Troll Farms and Voter Disinformation[*]

Philipp Denter[†] Boris Ginzburg[‡]

23rd May 2023

Abstract

Political agents often attempt to influence elections through "troll farms" – organisations that flood social media platforms with messages emulating genuine information. We model the behaviour of a troll farm that faces a heterogeneous electorate of partially informed voters, and aims to achieve a desired political outcome by targeting each type of voter with a specific distribution of messages. We show that such tactics are more effective when voters are otherwise well-informed. Consequently, societies with high-quality media are more vulnerable to electoral manipulation, and counteracting troll farms may require promotion of informative but non-expert opinions. At the same time, increased polarisation, as well as deviations from Bayesian rationality, can reduce the negative effect of troll farms and restore the efficiency of electoral outcomes.

Keywords: social media, fake news, elections, polarisation, bounded rationality.
JEL Codes: D72, D83, D91

[*] We thank seminar audiences in Bielefeld and Lancaster for valuable feedback. We also gratefully acknowledge the support from the Ministerio Economia y Competitividad (Spain) through grant PGC2018-098510-B-I00 and of the Comunidad de Madrid (Spain) through grants EPUC3M11 (V PRICIT) and H2019/HUM-5891.

[†] Universidad Carlos III de Madrid, Department of Economics, Calle de Madrid 126, 29803 Getafe, Spain. E-Mail: pdenter@eco.uc3m.es.

[‡] Universidad Carlos III de Madrid, Department of Economics, Calle de Madrid 126, 29803 Getafe, Spain. E-Mail: bginzbur@eco.uc3m.es.

1 Introduction

"*The Democrats don't matter. The real opposition is the media. And the way to deal with them is to flood the zone with shit.*"

Attributed to Steve Bannon

One of the consequences of the rise of social media is that voters receive a significant amount of politically relevant information through social media platforms. This has enabled voters to acquire information from many sources, but has also given malicious agents new tools to manipulate information.

One growing concern are the so-called troll farms, also known as bot factories or keyboard armies – groups of coordinated social media accounts that disseminate propaganda by emulating messages from real individuals or reliable news sources. Such troll farms are increasingly common – one report estimates that in 38 out of 65 surveyed countries, political leaders use them to manipulate elections and other domestic political events (Freedom House, 2019).[1] Trolls farms have also used to manipulate elections abroad by governments such as Russia, China, Iran, and Saudi Arabia (Martin et al., 2019).

Several features of social media troll farms can make them more effective than more traditional propaganda tools such as biased media. First, unlike messages delivered via traditional channels, messages from troll farms shared via social media are able to emulate genuine information. Thus, the target audience is left uncertain whether a particular message comes from a genuine individual communicating a piece of news or a personal experience, or from a bot. Second, social media platforms collect a considerable amount of personal data, including information about political preferences. This makes it possible to target particular voters with messages designed to maximise the persuasive effect on a given voter. In fact, the ability of social media influence operations to exploit user characteristics in order to target them with custom-made political messages has been demonstrated by the Cambridge Analytica scandal in the 2010s as well as by other examples of troll farms microtargetting voters.[2] Third, modern technology allows large numbers of trolls to be deployed at almost no cost, as multiple fake accounts can be controlled by a single user or even by automated algorithms.[3]

[1] These countries include Iran, Philippines, Thailand, and others (The New Republic, 2017; Reuters, 2020).

[2] For example, prior to the 2020 presidential election in the US, troll farms controlled all of the ten most important Christian American Facebook pages as well as six of the ten most important African American Facebook pages (Hao, 2021).

[3] Existing software allows a single user to manage multiple "sock puppet" accounts (The Guardian, 2011). Furthermore, in at least 20 countries, fully automated bots appear to be used to manipulate online opinions (Freedom House, 2017).

The rise of generative language models such as ChatGPT can make it particularly easy to flood social media platforms with content generated by fake accounts (Goldstein et al., 2023), potentially drowning out other messages.

In this paper, we analyse the impact of troll farms on voting outcomes. We develop a model in which a continuum of voters need to choose between two actions – for example, whether to vote for or against the government. There is a binary state of the world which indicates, for example, whether the government is competent. All voters share a common prior about the state, but they differ in their political preferences, or *types.* The type of each voter corresponds to the minimum probability that the voter needs to put on the high state to be willing to vote for the government. Each voter receives an imperfect continuous signal about the state. In addition, there is a sender, who wants the government to be reelected. She can organise a troll farm, which sends messages mimicking the informative signals. For each voter type, the sender chooses the number of trolls targeting that type – that is, the share of voters of a given type that receive messages from the troll farm instead of a genuine informative message – and the distribution of the message. Each voter receives exactly one message, and does not know whether it is an informative signal or comes from the troll farm. The aim of the sender is to maximise the expected share of voters voting for the government.

The optimal strategy of the sender takes into account the preferences of each type of voter. Some voters are willing to vote for the government when their belief equals the prior. The sender can persuade these voters to vote for the government by "flooding the zone" – increasing the number of trolls targeting these voters so much that the trolls completely displace informative signals, leaving these voters with no information.

On the other hand, voters who are unwilling to back the government in the absence of additional information cannot be persuaded in this manner. For these voters, the sender faces a tradeoff. On the one hand, increasing the mass of trolls targeting these voters, as well as shifting the distribution of trolls' messages towards more favourable signals, increases the likelihood that a given voter receives a signal that is favourable for the sender. On the other hand, this weakens the signals that these voters receive, making it harder to convince them to vote for the government. For such voters, we describe a unique optimal strategy of the sender – that is, a unique combination of the mass of trolls targeting each voter, and the distribution of the trolls' signals for each voter – that maximises the government's vote share in both states. This strategy is chosen in such a way as to ensure that the trolls only send messages from a certain interval, and within this interval any signal induces a posterior belief that leaves the voter indifferent between voting for and against the government.

We then derive three results related to the ability of the sender to manipulate electoral outcomes.

First, we show that the presence of a troll farm changes the role of the voters' independent signals. In the absence of trolls, making voters' independent signals more informative increases the share of voters voting for the government in the high state, and reduces it in the low state. Thus, increased signal precision improves the efficiency of the political outcome. However, the presence of trolls changes this picture: increased precision of voters' independent signals helps the sender manipulate the election, and increases the share of the voters who back the government in *both* states.

The reason why increased signal precision raises the government's vote share not only in the low state, but also in the high state has to do with the way troll farms emulate the independent signals. Increased signal precision means that in the low state, more voters receive a signal that induces them to vote against the government. It also means that such signals are more persuasive. But because the messages coming from the troll farm are indistinguishable from the genuine signals, this also makes the troll farm's messages more persuasive. Hence, the sender is able to increase the mass of trolls targeting each type of voter while also making sure that the messages retain sufficient persuasive power to induce these voters to vote for the government. Consequently, increased signal precision makes it easier to manipulate the electorate. This contrasts with the standard literature on persuasion, in which receivers observe where signals originate, and hence more precise independent signals make it harder for the sender to manipulate their beliefs (see Denter et al., 2021 on persuasion of voters; as well as Bergemann and Morris, 2016, and Matyskova, 2018).

Second, we analyse how increased polarisation of the electorate affects the power of the troll farm. Existing evidence suggests that polarisation of voters is increasing in a number of democracies.[4] There is considerable discussion of the potential negative effects of increased polarisation.[5] We show, however, that in the presence of a troll farm, polarisation can be beneficial. When the society becomes more ideologically polarised, the sender finds it harder to ensure that the government wins the election. In fact, we show that whenever the sender can ensure that the government wins the election in both states, an increase in polarisation can always move the society to an efficient equilibrium in which the government wins in the high state only.

