Curated Truths: A Model of Media Bias under Social Media Algorithms and Fake News Competition*

Anirban Kar[†] Shubhro Sarkar[‡]

December 6, 2025

This version is incomplete and work is in progress.

Abstract

We examine the reporting strategy of an unbiased mainstream media outlet that competes for user endorsements on a social media platform against a bot disseminating fake news. Platform users comprise quasi-rational voters who value both accurate information and the endorsement of reports consistent with their ideologies, and who are, on average, nonpartisan. The bot generates right-leaning reports with an exogenously given probability, while the platform employs a personalized algorithm designed to maximize user engagement. Our model predicts that, in equilibrium, the media outlet adopts a left-leaning bias to differentiate itself from the bot. This bias diminishes as (i) the algorithm places greater weight on content previously endorsed by users and (ii) voters become polarized towards the right. The relationship between media bias and the right-bias of the bot follows a U-shape (is positive) for higher (lower) degrees of personalization.

Keywords: Media bias, fake news, social media algorithms, polarization.
JEL codes: D72, D82, D83, L82, C72.

*We thank Sourav Bhattacharya, Parikshit Ghosh and Srijita Ghosh for helpful comments and suggestions. All errors are our own.

[†]Department of Economics, Delhi School of Economics, University of Delhi, New Delhi, India. Email: anirban@econdse.org

[‡]Indira Gandhi Institute of Development Research, Goregaon East, Mumbai, India. Email: shubhro@igidr.ac.in

1 Introduction

The 2020 U.S. presidential elections proved to be a pivotal moment in American history – not only for the deep polarization that it revealed but also for the election security and misinformation debates that accompanied it. In the weeks leading up to and following the elections, President Trump, along with several renowned conservative pundits, repeatedly advanced the claim that the Democrats had 'stolen' the election, that mail-in voting had led to fraud, and that ballots had been tampered with. The President's claims were amplified by several right-wing media outlets, which appeared to compete with Fox News for conservative viewership.[1] In the middle of this misinformation mayhem, Fox News and other prominent conservative outlets faced a predicament. They could either protect their reputation and the faith in the democratic processes of the country by resisting such unfounded theories or risk losing viewership of Trump supporters to fringe competitors.

While most of the extant literature focuses on competition between mainstream outlets that are driven by profit or reputation motives, we choose to focus on how mainstream outlets respond to competition from fringe outlets that face no reputational constraints. The advent of the internet lowered entry barriers into the news media industry and facilitated the entry of non-mainstream sources, which are driven by pecuniary or ideological motives. Such outlets frequently engage in the dissemination of "fake news" as they simultaneously seek to provide entertainment and generate content that indulges the tastes of their audience.[2]

The problem of the spread of misinformation has been aggravated by the fact that an increasing fraction of the population gets their information on social media platforms, and that the content circulated on such platforms has "no significant third-party filtering, fact-checking, or editorial judgment"

[1] For instance, YouTube host Steven Crowder and a far-right outlet, Newsmax, consistently propagated conspiracy theories that questioned the election results, fueling a surge in Facebook engagement with the Newsmax content. Similarly, The Gateway Pundit posted claims of mysterious spikes in voter registrations in Michigan and Wisconsin, and falsely accused election officials in Georgia of rigging the election results. Briebart published early stories implying that there were irregularities in Pennsylvania and Arizona, that a large number of dead people had voted, and that voting machines were switching votes from Trump to Biden.

[2] Fake news is "purposefully crafted, sensational, emotionally charged, misleading or totally fabricated information that mimics the form of mainstream news"(Zimdars and McLeod (2020)).

(Allcott and Gentzkow (2017)). To make matters worse, social media platforms are known to use machine learning to personalize users' feeds based on past behavior, leading to the creation of "echo chambers" or "filter bubbles". A collaborative investigation by The Guardian and a former Google engineer revealed systematic biases in YouTube's recommendation system, which disproportionately promoted videos that were "divisive, sensational, and conspiratorial" (Lewis (2018), Lewis and McCormick (2018)). Similarly, in response to declining user engagement, Facebook restructured its recommendation algorithms, inadvertently accelerating the circulation of exaggerated and attention-grabbing content.

