ldots, r_n\}$. Then there exist signals s_1, s_2 with the following properties:

1. $s \equiv \eta s_1 + (1 - \eta) s_2$, where $\eta \in (0, 1)$,

2. $|\text{supp}(s_1)|, |\text{supp}(s_2)| < n$.

3. $|\text{supp}(s_1) \cap \text{supp}(s_2)| \leq 1$, and

4. if $r_i \in \text{supp}(s_1) \cap \text{supp}(s_2)$ then $\text{P}\left[\theta_B | s_1 = r_i\right] = \text{P}\left[\theta_B | s_2 = r_i\right]$.

Proof. Denote by $(\alpha_1, \ldots, \alpha_n)$ the posteriors induced by the n signal realizations of s on state θ_A, starting with prior $1/2$. This is simply the generalization of the (α, β) notation for binary signals to signals with more than two realizations. Suppose r_i is the signal realization associated with posterior α_i, and denote by $p_i = \text{P}\left[s = r_i\right]$ (the *unconditional* probability of this realization).

If signal s is non-informative, then the result is immediate (as each s_i can be a non-informative signal with a single realization). Otherwise, there must be at least one α_i above $1/2$ and at least one α_j below $1/2$. Without loss of generality, assume that $\alpha_1 > \ldots > \alpha_n$, and note that $\alpha_1 > 1/2$ and $\alpha_n < 1/2$.

Consider the binary signal $s_1 = (\alpha_1, \alpha_n)$ with $\text{supp}(s_1) = \{r_1, r_n\}$. For this signal, let

$$p_{\mathfrak{a}} = \text{P}\left[s_1 = r_1\right] = \frac{\frac{1}{2} - \alpha_n}{\alpha_1 - \alpha_n} \quad \text{and} \quad p_{\mathfrak{b}} = \text{P}\left[s_1 = r_n\right] = \frac{\alpha_1 - \frac{1}{2}}{\alpha_1 - \alpha_n},$$

where the respective probabilities are computed as in Section 4.1.

We now consider three cases. First, if $\frac{p_{\mathfrak{a}}}{p_{\mathfrak{b}}} = \frac{p_1}{p_n}$, then let $\eta = p_1 + p_n$, and let s_2 be the signal whose realizations are $\{r_2, \ldots, r_{n-1}\}$, and where $\text{P}\left[s_2(\theta) = r_i\right] = \text{P}\left[s(\theta) = r_i\right]/(1 - \eta)$ for each $i \in \{2, \ldots, n - 1\}$ and $\theta \in \{\theta_A, \theta_B\}$. Observe that s_2 yields posterior beliefs $(\alpha_2, \ldots, \alpha_{n-1})$. It is straightforward to verify that these satisfy the claim of the lemma.

Second, suppose $\frac{p_{\mathfrak{a}}}{p_{\mathfrak{b}}} < \frac{p_1}{p_n}$. In this case, let $\eta = \frac{p_{\mathfrak{a}}}{p_{\mathfrak{b}}} p_n + p_n$, and let s_2 be the signal whose realizations are $\{r_1, r_2, \ldots, r_{n-1}\}$, where $\text{P}\left[s_2(\theta) = r_i\right] = \text{P}\left[s(\theta) = r_i\right]/(1 - \eta)$ for each for each $i \in \{2, \ldots, n-1\}$ and $\theta \in \{\theta_A, \theta_B\}$, and where $\text{P}\left[s_2(\theta) = r_1\right] = (\text{P}\left[s(\theta) = r_1\right] - \eta \text{P}\left[s_1(\theta) = r_1\right])/(1 - \eta)$ for each $\theta \in \{\theta_A, \theta_B\}$. Observe that s_2 yields posterior beliefs $(\alpha_1, \alpha_2, \ldots, \alpha_{n-1})$. Again, it is straightforward to verify that these satisfy the claim of the lemma. In particular, the only

shared realization r_1 induces the same posterior belief under both signals: $\mathrm{P}\left[\theta_A|s_1 = r_1\right] = \mathrm{P}\left[\theta_A|s_2 = r_1\right]$.

Finally, suppose $\frac{p_{\mathfrak{a}}}{p_{\mathfrak{b}}} > \frac{p_1}{p_n}$. This case is analogous to the second case, except that $\eta = p_1 + \frac{p_{\mathfrak{b}}}{p_{\mathfrak{a}}}p_1$, r_n replaces r_1 in $\mathrm{supp}(s_2)$, the signal s_2 yields beliefs $(\alpha_2, \ldots, \alpha_{n-1}, \alpha_n)$, and s_1 and s_2 share the realization r_n. $\qquad\square$

References

ACEMOGLU, D., DAHLEH, M. A., LOBEL, I. and OZDAGLAR, A. (2011). Bayesian learning in social networks. *The Review of Economic Studies*, **78** 1201–1236.

ALONSO, R. and CÂMARA, O. (2016). Persuading voters. *American Economic Review*, **106** 3590–3605.