Third, we discuss what happens when voters perceive information in a boundedly rational

[4] See Iyengar and Krupenkin (2018) for the United States, and Boxell et al. (2022) for several other OECD countries.

[5] See McCoy et al. (2018); Martherus et al. (2021).

way. There is considerable evidence that individuals can deviate from Bayesian updating. For example, individuals may exhibit probability weighting, overweighting or underweighting probabilities depending on their magnitude (Tversky and Kahneman, 1992). We show that a tendency of voters to undervalue signals relative to Bayesian updating, while preventing voters from forming correct beliefs, can nevertheless restore the efficiency of the political process by limiting the power of the troll farm.

Taken together, these results suggest that when troll farms exist, features of the political system and of the media environment that are usually considered beneficial can reduce the efficiency of the electoral process. In particular, societies in which voters are otherwise better informed – for example, due to a tradition of high-quality media – are in fact more vulnerable to manipulation by troll farms. At the same time, increased polarisation and biased information perception, often seen as problematic, can limit the power of troll farms to manipulate the political process.

At the same time, these results shed light on the effectiveness of various measures in overcoming disinformation. For example, should social media platforms modify their algorithms to promote messages coming from experts and other highly reliable sources? While at a first glance this may appear to be a useful approach, our results suggest that increasing precision of messages amplifies the power of troll farms which emulate them. Rather, effort should be made to promote messages which are moderately informative but not too informative.

Related literature. The paper contributes to the growing literature on disinformation on social media. A number of papers have documented the spread of misinformation online (see Zhuravskaya et al., 2020 for an overview, as well as Del Vicario et al., 2016; Allcott and Gentzkow, 2017; Vosoughi et al., 2018). In particular, prior research has documented extensive use of troll farms in the 2016 Brexit referendum and the 2016 US presidential election (Gorodnichenko et al., 2021); as well as in the online debate in China (King et al., 2017). Our paper complements this empirical literature by providing a theoretical framework for analysing the choice of strategy for the trolls, as well as for evaluating the impact of various features of the political process on the ability of troll farms to influence political outcomes.

Our model also contributes to the literature on information design. A number of papers have adapted the Bayesian persuasion (Kamenica and Gentzkow, 2011) to political contexts (see Alonso and Câmara, 2016; Wang, 2015; Bardhi and Guo, 2018; Ginzburg, 2019; Kolotilin et al., 2022; Sun et al., 2022; Mylovanov and Zapechelnyuk, 2023). Several other models of Bayesian persuasion examine private persuasion, under which, as in our paper, the sender can target different receivers with different persuasion schemes (Arieli and Babichenko, 2019;

Chan et al., 2019). The closest papers within this literature are those that study persuasion of voters who, as in our paper, also receive independent private signals in addition to the signal from the sender (Denter et al., 2021; Gradwohl et al., 2022; Heese and Lauermann, 2021) . In these settings, increased precision of voters' private signals constrains the sender's ability to persuade them. Our paper has two crucial differences from the prior literature on persuasion. First, in our model the sender does not have commitment power. Second, the signal of the sender does not complement voters' private signals, but replaces them in such a way that the voter is uncertain as to the source of the signal she receives. The fact that the sender's signals mimic voters' informative signals implies the result that higher precision of voters' signals *helps* the sender to achieve her desired outcome. We discuss the role of the key assumptions of our model in more detail in Section 6.6.

A number of papers have also looked at the effect of bounded rationality on voters' information processing and on voting outcomes. Levy and Razin (2015) show that, in the absence of a sender, correlation neglect, which causes voters to overvalue the signals they receive, can improve efficiency of the electoral outcome. In contrast, Denter et al. (2021) show that when a strategic sender is present, correlation neglect increases the persuasive power of the sender, and hence is harmful for information aggregation. Our paper analyses the role of bounded rationality when receivers cannot distinguish between the sender's message and genuine signals. We show that in this context, a tendency to *under*value signals – the opposite of what correlation neglect induces – makes electoral outcomes more efficient.

Finally, our paper is related to models of persuasion as signal distortion (Kartik et al., 2007; Edmond and Lu, 2021). In these models, a receiver obtains an informative signal about the state, and a sender can shift its realisation at a cost. In contrast, in our model the sender replaces the receiver's signal rather than distorting it.

2 Model

A continuum of voters of mass one need to choose whether to reelect the government. There is an unknown state of the world $\theta \in \{0, 1\}$, which indicates, for example, whether the government is competent. The preferences of each voter i are characterised by a type $x_i \in \mathbb{R}$. If voter i votes for the government, she receives a payoff $1 - x_i$ if the state turns out to be 1 – that is, if the government is competent – and a payoff $-x_i$ is the state turns out to be 0. The payoff of a voter who votes against the government is normalised to zero.[6] Thus, a voter with

[6] Thus, voters receive payoffs from their actions, and not from the outcome of the election. Since the set of voters is a continuum, each voter is pivotal with probability zero. Hence, allowing voters' payoffs to also

a higher type is more opposed to the government. We assume that x_i is distributed across voters according to a cdf H (with the associated density h) with full support on $\mathbb{R}$. Note that we allow for types to be outside the $[0, 1]$ interval, that is, for a voter to be a partisan – in other words, to prefer to vote the same way irrespective of the state. We assume that a voter who is indifferent votes for the government.

The government is reelected if the share of voters who vote for it is at least $\frac{1}{2}$. We will say that the election aggregates information if it achieves the "correct" outcome – that is, if the government is reelected in state one, and is not reelected in state zero.

Voters share a common prior belief about the state of the world being 1. We normalise that belief to $\frac{1}{2}$. Note that this is without loss of generality: changing the belief is equivalent to shifting the distribution H.

At the beginning of the game, each voter i receives a private signal s_i. These signals are independent across voters. In each state $\theta \in \{0, 1\}$, the signal of each voter is drawn from a cdf F_θ with density f_θ. We assume, without loss of generality, that $f_0(0) = f_1(0)$. Let $m(s) := \frac{f_1(s)}{f_0(s)}$ denote the likelihood ratio. We will assume that $m(s)$ is strictly increasing. This implies that a higher realisation of the signal makes a voter believe that the state equals one with a higher probability.

A political operator, whom we will call the sender, is trying to help the government. She can do it by setting up a troll farm, that is, by flooding the information environment with messages that imitate the informative signals but are not correlated with the true state. For each type of voter, the sender can select the intensity of trolls' messages, as well as the distribution of the trolls' messages. Formally, for each type of voter x, the sender chooses the probability α_x with which the voter observes a signal from trolls instead of an informative signal; and a distribution $\tilde{F}_x$ of the trolls' signals (with the associated density $\tilde{f}_x$). For example, $\alpha_x = 0$ means that no trolls are targeting voters with type x, and thus all signals observed by these voters are coming from informative sources. Similarly, $\alpha_x \to 1$ means that the number of trolls targeting voters with type x tends to infinity, and hence a signal is almost surely coming from a troll. Setting up any number of trolls is costless.

The timing of the game is as follows. First, for each x, the sender selects α_x and $\tilde{F}_x$. Then, nature draws the state θ and the signals. Each voter then receives either a signal from the troll farm or an informative signal, without being able to distinguish between the two. With probability α_x, a given voter with type x observes a signal from a troll drawn from the cdf $\tilde{F}_x$. With probability $1 - \alpha_x$ she observes a signal drawn from the cdf F_θ. Voters then

depend on the voting outcome has no effect on their behaviour at an equilibrium.

form their posterior beliefs and vote. Afterwards, payoffs are realised. The payoff of each voter is as described above, while the payoff of the sender is $u(V)$, where V is the share of voters that vote for the government, and u is a strictly increasing function.