As mainstream firms compete with fringe outlets through a social media platform over viewership of an audience with heterogeneous preferences, we investigate four central questions:

(i) As the *fake news* circulated on social media by non-mainstream sources (either individuals or organizations) becomes more ideologically skewed, does it amplify mainstream media bias?

(ii) Does the algorithm used by social media platforms to generate individual-specific "feeds"exacerbate the proclivity of mainstream outlets to slant information in the face of competition from such non-mainstream outlets?

(iii) To what extent does polarization drive the mainstream media bias in a market that is catered to by mainstream and fringe outlets?

(iv) In a political economy context, how does the interplay of media bias, incidence of fake news and social media algorithms affect the vote shares of the candidates?

To answer these research questions, we develop a model that links mainstream media bias to the incidence of fake news in a setup where the electorate obtains information through a social media platform that is driven to maximize engagement. We assume that a continuum of voters who are unaware of the true state rely on (noisy) reports over two stages before they cast their vote. These voters get utility one if their vote matches the true state and zero otherwise.

In the first stage, a mainstream media firm (media) and an online bot (bot) furnish two reports to the voter. While the media observes the true state, which could be left (L) or right (R), it commits to an editorial policy and could choose to slant its report for both stages through its reporting strategy, σ (Gentzkow and Shapiro (2006)). Our paper is therefore closer to the information design literature in which the sender commits to a signaling strategy (Kamenica and Gentzkow (2011)) and is distinct from the literature with cheap talk (Crawford and Sobel (1982)).[3] The bot, on the other hand, regardless of the state, always reports R with probability $\lambda \geq \frac{2}{3}$. While voters observe the first stage reports, they are unable to distinguish which report came from which source. Voters endorse one of the reports, following which the social media algorithm displays a report in the second stage, from the same outlet that the voter endorsed in the first stage, with probability $\alpha \geq \frac{2}{3}$. Both λ and α are given exogenously.

We assume that voters are quasi-rational in the sense that they care about the true state of the world, but get an additional (dis)utility from endorsing a right-leaning report that reflects the strength of their ideological preferences. In our baseline model, we assume that the (net) idiosyncratic utility to a voter from endorsing report R is given by γ and the population of voters is unbiased on average.[4] In this setting, the media has to resolve the tension between reporting the true state truthfully and conforming to the prior beliefs of their readers. While it is natural to expect the media to cater to the preferences of a right-leaning audience, it is not immediately obvious what the optimal reporting strategy would be for a firm aiming to maximize endorsements from an unbiased electorate.

It is noteworthy that while voters do not face a dilemma while voting as they choose the more likely option, they do so while making endorsements. On the one hand, we expect voters to endorse the report that aligns with their ideological preferences. On the other hand, voters would want to endorse the more informative report of the media given the platform's algorithm, as it facilitates access to a second informative report and improves their inference about the true state of the world.

We show that the optimal reporting strategy of the media comprises generating *left-biased* reports as it tries to distinguish itself from the bot. It reports the state L truthfully but reports R with probability less than one when the true state is R (bias). The intuition behind the above result is as follows. When the state is R, the media could report the true state truthfully or report L with some probability. The likelihood of endorsement from reporting R is $\frac{1}{2}$, as the bot frequently reports R. On the contrary, if the media reports L, it attempts to signal its quality to the voter and significantly improves its chances of endorsement. At this juncture, the media faces a tradeoff. If it

[3] As an extension, we also analyze the cheap-talk version of our model.

[4] Probabilistic voting models usually have a factor similar to γ.

misreports R (as L) too often, then the quality of its report falls, and the media loses some (moderately) right-biased readers, who would have endorsed the media for its informativeness.

The signaling incentive becomes stronger in an alternative setup in which the media uses a reporting strategy in the first stage and reports truthfully in the second. In this case, we show that the media always reports L. The media follows the same strategy when there are homogeneous voters with $\gamma = 0$. In this case, the firm does not face any incentive to slant its report in favor of the voters' ideological leanings. Our analysis shows that the media firm faces a tradeoff in our baseline model as it commits to a reporting strategy - in its attempt to distinguish itself from the bot, it could report L for both states of the world. However, by doing so, it would risk mirroring the behavior of the bot in the other direction, thereby undermining the credibility of its information in the eyes of the voter. Our insight that the media may choose to slant the news follows that of Morris (2001), where an informed advisor having identical preferences to an uninformed decision maker may counterintuitively choose to misreport his information to preserve his reputation of being unbiased.