ARIELI, I. and BABICHENKO, Y. (2019). Private bayesian persuasion. *Journal of Economic Theory*, **182** 185–217.

ARIELI, I., BABICHENKO, Y. and SMORODINSKY, R. (2020). Identifiable information structures. *Games and Economic Behavior*, **120** 16–27.

ARIELI, I., GRADWOHL, R. and SMORODINSKY, R. (2021). Herd design. *Available at SSRN 3917729*.

AUSTEN-SMITH, D. and BANKS, J. S. (1996). Information aggregation, rationality, and the Condorcet jury theorem. *American Political Science Review*, **90** 34–45.

BERMAN, R., ZHAO, H. and ZHU, Y. (2021). When (not) to persuade consumers: Persuasive and demarkting information designs. *mimeo*.

CHAN, J., GUPTA, S., LI, F. and WANG, Y. (2019). Pivotal persuasion. *Journal of Economic theory*, **180** 178–202.

DE CONDORCET, N. (1785). *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.

DENTER, P., DUMAV, M. and GINZBURG, B. (2021). Social connectivity, media bias, and correlation neglect. *The Economic Journal*, **131** 2033–2057.

DOWNS, A. (1957). *An Economic Theory of Democracy*. Harper & Row New York, NY.

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

FEDDERSEN, T. and PESENDORFER, W. (1997). Voting behavior and information aggregation in elections with private information. *Econometrica: Journal of the Econometric Society* 1029–1058.

FEDDERSEN, T. and PESENDORFER, W. (1998). Convicting the innocent: The inferiority of unanimous jury verdicts under strategic voting. *American Political science review*, **92** 23–35.

GENTZKOW, M., SHAPIRO, J. M. and STONE, D. F. (2015). Media bias in the marketplace: Theory. In *Handbook of Media Economics*, vol. 1. Elsevier, 623–645.

GITMEZ, A. A. and MOLAVI, P. (2022). Polarization and media bias. *arXiv preprint arXiv:2203.12698*.

GROSSMAN, G. M. and HELPMAN, E. (1999). Competing for endorsements. *American Economic Review*, **89** 501–524.

GROSSMAN, G. M. and HELPMAN, E. (2001). *Special Interest Politics*. MIT press.

HEESE, C. and LAUERMANN, S. (2021). Persuasion and information aggregation in elections. Tech. rep., ECONtribute Discussion Paper.

HILLMAN, A. L. (2010). Expressive behavior in economics and politics. *European Journal of political economy*, **26** 403–418.

HILLMAN, A. L. and URSPRUNG, H. W. (1988). Domestic politics, foreign interests, and international trade policy. *American Economic Review* 729–745.

HOTELLING, H. (1929). Stability in competition. *The Economic Journal*, **39** 41–57.

KAMENICA, E. (2019). Bayesian persuasion and information design. *Annual Review of Economics*, **11** 249–272.

KAMENICA, E. and GENTZKOW, M. (2011). Bayesian persuasion. *American Economic Review*, **101** 2590–2615.

KERMAN, T., HERINGS, P. J.-J. and KAROS, D. (2020). Persuading strategic voters.

KORIYAMA, Y. and SZENTES, B. (2009). A resurrection of the Condorcet jury theorem. *Theoretical Economics*, **4** 227–252.

KURAN, T. (1997). *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Harvard University Press.

LEVY, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, **111** 831–70.

LICHTER, S. R. (2017). Theories of media bias. In *The Oxford Handbook of Political Communication*. Oxford University Press New York, NY, 403–416.

MUKHOPADHAYA, K. (2003). Jury size and the free rider problem. *Journal of Law, Economics, and Organization*, **19** 24–44.

PONS, V. and TRICAUD, C. (2018). Expressive voting and its cost: Evidence from runoffs with two or three candidates. *Econometrica*, **86** 1621–1649.

PRAT, A. (2002). Campaign spending with office-seeking politicians, rational voters, and multiple lobbies. *Journal of Economic Theory*, **103** 162–189.

PUGLISI, R. and SNYDER JR, J. M. (2015). Empirical studies of media bias. In *Handbook of media economics*, vol. 1. Elsevier, 647–667.

SPENKUCH, J. L. (2018). Expressive vs. strategic voters: An empirical assessment. *Journal of Public Economics*, **165** 73–81.

STRÖMBERG, D. (2015). Media and politics. *Annual Review of Economics*, **7** 173–205.

SUN, J., SCHRAM, A. and SLOOF, R. (2019). A theory on media bias and elections. Tech. rep., Tinbergen Institute Discussion Paper.

TANEVA, I. (2019). Information design. *American Economic Journal: Microeconomics*, **11** 151–85.

TWITTER (2022a). Help with locked or limited account. https://help.twitter.com/en/managing-your-account/locked-and-limited-accounts. [Online; accessed 12-June-2022].

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

WANG, Y. (2013). Bayesian persuasion with multiple receivers. *Available at SSRN 2625399*.