3 Benchmark: No Trolls

Take a voter with type x. Let $\pi(s)$ be the probability that a voter assigns to the government being competent when she observes signal s. Her expected payoff is $\pi(s) - x$ if the she votes for the government, and zero otherwise. Hence, she votes for the government if and only if $\pi(s) \geq x$.

As a benchmark, consider the case when the troll farm is not operating. Then voters who observe signal s form a belief

$$\frac{f_1(s)}{f_1(s) + f_0(s)} = \frac{m(s)}{m(s)+1}.$$

Hence, a voter with type x votes for the government if and only if $x \leq \frac{m(s)}{m(s)+1}$. If $x \leq 0$, the voter votes for the government regardless of the signal, while if $x > 1$, she voters against the government regardless of the signal. If $x \in (0,1]$, the voter votes for the government if and only if $m(s) \geq \frac{x}{1-x}$, or, equivalently, if and only if

$$s \geq s^*(x)\,, \text{where } s^*(x) := m^{-1}\left(\frac{x}{1-x}\right). \tag{1}$$

Given the distribution of the signal in each state, the share of voters voting for the government in state $\theta \in \{0,1\}$ then equals

$$H(0) + \int_0^1 \left(1 - F_\theta\left[s^*(x)\right]\right) dH(x) = H(1) - \int_0^1 F_\theta\left[s^*(x)\right] dH(x)\,. \tag{2}$$

Then the election aggregates information if and only if

$$\int_0^1 F_1\left[s^*(x)\right] dH(x) \leq H(1) - \frac{1}{2} < \int_0^1 F_0\left[s^*(x)\right] dH(x)\,.$$

4 Equilibrium with Trolls

Suppose the sender has chosen α_x and $\tilde{F}_x$. A voter with type x who observes signal s forms the following posterior belief:

$$\pi(s) = \frac{(1-\alpha_x) f_1(s) + \alpha_x \tilde{f}_x(s)}{(1-\alpha_x) f_1(s) + \alpha_x \tilde{f}_x(s) + (1-\alpha_x) f_0(s) + \alpha_x \tilde{f}_x(s)} = \frac{1}{1 + \frac{(1-\alpha_x) f_0(s) + \alpha_x \tilde{f}_x(s)}{(1-\alpha_x) f_1(s) + \alpha_x \tilde{f}_x(s)}}.$$

She then votes for the government whenever $\pi(s) \geq x$.

The sender chooses α_x and $\tilde{F}_x$ for each voter to maximise the expected mass of votes that the government receives. As before, the actions of voters with types $x \notin [0,1]$ do not depend on their beliefs about the state. From now on, we will focus on voters with types $x \in [0,1]$, who can be persuaded to vote for the government.

Consider first a voter with type $x \leq \frac{1}{2}$. When that voter's belief equals the prior, she is willing to vote for the government. By setting $\alpha_x = 1$ the sender ensures that her posterior belief $\pi(s)$ equals the prior regardless of the state. Hence, this voter will vote for the government in either state with probability one. Intuitively, when the voter is ex ante willing to vote for the government, the sender can ensure that she does so ex post by "flooding the zone" – overloading the environment with trolls' messages to such an extent that they drown out informative messages and prevent the voter from learning any information.

Now consider a voter with type $x > \frac{1}{2}$. This voter is ex ante opposed to the government. Recall that without trolls, a voter with this type will vote for the government if and only if she receives signal $s \geq s^*(x)$. Introducing trolls weakens the signal. Since a signal $s < s^*(x)$ cannot persuade this voter to vote for the government even without trolls, it cannot make her willing to vote for the government with trolls either. It is then optimal for the sender to minimise the probability that these voters receive such a signal. Thus, the sender sets $\tilde{f}(s) = 0$ for all $s < s^*(x)$.

On the other hand, a signal $s \geq s^*(x)$ will, in each state, induce a posterior belief greater than x if there are no trolls, that is, if $\alpha_x = 0$. Introducing trolls – that is, increasing α_x – allows the sender to increase the probability that the voter receives a signal greater than $s^*(x)$. At the same time, increasing α_x weakens the signal, reducing the posterior belief that it induces. Hence, it is optimal for the sender to set α_x and $\tilde{f}(s)$ such that the posterior belief after observing signal $s \geq s^*(x)$ equals exactly x. This requirement, together with the requirement that $\tilde{f}(s)$ remains a density implies that there is a unique optimal pair $\left(\alpha_x, \tilde{F}_x\right)$ that maximises the vote share of the government in both states. This optimal strategy is presented in the following result, the proof of which is based on the intuition described above:

Lemma 1. *For $x \leq \frac{1}{2}$ the sender selects $\alpha_x = 1$, with any $\tilde{f}_x$. For $x > \frac{1}{2}$, the sender selects*

$$\alpha_x = \frac{\int_{s^*(x)}^{+\infty} \left[(1-x) f_1(s) - x f_0(s)\right] ds}{2x - 1 + \int_{s^*(x)}^{+\infty} \left[(1-x) f_1(s) - x f_0(s)\right] ds}$$

and

$$\tilde{f}_x(s) = \begin{cases} 0 & \text{if } s < s^*(x) \\ \frac{(1-x) f_1(s) - x f_0(s)}{\int_{s^*(x)}^{+\infty} [(1-x) f_1(s) - x f_0(s)] ds} & \text{if } s \geq s^*(x). \end{cases}$$

This strategy enables the sender to ensure that in each state, all voters with $x \leq \frac{1}{2}$ vote for the government. At the same time, a voter with $x > \frac{1}{2}$ votes for the government when she receives a signal from a troll, or when she receives an informative signal that happens to be greater than $s^*(x)$. The former event happens with the same probability in both states, while the latter is more likely to happen in state 1 – therefore, the share of voters voting for the government is greater in state 1. Using this logic, we can find the government's vote share in each state when the sender is following her optimal strategy. These are defined as follows:

Lemma 2. *The share of voters voting for the government in state $\theta = 0$ equals*

$$V_0 = H\left(\frac{1}{2}\right) + \int_{\frac{1}{2}}^{1} \frac{F_0[s^*(x)] - F_1[s^*(x)]}{\frac{x}{1-x} F_0[s^*(x)] - F_1[s^*(x)]} dH(x),$$

while in state $\theta = 1$ it equals

$$V_1 = H\left(\frac{1}{2}\right) + \int_{\frac{1}{2}}^{1} \frac{F_0[s^*(x)] - F_1[s^*(x)]}{F_0[s^*(x)] - \frac{1-x}{x} F_1[s^*(x)]} dH(x);$$

furthermore, $V_0 < V_1$.

The election will aggregate information whenever the government wins in state 1 but not in state 0. This happens if and only if $V_0 < \frac{1}{2} \leq V_1$. On the other hand, if $V_0 \geq \frac{1}{2}$, the sender is able to ensure the government's victory in both states.

5 Polarisation, Informativeness, and Voting Outcomes

In this section we analyse the effects of the features of the voting process on election outcomes.

We will start by looking at the effect of polarisation. Polarisation is described by the shape of the distribution H of voters' types. We will define it as follows:

Definition 1. A distribution $\hat{H}$ *admits greater polarisation* than distribution H if and only if:

- $\hat{H}(x) \geq H(x)$ for all $x \leq \frac{1}{2}$; and
- $\hat{H}(x) \leq H(x)$ for all $x \geq \frac{1}{2}$.

This defines a partial order on the set of distributions under which greater polarisation means more mass away from $\frac{1}{2}$. To see the intuition behind this definition, recall that a voter whose type x equals $\frac{1}{2}$ is indifferent between supporting and opposing the government at the prior belief. More generally, voters whose types are close to $\frac{1}{2}$ can be convinced to shift their vote by a relatively weak signal, while voters whose types are far from $\frac{1}{2}$ need a strong signal to be convinced to change her vote. We can then say that for any two voters who are on the same side of the political divide (that is, whose types are on the same side of $\frac{1}{2}$), the voter whose type is closer to $\frac{1}{2}$ is more moderate or more centrist, while the voter whose type is further from $\frac{1}{2}$ is more extreme. Definition 1 then says that polarisation is higher if, for any given type x of voter, there are more voters who are more extreme than x.