The raison d'être of our study is to elucidate how the intensity of right-leaning bias of the fake news competitor, λ, and the frequency with which the social media platform shows content from a source previously endorsed by a voter, α, affects the bias of the media. We find that for α closer to the lower end of the spectrum, there is a positive relationship between the left-bias of the media and λ, such that as the intensity of right-bias of the bot rises, the media tries to distinguish itself by leaning further to the left. For higher values of α, the relationship between media bias and λ becomes non-monotonic (Figure 2b). For lower values of λ, the media reduces its bias by reporting R when the state is R more frequently as the right-bias of the bot increases. However, as λ crosses a threshold, the media chooses to raise its left-bias as the right-bias of the bot rises.

We also find that the left bias of the media is falling in α. Therefore, as the algorithm increasingly shows content from the endorsed outlet in the second stage as it attempts to maximize user engagement, it counterintuitively ends up reducing mainstream media bias. This strategy incentivizes the voter to endorse the more informative report, as he wishes to gain access to a high-quality report in the second stage. With rising polarization (i.e., as the upper bound of γ rises), we show that the bias of the media reduces, as the likelihood of being endorsed by an increasingly right-biased electorate decreases. In the extreme case where the upper bound of the support of γ

approaches infinity, the media becomes right-leaning by always reporting R. Finally, with $\lambda = 1$, we show that the vote share for R is rising in α. The algorithm, therefore, improves the chances of the R option being chosen in equilibrium.

Our paper is closest to Yea (2018), in which two ideologically driven outlets (having opposing preferences), one mainstream and the other circulating fake news, sequentially attempt to persuade a decision-maker, who receives one of the two reports. He finds that when the probability that the decision-maker is exposed to fake news is low (high), a rising prevalence of fake news leads the real news sender to generate more (less) informative news. In a recent paper, Chowdhury (2024) demonstrates that when epistemic institutions such as academia, the judiciary, and civil society operate with moderate efficiency, a social media platform adopting identity-based censoring as an equilibrium strategy disseminates both confirmatory and contrarian reports.

2 Literature Review

The causes and effects of media bias has been examined extensively in the literature, both theoretically and empirically (Groseclose and Milyo (2005), Gentzkow and Shapiro (2010), DellaVigna and Kaplan (2007), Gerber et al. (2009)). Some studies adopt a spply side approach (Baron (2006), Besley and Prat (2006), Djankov et al. (2003), Anderson and McLaren (2012)) while others emphasize the role of consumers' innate preferences in driving such bias (Mullainathan and Shleifer (2005), Gentzkow and Shapiro (2006), Foerster (2023), Guo and Lai (2014)).[5] While a supply-side perspective shows that media bias reduces with an increase in competition, Mullainathan and Shleifer (2005) establish that reader heterogeneity is more important than competition in improving the accuracy of reported news and that competition leads to a reduction in prices but not bias. Xiang and Sarvary (2007) counterintuitively show that an increase in the number of conscientious consumers may increase media bias, even when there are limits on the extent to which the media can slant information. Gentzkow and Shapiro (2006), on the other hand, find that a media firm may slant its reports to conform to the prior of a set of homogeneous consumers as it seeks to build its reputation, and that competition would reduce such bias. Our paper

[5]For a review of the theoretical literature on how market forces drive media bias, see Gentzkow et al. (2015)

contributes to this body of work by offering a fresh perspective through the lens of misinformation and the dynamics of social media algorithms.

Since the media firm commits to an editorial policy before Nature selects the true state of the world, our paper connects to the literature on persuasion games. In their seminal contribution, Kamenica and Gentzkow (2011) derive necessary and sufficient conditions under which a sender can choose an information structure that induces the receiver to take an action strictly beneficial to the sender. Their analysis anchors information design in convex geometry and demonstrates that partial, rather than full disclosure, is often optimal. Rayo and Segal (2010) characterize the optimal disclosure rule when the sender's information is two-dimensional and the receiver's opportunity cost is private; the authors demonstrate that the principal's utility is maximized under a partial information disclosure scheme in which prospects with heterogeneous values to agents and profit implications for the principal are pooled into the same signal. Gentzkow and Kamenica (2017) extend Kamenica and Gentzkow (2011) to environments with multiple senders and allow for arbitrary correlation among senders' signals, finding that competition leads to full revelation.[6] Our setting differs from standard persuasion games in that the receiver (voter) takes two actions, only one of which affects the payoffs of both the sender (media) and the receiver. It also departs from the multi-sender persuasion literature because one of the senders in our framework is non-strategic.

Recent research has highlighted the role of social media across a multitude of contexts, including political campaigns (Aziz and Bischoff (2025); Liberini et al. (2025), gender-based violence (Battisti et al. (2024), and social movements (Flückiger and Ludwig (2025). Yet the political-economy literature has devoted comparatively little attention to the pernicious effects of social media platforms—and their underlying algorithms—in amplifying misinformation and thereby undermining information disclosure. Azzimonti and Fernandes (2023) study a model in which agents embedded in a network acquire information either from their social contacts or from a bot disseminating fake news. Agents can thus be exposed to misinformation directly from the bot or indirectly through their network neighbors. They show that greater bot centrality increases polarization but reduces overall

[6] Brocas et al. (2012) and Gul and Pesendorfer (2012) analyze settings with two senders holding opposed interests, each generating costly signals about a binary state. For additional work on multi-sender communication games, see Battaglini (2002), Krishna and Morgan (2001), and Milgrom and Roberts (1986).

misinformation, while asymmetric bot centrality generates the opposite pattern. Acemoglu et al. (2024) set up a model that explicates how misinformation spreads over social media populated by Bayesian agents who sequentially observe an article and decide whether to share it with other directly connected agents. Users possess ideological biases and simultaneously derive utility from positive interactions and disutility from sharing misinformation. In equilibrium, a platform maximizing engagement chooses a network structure with minimal homophily when content is highly reliable, but switches to maximal homophily when reliability is low—thereby generating echo chambers of like-minded users.

3 Model with Personalization Algorithm

States, reports and election: We have a binary state of nature - left (L) and right (R) - realized with equal probability. The realized state remains the same throughout the game. Voters do not observe the state directly, they receive information about it only through media reports, which may or may not be correct. Reports are also denoted by L and R. There are two stages; voters see media report(s) in each stage and subsequently vote for either L or R.

Media: There are two kinds of media - 'mainstream media' and 'on-line bots'. A mainstream media (henceforth media, for brevity) can observe the state perfectly. However, it reports as per its editorial policy $\sigma = \{\sigma_{LL}, \sigma_{RR}\}$, where σ_{LL} represents the probability of reporting L upon observing state L and similarly σ_{RR} represents the probability of reporting R upon observing state R. σ is interpreted as media slant. It is common in the information design literature to assume that the information designer (here the media) commits to σ. Following Gentzkow and Shapiro (2006), we also assume that $\sigma_{RR} + \sigma_{LL} \geq 1$.

The on-line bot (henceforth bot) reports R with probability λ irrespective of the true state. We assume that the bot is biased towards R. Without loss of generality, $\lambda > \frac{1}{2}$. However, to avoid unnecessary computation, we assume that the bot is strongly biased, that is $\lambda > \frac{2}{3}$. Automation strategy of the bot and the value of λ are common knowledge.

The media and the bot post one report each simultaneously, at each stage. They draw stage 2 reports independently of their stage 1 reports. Hence, it is possible that the reports from a particular source differ across stages.

Platform: Voters receive reports from the media and the bot through a platform. In stage 1, each posts a report on the platform. Voters know that one report comes from the media and the other from the bot but they cannot make out which is which. Each voter endorses one of the two reports. Since the source is not known, voters endorsement is based only on the content. The personalization algorithm (PA) works as follows. In the second stage, each voter gets only one report and it is more likely to come from the source she had endorsed earlier. To be concrete, suppose that in stage 1, voter i has endorsed a report that was posted by the bot (resp. media). Then in the second stage, i sees the bot's report with probability α and media's report with $(1-\alpha)$. We assume that the personalization mechanism is sufficiently strong. To maintain the same cut-off as λ, we take $\alpha \geq \frac{2}{3}$. The objective of the media is to maximize endorsement in stage 1.