If $H\left(\frac{1}{2}\right) \geq \frac{1}{2}$, Lemma 2 implies that the government always wins the election in both states. A change in polarisation does not affect this. In the more interesting case when $H\left(\frac{1}{2}\right) < \frac{1}{2}$, polarisation has an effect. The next result shows that greater polarisation can limit the power of troll farms to manipulate elections, and restore information aggregation by preventing the sender from achieving government victory in state 0.

Proposition 1. *Suppose $H\left(\frac{1}{2}\right) < \frac{1}{2}$. Suppose further that the government wins the election in both states under some distribution $H(x)$. There exists a distribution $\hat{H}(x)$ that admits greater polarisation than $H(x)$ and under which the election aggregates information.*

Intuitively, increased polarisation means that for each voter on either side of $\frac{1}{2}$, there are more voters that are further away from $\frac{1}{2}$ – that is, from the most moderate voter. This means that there are more voters who very supportive of the government, as well as more voters who are very opposed to it. However, these two changes affect the power of the sender in different ways. As Lemma 1 shows, voters who are ex ante supportive of the government – that is, voters with types $x < \frac{1}{2}$ – can always be persuaded to vote for the government. Hence, an increase in the mass of extreme pro-government voters compared to the mass of moderate pro-government voters does not change the ability of the troll farm to manipulate elections. On the other hand, a shift of anti-government voters towards the more extreme positions makes it harder for the sender to achieve the outcome that it is aiming at.

Consider now the effect of information precision. Information precision depends on the shapes of signal distributions F_0 and F_1. We will refer to the pair (F_0, F_1) as *information structure*. Recall that in the absence of trolls, in state $\theta \in \{0,1\}$ a voter with type x votes for the government with probability $1 - F_\theta\left[s^*(x)\right]$. A change in the information structure changes both the distributions F_0 and F_1 and the cutoff $s^*(x)$, as the latter depends on the likelihood ratio. Intuitively, we can say that signals are more informative for a voter with type x if she is more likely to make the correct decision – that is, vote for the government in state 1, and against the government in state 0. We can define a more informative information structure as one in which every voter is more likely to make the correct decision in both states. Formally, we define informativeness as follows:

Definition 2. Information structure $\left(\hat{F}_0, \hat{F}_1\right)$ is more informative than information structure (F_0, F_1) if and only if for all $x \in [0,1]$ we have:

- $\hat{F}_0\left[\hat{s}^*(x)\right] \geq F_0\left[s^*(x)\right]$, and
- $\hat{F}_1\left[\hat{s}^*(x)\right] \leq F_1\left[s^*(x)\right]$,

where $s^*(x)$ and $\hat{s}^*(x)$ are defined as in (1).

Recall that a voter receives a payoff of $-x$ if she votes for the government in state 0, and a payoff of x if she votes for the government in state 1. Then her expected utility in state 0 equals $-x\left(1 - F_0\left[s^*(x)\right]\right)$, while in state 1 it equals $(1-x)\left(1 - F_1\left[s^*(x)\right]\right)$. Then under our definition, an information structure is more informative if and only if the utility of each voter in each state is higher. Thus, an information structure that is more informative under our definition is also more informative under the Blackwell information criterion.

From (2) it is easy to see that without interference from the sender, under a more informative information structure, the government receives more votes in state 1, and fewer votes in state 0. With trolls, however, the dynamic is different, as the next result shows:

Proposition 2. *If information structure* $\left(\hat{F}_0, \hat{F}_1\right)$ *is more informative than information structure* (F_0, F_1)*, then in each state more voters vote for the government under* $\left(\hat{F}_0, \hat{F}_1\right)$ *than under* (F_0, F_1)*.*

Hence, a change in the information structure which, on its own, increases the probability that a voter makes the correct decision in each state also helps the sender manipulate the voters' decisions. In other words, while in the absence of the sender greater informativeness increases the share of votes that the government receives in state 1 and reduces that share in state 0, in presence of the sender greater informativeness helps the government in both states.

Intuitively, for voters increased informativeness has two effects. First, for those voters who are not exposed to trolls, increased informativeness raises the probability that a voter votes in the correct way, which makes it harder for the sender to manipulate the voting outcome. Second, for voters who do receive a message from trolls, increased informativeness makes signals more believable, making it easier for the sender to persuade voter. However, increased informativeness also allows the sender to increase the mass of trolls targeting each voter while keeping the signal sufficiently convincing. Hence, the sender is able to partially negate the first effect of informativeness, and so the overall share of voters voting for the government increases.

Now we can look at the effect of information precision on the ability of the election to achieve the optimal outcome. Higher informativeness implies that for all x, $\frac{F_0[s^*(x)]}{F_1[s^*(x)]}$ is closer to infinity, while lower informativeness implies that it is closer to one.[7] One can define sequences of information structures such that along the sequence, the ratio $\frac{F_0[s^*(x)]}{F_1[s^*(x)]}$ increases from one to infinity. As we move along this sequence and informativeness increases, what happens to voting outcomes?

As before, if $H\left(\frac{1}{2}\right) \geq \frac{1}{2}$, then by Lemma 2 the government always wins the election in both states. When $H(1) < \frac{1}{2}$, the government always loses the election because the share of voters with type $x > 1$ who always vote against the government is greater than $\frac{1}{2}$. Consider now the more interesting case when $H\left(\frac{1}{2}\right) < \frac{1}{2}$ and $H(1) \geq \frac{1}{2}$. Then, as information structure moves from being completely uninformative towards being perfectly informative, we have the following result:

Proposition 3. *Suppose that $H\left(\frac{1}{2}\right) < \frac{1}{2}$. Take a sequence of information structures $(F_0, F_1)_r$ indexed by $r \in (-\infty, +\infty)$, such that informativeness increases with r, $\lim_{r\to-\infty} \frac{F_0(s)}{F_1(s)} = 1$ for all s, and $\lim_{r\to+\infty} \frac{F_0(s)}{F_1(s)} = +\infty$ for all s. Then there exist r', r'' with $r' < r''$, such that:*

- *for $r < r'$, the government loses the election in both states;*
- *for $r \in (r', r'')$, the government wins the election in state 1 only, and the election aggregates information;*

[7] Note that by our monotone likelihood ratio property, $F_0(s) > F_1(s)$ for all s.

- *for $r > r''$, the government wins the election in both states whenever the distribution of types admits sufficiently high polarisation; and wins the election in state 1 only otherwise.*

In words, when signals are not very informative, they cannot induce sufficiently many voters to move their belief sufficiently far from the prior, so the government cannot get enough votes even in state 1. If signals are moderately informative, the government can receive enough votes to win the election in state 1 but not in state 0. A more interesting case emerges when signals are very informative. If polarisation is very high, Proposition 1 implies that the sender cannot manipulate the election to ensure government victory in both states. However, when polarisation is not too high, the government wins the election irrespective of the state – thus, information aggregation is prevented.

Intuitively, increased informativeness raises the share of voters voting for the government in both states. Thus, completely uninformative signals mean that the government cannot win in either states, while if signals are very informative, Propositions 1 and 2 imply that the government wins the election in both states unless polarisation is high. At the same time, because the share of voters who vote for the government is lower in state 0 than in state 1, there is an intermediate range of information structures under which the government wins the election in state 1 only.

Proposition 3 means that when polarisation is relatively low, an increase in informativeness can move the voting outcome away from information aggregation. Hence, increased informativeness, beyond a certain level, is harmful.