Voters: Each voter takes two decisions - at stage 1, she endorses exactly one report on the platform (as mentioned above), and after reading the second report at stage 2, she votes for either L or R. Voting and the endorsement, each action gives her a certain pay-off. By endorsing report R, she gets a pay-off of γ while endorsing report L gives 0. γ is individual specific and can be interpreted as a voter's (net) bias for R. Overall, in the society, $\gamma \sim Uniform\ [-1, d]$. An agent is left-biased if her γ is negative and is right-biased otherwise. However, the societal bias, on average, depends on the parameter d. Finally, a voter gets 1 if her vote matches with the state and 0 otherwise.

Time-line: First, the media commits to σ. Next, the state of nature is revealed to the media. Voters observe two stage 1 reports on the platform and endorse one of the reports. Then they see the stage 2 report and vote. All pay-off are realized after election.

We compute PBE of this game and check (i) how the equilibrium media slant changes with exogenous increase in λ (bias of the bot) and α (PA intensity) and (ii) how the equilibrium vote share of R changes with an exogenous increase in λ and α.

4 Preliminaries

Second stage history for a voter i consists of the two first stage media reports, her own endorsement and one second stage report. Note that the second stage report is individual specific due to PA while the first stage reports are the same for all voters. Stage 1 reports can take three possible values $(L,L),(R,R),(R,L)$. Facing (R,L), a voter endorses either a L report or a R report. When two reports are identical, that is, first period reports are either (R,R) or (L,L), a voter cannot distinguish between the two and chooses one with equal probability. To distinguish an endorsement from the reports, we use small letters, l and r, where l (resp. r) means endorsement of L (resp. R) report.

Take the following history for voter i: first period reports are (R,L), i has endorsed report L and second period personalized report to i is R. Probability of this history given true state R is given below. We denote it by $Z^R_{(R,L),l,R}$, where the superscript R stands for true state R and the subscript describes the history.

$$Z^R_{(R,L),l,R} = (1-\lambda)\sigma_{RR}[\alpha\lambda + (1-\alpha)\sigma_{RR}] + \lambda(1-\sigma_{RR})[\alpha\sigma_{RR} + (1-\alpha)\lambda]$$

This can be calculated as follows. The first period reports (R,L) can result from either the media publishing R and the bot publishing L or the reverse. The former occurs with probability $(1-\lambda)\sigma_{RR}$ because the bot publishes L with probability $(1-\lambda)$ and the media publishes R, given state R, with probability σ_{RR}. As voter i endorsed L, which was reported by the bot, i gets the bot in the second stage with probability α. Hence the second period report is R with probability $[\alpha\lambda + (1-\alpha)\sigma_{RR}]$. Alternatively, in stage 1, the media published L and the bot published R. This event has probability $\lambda(1-\sigma_{RR})$. This time, endorsement of L leads voter i to observe the media in stage 2 with probability α. Thus the probability of R in stage 2 is $[\alpha\sigma_{RR} + (1-\alpha)\sigma_{RR}]$.

By similar arguments, probability of the previous history given state L is

$$Z^L_{(R,L),l,R} = (1-\lambda)(1-\sigma_{LL})[\alpha\lambda + (1-\alpha)(1-\sigma_{LL})] + \lambda\sigma_{LL}[\alpha(1-\sigma_{LL}) + (1-\alpha)\lambda]$$

It would be convenient to write $Z^R_{(R,L),l,R}$ and $Z^L_{(R,L),l,R}$ as the value of the following function computed at $x = \sigma_{RR}$ and $x = 1-\sigma_{LL}$ respectively. We abuse notation and name the function $Z_{(R,L),l,R}$.

$$Z_{(R,L),l,R}(x) = (1-\lambda)x[\alpha\lambda + (1-\alpha)x] + \lambda(1-x)[\alpha x + (1-\alpha)\lambda]$$

Since voters have equal prior, conditional probability of state R, given the history $(R,L),l,R$, is

$$\frac{Z^R_{(R,L),l,R}}{Z^R_{(R,L),l,R}+Z^L_{(R,L),l,R}}$$

where the denominator gives twice the unconditional probability of the history $(R,L),l,R$. Since pay-off from matching the state is 1 under both L and R, voter i votes R iff $Z^R_{(R,L),l,R}\geq Z^L_{(R,L),l,R}$. Her voting pay-off at this history is

$$\frac{\max\left\{Z^R_{(R,L),l,R},Z^L_{(R,L),l,R}\right\}}{2P\left((R,L),l,R\right)}$$