6 Discussion

6.1 Non-Bayesian Information Perception

Substantial theoretical and empirical research has pointed to the fact that individuals may systematically deviate from Bayesian updating. In this section, we analyse how such cognitive limitations on part of the voters affect information aggregation and the ability of the sender to manipulate election outcomes.

Deviations from Bayesian rationality can take different forms. Under probability weighting (Tversky and Kahneman, 1992), individuals may overweigh the probabilities that are higher than some reference point, and underweigh probabilities that are below it. This might mean that, after receiving a signal, voters place their posteriors closer to the prior than

Bayesian updating would suggest.[8]. In terms of our model, this would mean that voters perceive signal realisation s to be closer to zero than it actually is.

An opposite phenomenon can be caused by, for example, correlation neglect.[9] Suppose that each voter observes not one, but two identical signal realisations s, which are perfectly correlated. If voters do not realise that these signals are correlated, and treat them as independent, they would exaggerate their value, forming more extreme posterior beliefs. In terms of our model, this would mean that perceived signal realisation is further from zero than s.

To account for such phenomena, suppose that upon observing signal realisation s, a voter forms her posterior belief as if she observed signal realisation $\beta(s)$, where $\beta(\cdot)$ is a function that distorts the voter's belief according to a particular type of belief updating. We will assume that the belief distortion function β is strictly increasing, and that $\beta(0) = 0$ (thus, not receiving an informative signal does not lead the voter to change her belief). Phenomena such as probability weighting would correspond to the case when $\beta(s) > s$ for $s < 0$, and $\beta(s) < s$ for $s > 0$. Under such types of β, the posterior belief for a given s is more conservative – that is, closer to the prior – than the Bayesian posterior belief. On the other hand, correlation neglect would imply the opposite setting, in which $\beta(s) < s$ for $s < 0$, and $\beta(s) > s$ for $s > 0$ – this kind of β functions induce less conservative beliefs than the Bayesian updating does.

More generally, we can introduce a partial order on belief distortion functions β in terms of how conservative the resulting posterior beliefs are. We define it as follows:

Definition 3. A belief distortion function $\hat{\beta}$ is *more conservative* than belief distortion function β if and only if:

- $\hat{\beta}(s) > \beta(s)$ for all $s < 0$; and
- $\hat{\beta}(s) < \beta(s)$ for all $s > 0$

We can then show that deviations from Bayesian updating can improve information aggregation. More generally, the next result shows that whenever the sender can prevent information aggregation by ensuring government victory in both states, there exists a more conservative belief distortion function under which information aggregation is restored:

Proposition 4. *Suppose $H\left(\frac{1}{2}\right) < \frac{1}{2}$. Suppose further that the government wins the election in both states under some belief distortion function $\beta(s)$. There exists a belief distortion*

[8] See also Enke and Graeber (2019), which produces similar comparative statics in a Bayesian framework.

[9] See Enke and Zimmermann (2019) for empirical evidence of correlation neglect.

function $\hat{\beta}(s)$ that is more conservative than $\beta(s)$ and under which the election aggregates information.

The intuition is similar to that of Proposition 1: a more conservative distortion function moves the posterior belief of a given voter towards the prior. At the level of an individual voter, this has the same effect as a decrease in polarisation at the aggregate level.

6.2 Naive Voters

Another possibility is that voters are strategically naive, and do not understand the incentives of the sender. This would mean that they are unaware of existence of trolls, and update their beliefs thinking that all signals are informative. Experimental literature has shown evidence of strategic naivete in communication (Cai and Wang, 2006; Jin et al., 2021), and in voting interactions (Patty and Weber, 2007; Ginzburg and Guerra, 2019).

Formally, suppose that a fraction $\phi \in [0,1]$ of voters update their beliefs assuming that $\alpha = 0$. When $\phi = 0$, the setting is equivalent to our baseline model. If these voters can be targetted – that is, if for every type x of the voter the sender can condition her choice of α_x and $\tilde{F}_x$ on whether the voter is naive – then any naive voter can be persuaded to vote for the government by setting $\alpha_x \to 1$ and choosing $\tilde{F}_x$ in such a way that $\tilde{F}[s^*(x)] = 0$. Then every naive voter will, with probability one, receive a signal that induces her to vote for the government. Hence, all naive voters vote for the government, while for the remaining voters, the sender chooses a strategy equivalent to the one described in Section 4. Hence, adding naive voters is isomorphic to increasing $H(0)$; thus, the results of Section 5 remain unchanged.

6.3 Anti-Government Trolls

The government may be not the only side that has trolls on its side. Suppose that the opposing side can use the same tactic. In terms of our model, suppose that there are two senders, one of which is trying to ensure that the government wins, and the other is trying to achieve the opposite outcome. Each sender can set up a troll farm that would, as before, send messages uncorrelated with the state. The setting then becomes a zero-sum game between the two senders.

AS in the baseline model, the pro-government sender can persuade any voter with $x \leq \frac{1}{2}$ – that is, any voter who is in favour of the government at her prior belief – to vote for the government by setting $\alpha_x \to 1$, which eliminates the informative signal for this voter. On the other hand, by similar logic the anti-government sender can persuade any voter with $x > \frac{1}{2}$ to vote against the government by also setting $\alpha_x \to 1$. Hence, in this setting, the government wins the election if and only if the mass of senders with $x \leq \frac{1}{2}$ is greater than $\frac{1}{2}$. Because the outcome of the election does not depend on the state, the election does not aggregate information.

6.4 Limited Reach

6.5 Sender Observes State

Our baseline model assumed that the sender cannot observe the state of the world, and hence chooses the same action in each state. Suppose instead that she is informed about the state. How does this affect the equilibria? In this case she can condition the number of trolls and the distribution of trolls' messages – that is, her choice of $\left(\alpha_x, \tilde{F}_x\right)$ – on it. But then $\left(\alpha_x, \tilde{F}_x\right)$ serves as a signal about the state, and voters can update their beliefs after observing $\left(\alpha_x, \tilde{F}_x\right)$.

Since trolls are costless, the setting becomes that of a cheap talk game. As usual, there exists a babbling equilibrium, in which the sender selects the same $\left(\alpha_x, \tilde{F}_x\right)$ in each state. This equilibrium is equivalent to the setup of the baseline model, and hence the equilibrium outcomes are the same as the ones derived in Section 4.

On the other hand, a separating equilibrium, in which the sender with positive probability chooses different values of $\left(\alpha_x, \tilde{F}_x\right)$ in different states, cannot exist. The reason is that if the sender's choice reveals some information about the state, the government will receive more votes in state 0 than in state 1. Then in state 0 the sender will prefer to deviate to the same choice of $\left(\alpha_x, \tilde{F}_x\right)$ that she makes in state 1.

Hence, the pooling equilibrium, in which the sender does not condition her choice of $\left(\alpha_x, \tilde{F}_x\right)$ on the state, is the only equilibrium that can exist. Therefore, allowing the sender to observe the state does not change the outcome of the game.

6.6 Commitment, Source Uncertainty, and Comparison with Information Design

In this section we discuss the relation between our model and the broader literature on information design.

Our approach is a particular kind of information design in which the sender is constrained in several significant ways. First, she does not have commitment power, and hence cannot select an arbitrary mapping from states to signal distributions. Second, the sender cannot send signals of her own, but she can replace the receivers' independent signals. At the same time, unlike in standard Bayesian persuasion models, receivers are uncertain as to the source from which signals originate – this uncertainty about the source implies that precision of the voters' independent signals is a key factor that determines the sender's ability to manipulate the voters' beliefs.