We do similar calculation for all histories with mixed reports in stage 1. These are $(R,L),l,L$ (mixed report in stage 1, voter picks L and obtains report L in stage 2), $(R,L),r,R$ (mixed report in stage 1, voter picks R and obtains report R in stage 2), and $(R,L),r,L$ (mixed report in stage 1, voter picks R and obtains report L in stage 2). In each case, we identify the function that has to be evaluated at $x=\sigma_{RR}$ and $x=1-\sigma_{LL}$, to obtain the probability of that history given state R and state L respectively.

$$Z_{(R,L),l,L}(x)=(1-\lambda)\,x\,[\alpha(1-\lambda)+(1-\alpha)(1-x)]+\lambda(1-x)[\alpha(1-x)+(1-\alpha)(1-\lambda)]$$
$$Z_{(R,L),r,R}(x)=(1-\lambda)\,x\,[\alpha x+(1-\alpha)\lambda]+\lambda(1-x)[\alpha\lambda+(1-\alpha)\,x]$$
$$Z_{(R,L),r,L}(x)=(1-\lambda)\,x\,[\alpha(1-x)+(1-\alpha)(1-\lambda)]+\lambda(1-x)[\alpha(1-\lambda)+(1-\alpha)(1-x)]$$

Now we can calculate the expected pay-off from endorsement at stage 1. Endorsing report L (that is choosing action l) gives

$$P(R\mid(R,L),l)\frac{\max\left\{Z^R_{(R,L),l,R},Z^L_{(R,L),l,R}\right\}}{2P\left((R,L),l,R\right)}+P(L\mid(R,L),l)\frac{\max\left\{Z^R_{(R,L),l,L},Z^L_{(R,L),l,L}\right\}}{2P\left((R,L),l,L\right)}$$
$$=\frac{1}{2P\left((R,L)\right)}\left[\max\left\{Z^R_{(R,L),l,R},Z^L_{(R,L),l,R}\right\}+\max\left\{Z^R_{(R,L),l,L},Z^L_{(R,L),l,L}\right\}\right]$$

This is weighted average of the voting pay-off at the history $(R,L),l,R$ and the history $(R,L),l,L$. Expected pay-off of from endorsing R (choosing action r) at (R,L)

$$\gamma_i+\frac{1}{2P\left((R,L)\right)}\left[\max\left\{Z^R_{(R,L),r,R},Z^L_{(R,L),r,R}\right\}+\max\left\{Z^R_{(R,L),r,L},Z^L_{(R,L),r,L}\right\}\right]$$

The first term is voter i's net bias for R-report and the second term is her expected voting pay-off. Let γ^* be the agent who is indifferent between endorsing L and R. Then

$$\gamma^* = \frac{1}{2P((R,L))}\Big[\Big[\max\Big\{Z^R_{(R,L),l,R}, Z^L_{(R,L),l,R}\Big\} + \max\Big\{Z^R_{(R,L),l,L}, Z^L_{(R,L),l,L}\Big\}\Big] - \Big[\max\Big\{Z^R_{(R,L),r,R}, Z^L_{(R,L),r,R}\Big\} + \max\Big\{Z^R_{(R,L),r,L}, Z^L_{(R,L),r,L}\Big\}\Big]\Big] \tag{1}$$

The numerator will be denoted by Δ. All agents with γ below γ^* endorse L, and those above, endorse R. At the remaining first stage histories, (R,R) and (L,L), voters pick the media with probability $\frac{1}{2}$.

Now we compute the expected pay-off of the media as follows: If the state is L (and hence the media observes L) then the media reports L with probability σ_{LL}. The bot reports R with probability λ and L with $(1-\lambda)$. The former leads to (R,L) and size of endorsement for media is $\frac{\gamma^*+1}{1+d}$, because those below γ^* endorses L. For the latter, at (R,R) history endorsement size is $\frac{1}{2}$. This explains the first term in the following expected pay-off expression. The second term corresponds to the media reporting R at state L. In this case, the first stage reports are mixed if the bot reports L and voters above γ^* endorse the media. The third and the fourth terms correspond to media reporting R and L respectively at state R.

$$\begin{aligned}\Pi(\sigma) =& \frac{1}{2}\left[\sigma_{LL}\left(\lambda\frac{(\gamma^*+1)}{1+d} + (1-\lambda)\frac{1}{2}\right) + (1-\sigma_{LL})\left(\lambda\frac{1}{2} + (1-\lambda)\frac{(d-\gamma^*)}{1+d}\right)\right] + \\ & \frac{1}{2}\left[\sigma_{RR}\left(\lambda\frac{1}{2} + (1-\lambda)\frac{(d-\gamma^*)}{1+d}\right) + (1-\sigma_{RR})\left(\lambda\frac{(\gamma^*+1)}{1+d} + (1-\lambda)\frac{1}{2}\right)\right] \\ =& \frac{1}{2}\left[(1+\sigma_{RR}-\sigma_{LL})\left(\lambda\frac{1}{2} + (1-\lambda)\frac{(d-\gamma^*)}{1+d}\right) + (1-\sigma_{RR}+\sigma_{LL})\left(\lambda\frac{(\gamma^*+1)}{1+d} + (1-\lambda)\frac{1}{2}\right)\right]\end{aligned}$$

To obtain the PBE of this game, we maximize $\Pi(\sigma)$ with respect to σ_{RR}, σ_{LL}, where $\sigma_{RR} + \sigma_{LL} \geq 1$.

Before proceeding further, we re-parametrize the optimization problem. Instead of maximizing the objective function over σ_{RR} and σ_{LL}, we change the variables to $a = [\sigma_{RR} - \sigma_{LL}]$ and $b = [\sigma_{RR} + \sigma_{LL} - 1]$. This is possible because there is one-to-one mapping between $(\sigma_{RR}, \sigma_{LL}) \longleftrightarrow (a, b)$. Moreover, a and b have natural interpretations: a measures the media bias towards R while b

can be interpreted as informativeness of the media report. When $b = 0$, that is, $\sigma_{RR} = 1 - \sigma_{LL}$, then media reports R with equal probability irrespective of its signal, hence the report is not informative at all. On the other hand b is highest when media truthfully reports its signal, that is, $\sigma_{RR} = \sigma_{LL} = 1$. The range of a and b is as follows: $-1 \leq a \leq 1$ and $0 \leq b \leq 1 - |a|$. The following diagram depicts the re-parametrization.

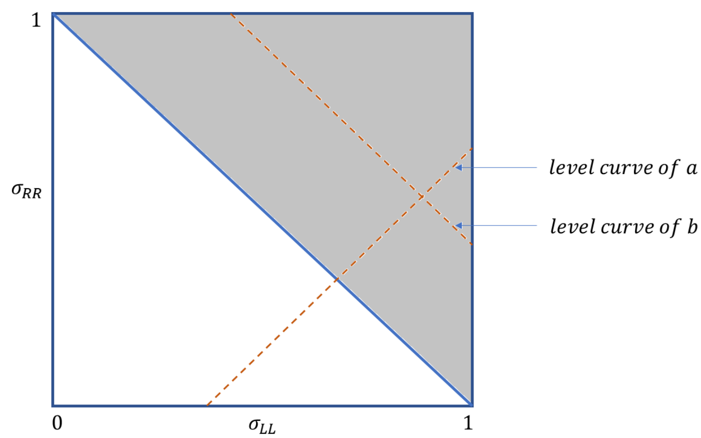


Figure 1: Re-parametrization of the choice space

The re-parametrized objective function is

$$\begin{aligned}\Pi(\sigma) &= \frac{1}{2}\left[(1+a)\left(\lambda\frac{1}{2}+(1-\lambda)\frac{\left(d-\gamma^*\right)}{1+d}\right)+(1-a)\left(\lambda\frac{\left(\gamma^*+1\right)}{1+d}+(1-\lambda)\frac{1}{2}\right)\right] \\ &= \frac{1}{2}\left[\kappa_1 + a\kappa_2 + (2\lambda-1-a)\frac{\gamma^*}{1+d}\right] \qquad (2)\end{aligned}$$

where, $\kappa_1 = \left[\frac{1}{2} + \frac{1}{1+d}(d(1-\lambda)+\lambda)\right]$ and $\kappa_2 = \left[\frac{1}{2}(2\lambda-1) + \frac{1}{1+d}(d(1-\lambda)-\lambda)\right] = \left[\frac{d}{1+d} - \frac{1}{2}\right]$. Note further that κ_2 is positive if and only if $d \geq 1$.