Suppose that, in addition to source uncertainty, the sender also had commitment power, as in standard models of Bayesian persuasion. Because the sender can fully replace the independent signals with messages of her own, she would then be able to induce any posterior in each state. Hence, despite source uncertainty, the model would become a standard model of Bayesian persuasion.

On the other hand, suppose that the sender had neither commitment power nor was able to mimic and replace voters' private signals – thus, the sender's messages would complement voters' private signals, as in standard models of signalling games. If the sender, as in the baseline model, is unable to observe the state, then at the equilibrium voters would ignore her messages, and so the sender would have no commitment power. On the other hand, if the sender could observe the state, then, as explained in Section 6.5, the game would be a form of a cheap talk game, in which no separating equilibrium would exist, and hence the sender would also have no ability to persuade the voters.

Thus, both of the key differences – lack of commitment power, and source uncertainty – from the more standard information design framework are crucial for the mechanics of the model.

Mathematical Appendix

Proof of Lemma 1

Take a voter with type $x \in [0,1]$. Suppose that the sender has chosen α_x and $\tilde{f}_x$. Then the voter's belief after observing signal s equals $\pi(s) = \frac{1}{1+\frac{(1-\alpha_x)f_0(s)+\alpha_x\tilde{f}_x(s)}{(1-\alpha_x)f_1(s)+\alpha_x\tilde{f}_x(s)}}$.

If $x \leq \frac{1}{2}$, then for any $\tilde{f}_x$ the sender can ensure that $\pi(s) = \frac{1}{2} \geq x$ by setting $\alpha_x = 1$. Hence, this is part of an optimal strategy.

Now take a voter with type $x \in \left(\frac{1}{2}, 1\right]$. Given the sender's choice of α_x and $\tilde{f}_x$, let $A(x) \subseteq \mathbb{R}$ be the set of signals such that this voter votes for the government if and only if she receives a signal $s \in A(x)$. That is, $A(x)$ is a set of signals such that $\pi(s) \geq x$ if and only if $s \in A(x)$. Then the probability that the voter votes for the government in state $\theta \in \{0,1\}$ equals the probability that she observes a signal $s \in A(x)$. This probability equals

$$p_\theta(x) = \int_{s \in A(x)} \left[(1-\alpha_x) f_\theta(s) + \alpha_x \tilde{f}_x(s)\right] ds.$$

The sender aims to maximise $p_\theta(x)$, that is, the probability that the voter observes a signal $s \in A(x)$.

Recall that without trolls, after observing signal s she votes for the government if and only if $s \geq s^*(x) = m^{-1}\left(\frac{x}{1-x}\right)$. If $x > \frac{1}{2}$, then $s^*(x) > m^{-1}(1) = 0$. Note that $\pi(s)$ is decreasing in α_x and in $\tilde{f}_x(s)$ if and only if $f_1(s) > f_0(s)$, that is, if and only if $s > 0$.

Consider any signal s. If $s < 0$, then $f_0(s) > f_1(s)$, and hence $\pi(s) < \frac{1}{2} < x$ regardless of α_x and $\tilde{f}_x$. Hence, $s \notin A(x)$. On the other hand, if $s \in [0, s^*(x))$, then without trolls (that is, when $\alpha_x = \tilde{f}_x = 0$), the voter votes against the government. Since $\pi(s)$ is decreasing in α_x and in $\tilde{f}_x(s)$, the sender votes against the government for any α_x and $\tilde{f}_x$. Hence, $s \notin A(x)$ as well. We can conclude that $s \notin A(x)$ for any $s < s^*(x)$.

On the other hand, for $s \geq s^*(x)$, then without trolls, $\pi(s) \geq x$, so the voter is willing to vote for the government. At the same time, $\pi(s)$ is decreasing in α_x and in $\tilde{f}_x(s)$. We can show that it is optimal for the sender to set $\pi(s) = x$ for all $s \geq s^*(x)$. To see this, take a set $B \subset [s^*(x), \infty)$, and suppose that $\pi(s) > x$ for all $s \in B$. Then the sender can increase $\tilde{f}_x(s)$ for all $s \in B$ in such a way that $\pi(s) \geq x$ still holds. This would increase $p_\theta(x)$, implying that the original choice is not optimal.

Hence, at the optimum, for all $s \geq s^*(x)$, $\tilde{f}_x(s)$ is such that

$$\begin{aligned} \frac{1}{1 + \frac{(1-\alpha_x)f_0(s) + \alpha_x \tilde{f}_x(s)}{(1-\alpha_x)f_1(s) + \alpha_x \tilde{f}_x(s)}} &= x \\ \Leftrightarrow \tilde{f}_x(s) &= \frac{1-\alpha_x}{\alpha_x} \frac{(1-x) f_1(s) - x f_0(s)}{2x-1}. \end{aligned} \tag{3}$$

Note that $\tilde{f}_x(s)$ as defined above is positive. To see this, observe that at $s = s^*(x)$ we have $m(s) = \frac{f_1(s)}{f_0(s)} = \frac{x}{1-x}$, and hence $(1-x) f_1(s) - x f_0(s) = 0$. As $m(s)$ is increasing in s, we have $\frac{f_1(s)}{f_0(s)} > \frac{x}{1-x}$ for all $s > s^*(x)$, and hence $(1-x) f_1(s) - x f_0(s) > 0$.

Then the probability that the voter votes for the government in state θ equals

$$p_\theta(x) = (1-\alpha_x) \int_{s \in A(x)} \left[f_\theta(s) + \frac{(1-x) f_1(s) - x f_0(s)}{2x-1}\right] ds,$$

which is decreasing in α_x. Hence, it is optimal to select the smallest possible α_x under which $\tilde{f}_x(s)$ is a density. Therefore, the sender sets α_x such that $\tilde{f}_x(s) = 0$ for all $s < s^*(x)$, and $\int_{s^*(x)}^{+\infty} \tilde{f}_x(s)\, ds = 1$. Consequently, α_x is given by

$$\int_{s^*(x)}^{+\infty} \frac{1-\alpha_x}{\alpha_x} \frac{(1-x) f_1(s) - x f_0(s)}{2x-1} ds = 1$$
$$\Leftrightarrow \alpha_x = \frac{\int_{s^*(x)}^{+\infty} [(1-x) f_1(s) - x f_0(s)]\, ds}{2x - 1 + \int_{s^*(x)}^{+\infty} [(1-x) f_1(s) - x f_0(s)]\, ds}.$$

Substituting this into (3) yields

$$\tilde{f}_x(s) = \frac{(1-x) f_1(s) - x f_0(s)}{\int_{s^*(x)}^{+\infty} [(1-x) f_1(s) - x f_0(s)]\, ds} \text{ for all } s \geq s^*(x),$$

which together with the fact that $\tilde{f}_x(s) = 0$ for all $s < s^*(x)$ implies the result. □

Proof of Lemma 2

Under the strategy described in Lemma 1, each voter with type $x \leq \frac{1}{2}$ votes for the government with probability one. The mass of such voters is $H\left(\frac{1}{2}\right)$. A voter with type $x \in \left(\frac{1}{2}, 1\right]$ votes for the government if and only if she receives a signal $s \geq s^*(x)$. In state $\theta \in \{0, 1\}$, the probability of this is

$$\int_{s^*(x)}^{+\infty} \left[(1-\alpha_x) f_\theta(s) + \alpha_x \tilde{f}_x(s)\right] ds$$
$$= (1-\alpha_x)(1 - F_\theta[s^*(x)]) + \alpha_x \frac{(1-x)(1 - F_1[s^*(x)]) - x(1 - F_0[s^*(x)])}{\int_{s^*(x)}^{+\infty} [(1-x) f_1(s) - x f_0(s)]\, ds}$$
$$= \frac{x F_0[s^*(x)] - (1-x) F_1[s^*(x)] - (2x-1) F_\theta[s^*(x)]}{x F_0[s^*(x)] - (1-x) F_1[s^*(x)]}$$

Hence, the overall vote share of the government in state $\theta \in \{0, 1\}$ equals $H\left(\frac{1}{2}\right) + \int_{\frac{1}{2}}^{1} \frac{x F_0[s^*(x)] - (1-x) F_1[s^*(x)] - (2x-1) F_\theta[s^*(x)]}{x F_0[s^*(x)] - (1-x) F_1[s^*(x)]} dH(x)$. Substituting $\theta \in \{0, 1\}$ yields the expressions for V_0 and V_1.

Furthermore, for $x > \frac{1}{2}$, we have $\frac{x}{1-x} F_0[s^*(x)] - F_1[s^*(x)] > F_0[s^*(x)] - \frac{1-x}{x} F_1[s^*(x)]$, which follows from the fact that $F_0(s) > F_1(s)$, because monotone likelihood ratio property implies first-order stochastic dominance. As a consequence, $V_1 > V_0$. □

Proof of Proposition 1

For a given $H(x)$, take a family of distributions indexed by $r \in (0,1)$, of the form

$$H_r(x) := H\left[(1-r)x + r\frac{1}{2}\right].$$

It is easy to see that for each $r \in (0,1)$, the function $H_r(x)$ is a cdf, as it is increasing in x, with $\lim_{x\to-\infty} H_r(x) = 0$ and $\lim_{x\to+\infty} H_r(x) = 1$. Furthermore, $H_r(x) = H(x)$ for $r = 0$, and higher r means greater polarisation. Note also that $dH_r(x) = (1-r)h\left[(1-r)x + r\frac{1}{2}\right]dx$.

Using Lemma 2, we have for each r

$$V_0 = H\left(\frac{1}{2}\right) + \int_{\frac{1}{2}}^{1} \frac{F_0[s^*(x)] - F_1[s^*(x)]}{\frac{x}{1-x}F_0[s^*(x)] - F_1[s^*(x)]}(1-r)h\left[(1-r)x + r\frac{1}{2}\right]dx,$$

Hence,

$$\lim_{r\to1} V_0 = H\left(\frac{1}{2}\right) < \frac{1}{2}.$$

Thus, when r is sufficiently large, the government does not win the election in state 0. Because $V_1 > V_0$ for all r, and since $H_r(x)$ is continuous in r, there exists a value of $r \in (0,1)$ for which $V_0 < \frac{1}{2} < V_1$, that is, for which the election aggregates information. □

Proof of Proposition 2

Rewriting the expressions in Lemma 2, we obtain:

$$V_0 = H\left(\frac{1}{2}\right) + \int_{\frac{1}{2}}^{1} \frac{\frac{F_0[s^*(x)]}{F_1[s^*(x)]} - 1}{\frac{x}{1-x}\frac{F_0[s^*(x)]}{F_1[s^*(x)]} - 1}dH(x),$$

and

$$V_1 = H\left(\frac{1}{2}\right) + \int_{\frac{1}{2}}^{1} \frac{\frac{F_0[s^*(x)]}{F_1[s^*(x)]} - 1}{\frac{F_0[s^*(x)]}{F_1[s^*(x)]} - \frac{1-x}{x}}dH(x).$$

An increase in informativeness increases $F_0[s^*(x)]$ and reduces $F_1[s^*(x)]$ for all $x \in [0,1]$. Hence, $\frac{F_0[s^*(x)]}{F_1[s^*(x)]}$ increases for all $x \in \left[\frac{1}{2}, 1\right]$. Therefore, both V_0 and V_1 increase. □

Proof of Proposition 3

Rewriting the expressions in Lemma 2, we obtain:

$$V_0 = H\left(\frac{1}{2}\right) + \int_{\frac{1}{2}}^{1} \frac{\frac{F_0[s^*(x)]}{F_1[s^*(x)]} - 1}{\frac{x}{1-x}\frac{F_0[s^*(x)]}{F_1[s^*(x)]} - 1} dH(x),$$

and

$$V_1 = H\left(\frac{1}{2}\right) + \int_{\frac{1}{2}}^{1} \frac{\frac{F_0[s^*(x)]}{F_1[s^*(x)]} - 1}{\frac{F_0[s^*(x)]}{F_1[s^*(x)]} - \frac{1-x}{x}} dH(x).$$

Suppose $H\left(\frac{1}{2}\right) < \frac{1}{2}$. Take a sequence $(F_0, F_1)_r$ such that informativeness increases with r, $\lim_{r\to-\infty} \frac{F_0(s)}{F_1(s)} \to 1$ for all s, and $\lim_{r\to-\infty} \frac{F_0(s)}{F_1(s)} \to \infty$ for all s. Then we have

$$\lim_{r\to-\infty} V_1 = H\left(\frac{1}{2}\right) < \frac{1}{2},$$

hence the government loses the election in state 1 when r is sufficiently low. Furthermore,

$$\lim_{r\to\infty} V_1 = H\left(\frac{1}{2}\right) + \int_{\frac{1}{2}}^{1} dH(x) = H(1) \geq \frac{1}{2},$$

hence the government wins the election in state 1 when r is sufficiently high. Since V_1 is increasing with r by Proposition 2, there exists r' such that the government loses the election in state 1 for $r < r'$, and wins the election in state 1 for $r > r'$.

At the same time,

$$\lim_{r\to-\infty} V_0 = H\left(\frac{1}{2}\right) < \frac{1}{2},$$

hence the government loses the election in state 0 when r is sufficiently low. Furthermore,

$$\begin{aligned}\lim_{r\to\infty} V_0 =& H\left(\frac{1}{2}\right) + \int_{\frac{1}{2}}^{1} \frac{1-x}{x} dH(x) \\ =& H\left(\frac{1}{2}\right) + \frac{1-x}{x} H(x)\Big|_{\frac{1}{2}}^{1} + \int_{\frac{1}{2}}^{1} \frac{1}{x^2} H(x)\, dx \\ =& \int_{\frac{1}{2}}^{1} \frac{1}{x^2} H(x)\, dx.\end{aligned}$$

When polarisation is high enough, then $H(x)$ is close to $H\left(\frac{1}{2}\right)$ for almost all $x \in \left(\frac{1}{2}, 1\right)$. Hence, $\lim_{r\to\infty} V_0$ is close to $H\left(\frac{1}{2}\right) \int_{\frac{1}{2}}^{1} \frac{1}{x^2} dx = H\left(\frac{1}{2}\right) < \frac{1}{2}$. Therefore, the government loses the election in state 0 even when r is high. On the other hand, when polarisation is low enough, then $H(x)$ is close to 1 for almost all $x \in \left(\frac{1}{2}, 1\right)$. Hence, $\lim_{r\to\infty} V_0$ is close to $\int_{\frac{1}{2}}^{1} \frac{1}{x^2} dx = 1 > \frac{1}{2}$. Therefore, the government wins the election in state 0 when r is sufficiently high. Since V_0 is increasing with r by Proposition 2, when polarisation is sufficiently high, there exists r'' such that the government loses the election in state 0 for $r < r''$, and wins the election in state 0 for $r > r''$.

Finally, the fact that $V_0 < V_1$ implies that $r' < r''$. □

Proof of Proposition 4

In the absence of trolls, a voter, upon observing signal s updates her belief in a manner similar to the one described in Section 3, except that she perceives the signal to equal $\beta(s)$. She thus votes for the government if and only if $\beta(s) \geq s^*(x)$, where, as before, $s^*(x) := m^{-1}\left(\frac{x}{1-x}\right)$. The sender, as before, selects α_x and $\tilde{F}_x$ to make sure that a voter's posterior belief equals 0 for all $s < s^*(x)$, and equals x for all $s > s^*(x)$. Using the same steps as in Lemma 1and Lemma 2, with s replaced by $\beta(s)$, we can show that at the equilibrium, the share of voters voting for the government in state $\theta = 0$ equals

$$V_0 = H\left(\frac{1}{2}\right) + \int_{\frac{1}{2}}^{1} \frac{F_0\left(\beta^{-1}\left[s^*(x)\right]\right) - F_1\left(\beta^{-1}\left[s^*(x)\right]\right)}{\frac{x}{1-x}F_0\left(\beta^{-1}\left[s^*(x)\right]\right) - F_1\left(\beta^{-1}\left[s^*(x)\right]\right)} dH(x),$$

while in state $\theta = 1$ it equals

$$V_1 = H\left(\frac{1}{2}\right) + \int_{\frac{1}{2}}^{1} \frac{F_0\left(\beta^{-1}\left[s^*(x)\right]\right) - F_1\left(\beta^{-1}\left[s^*(x)\right]\right)}{F_0\left(\beta^{-1}\left[s^*(x)\right]\right) - \frac{1-x}{x}F_1\left(\beta^{-1}\left[s^*(x)\right]\right)} dH(x);$$

and that, furthermore, $V_0 < V_1$.

Suppose that when voters form their beliefs according to some distortion function $\beta(s)$, the government wins the election in both states, that is, $V_1 > V_0 \geq \frac{1}{2}$. To prove the proposition, we need to show that there exists a more conservative distortion function under which the election aggregates information.

Take a sequence of belief distortion functions indexed by $r = 1, 2, ...$, of the form

$$\beta(s) = \frac{\beta_r(s)}{r}.$$

Note that $\beta_r(s)$ is more conservative whenever r is higher. Furthermore, $\lim_{r\to\infty} \beta_r(s) = 0$ for all s. Since $\beta(0) = 0$, we have

$$\lim_{r\to\infty} V_0 = H\left(\frac{1}{2}\right) + \lim_{r\to\infty} \int_{\frac{1}{2}}^{1} \frac{F_0(0) - F_1(0)}{\frac{x}{1-x}F_0(0) - F_1(0)} dH(x) = H\left(\frac{1}{2}\right) < \frac{1}{2},$$

and

$$\lim_{r\to\infty} V_1 = H\left(\frac{1}{2}\right) + \lim_{r\to\infty} \int_{\frac{1}{2}}^{1} \frac{F_0(0) - F_1(0)}{F_0(0) - \frac{1-x}{x} F_1(0)} dH(x) = H\left(\frac{1}{2}\right) < \frac{1}{2}$$

Hence, in the limit, the government loses the election in both states. Since $V_0 < V_1$ for all r, by continuity there exists $r > 1$ at which $V_1 > \frac{1}{2} > V_0$, that is, at which the election aggregates information. □

References

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

Alonso, R. and Câmara, O. (2016). Persuading voters. *The American Economic Review*, 106(11):3590–3605.

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

Bardhi, A. and Guo, Y. (2018). Modes of persuasion toward unanimous consent. *Theoretical Economics*, 13(3):1111–1149.

Bergemann, D. and Morris, S. (2016). Information design, bayesian persuasion, and bayes correlated equilibrium. *American Economic Review*, 106(5):586–91.

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

Cai, H. and Wang, J. T.-Y. (2006). Overcommunication in strategic information transmission games. *Games and Economic Behavior*, 56(1):7–36.

Chan, J., Gupta, S., Li, F., and Wang, Y. (2019). Pivotal persuasion. *Journal of Economic theory*, 180:178–202.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., and Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559.

Denter, P., Dumav, M., and Ginzburg, B. (2021). Social Connectivity, Media Bias, and Correlation Neglect. *The Economic Journal*, 131:2033–2057.

Edmond, C. and Lu, Y. K. (2021). Creating confusion. *Journal of Economic Theory*, 191:105145.

Enke, B. and Graeber, T. (2019). Cognitive uncertainty.

Enke, B. and Zimmermann, F. (2019). Correlation neglect in belief formation. *The Review of Economic Studies*, 86(1):313–332.

Freedom House (2017). Manipulating social media to undermine democracy. Accessed on 26 April 2021.

Freedom House (2019). The crisis of social media. Accessed on 26 April 2021.

Ginzburg, B. (2019). Optimal information censorship. *Journal of Economic Behavior & Organization*, 163:377–385.

Ginzburg, B. and Guerra, J.-A. (2019). When collective ignorance is bliss: Theory and experiment on voting for learning. *Journal of Public Economics*, 169:52–64.

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., and Sedova, K. (2023). Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

Gorodnichenko, Y., Pham, T., and Talavera, O. (2021). Social media, sentiment and public opinions: Evidence from# brexit and# uselection. *European Economic Review*, 136:103772.

Gradwohl, R., Heller, Y., and Hillman, A. (2022). Social media and democracy. *arXiv preprint arXiv:2206.14430*.

Hao, K. (2021). Troll farms reached 140 million americans a month on facebook before 2020 election, internal report shows. *MIT Technology Review*.

Heese, C. and Lauermann, S. (2021). Persuasion and information aggregation in elections.

Iyengar, S. and Krupenkin, M. (2018). The strengthening of partisan affect. *Political Psychology*, 39:201–218.

Jin, G. Z., Luca, M., and Martin, D. (2021). Is no news (perceived as) bad news? an experimental investigation of information disclosure. *American Economic Journal: Microeconomics*, 13(2):141–73.

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

Kartik, N., Ottaviani, M., and Squintani, F. (2007). Credulity, lies, and costly talk. *Journal of Economic Theory*, 134(1):93–116.

King, G., Pan, J., and Roberts, M. E. (2017). How the chinese government fabricates social media posts for strategic distraction, not engaged argument. *American political science review*, 111(3):484–501.

Kolotilin, A., Mylovanov, T., and Zapechelnyuk, A. (2022). Censorship as optimal persuasion. *Theoretical Economics*, 17(2):561–585.

Levy, G. and Razin, R. (2015). Correlation neglect, voting behavior, and information aggregation. *American Economic Review*, 105(4):1634–45.

Martherus, J. L., Martinez, A. G., Piff, P. K., and Theodoridis, A. G. (2021). Party animals? extreme partisan polarization and dehumanization. *Political Behavior*, 43(2):517–540.

Martin, D. A., Shapiro, J. N., and Nedashkovskaya, M. (2019). Recent trends in online foreign influence efforts. *Journal of Information Warfare*, 18(3):15–48.

Matyskova, L. (2018). Bayesian persuasion with costly information acquisition. *CERGE-EI Working Paper Series*, 614.

McCoy, J., Rahman, T., and Somer, M. (2018). Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities. *American Behavioral Scientist*, 62(1):16–42.

Mylovanov, T. and Zapechelnyuk, A. (2023). Constructive vs toxic argumentation in debates. *American Economic Journal: Microeconomics*.

Patty, J. W. and Weber, R. A. (2007). Letting the good times roll: A theory of voter inference and experimental evidence. *Public Choice*, 130(3):293–310.

Reuters (2020). Facebook, twitter dismantle global array of disinformation networks. Accessed on 19 April 2021.

Sun, J., Schram, A. J., and Sloof, R. (2022). Public persuasion in elections: Single-crossing property and the optimality of censorship. *Available at SSRN 4028840*.

The Guardian (2011). Revealed: Us spy operation that manipulates social media. Accessed on 19 April 2021.

The New Republic (2017). Rodrigo duterte's army of online trolls. Accessed on 19 April 2021.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4):297–323.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.

Wang, Y. (2015). Bayesian persuasion with multiple receivers.

Zhuravskaya, E., Petrova, M., and Enikolopov, R. (2020). Political effects of the internet and social media. *Annual Review of Economics*, 12:415–438.