Next, we unpack γ^* using new parameters a and b. This requires a couple of preliminary steps. For any stage 2 history h, define,

$$\Delta(Z_h) = \left(Z_h^R - Z_h^L\right) = Z_h(\sigma_{RR}) - Z_h(1-\sigma_{LL})$$

Detail of the following calculations are given in the Appendix *A*.

$$\Delta\left(Z_{(R,L),l,R}\right) \geq 0 \text{ iff } a \leq \left[\frac{\lambda(2\alpha-\lambda)}{(\lambda+\alpha-1)}-1\right] = c_1$$

$$\Delta\left(Z_{(R,L),l,L}\right) \geq 0 \text{ iff } a \geq \left[1-\frac{(1-\lambda)(2\alpha-\lambda-1)}{(\alpha+\lambda-1)}\right] = c_2$$

$$\Delta\left(Z_{(R,L),r,R}\right) = \begin{cases} \leq 0 \text{ for all } a \text{ when } \alpha \leq \lambda \\ \geq 0 \text{ iff } a \geq \left[\frac{\lambda(2\alpha+\lambda-2)}{(\alpha-\lambda)}-1\right] = c_3 \text{ when } \alpha > \lambda \end{cases}$$

$$\Delta\left(Z_{(R,L),r,L}\right) = \begin{cases} \leq 0 \text{ for all } a \text{ when } \alpha \leq \lambda \\ \geq 0 \text{ iff } a \leq \left[1-\frac{(1-\lambda)(2\alpha+\lambda-1)}{(\alpha-\lambda)}\right] = c_4 \text{ when } \alpha > \lambda \end{cases}$$

These cut-off partition the domain into several intervals of a. Note also that these intervals are independent of b. Each interval has a different value of γ^*. Moreover, the partition itself depends on λ and α. Proof of Lemma 1 is given in Appendix *B*.

Lemma 1 $\gamma^* = \frac{\Delta}{2P((R,L))}$, *where* $P((R,L)) = \frac{1}{2}(1-(2\lambda-1)a)$. Δ *is given below*[7]

If $\alpha > \lambda$ *then*

$$\Delta = \begin{cases} b\left[2\lambda(1-\lambda)-(2\lambda-1)a\right] \textit{ if } a \in [-1, c_4] \\ b\left[\lambda(2\alpha-\lambda)-(\alpha+\lambda-1)(a+1)\right] \textit{ if } a \in [\max\{-1, c_4\}, c_1] \\ 0 \textit{ if } a \in [c_1, \min\{c_2, c_3, 1\}] \\ b\left[(\alpha-\lambda)-\lambda(2\alpha-\lambda)+(\alpha+\lambda-1)a\right] \textit{ if } a \in [c_2, \min\{c_3, 1\}] \\ b\left[\lambda(2\alpha+\lambda-2)-(\alpha-\lambda)(a+1)\right] \textit{ if } a \in [c_3, \min\{c_2, 1\}] \\ b\left[-2\lambda(1-\lambda)+(2\lambda-1)a\right] \textit{ if } a \in [\max\{c_2, c_3\}, 1] \end{cases}$$

If $\alpha \leq \lambda$ *then*

$$\Delta = \begin{cases} b\left[\lambda(2\alpha-\lambda)-(\alpha+\lambda-1)(a+1)\right] \textit{ if } a \in [-1, c_1] \\ 0 \textit{ if } a \in [c_1, 1] \end{cases}$$

No interval other than those listed above is possible.

[7] Since Δ at two neighbouring intervals match on the boundary due to continuity; we use closed intervals throughout to avoid unnecessary confusion.

5 Media Equilibrium

Our first result shows that the optimal choice of σ is not an interior solution, that is, either it reports R truthfully or it reports L truthfully.

Theorem 2 *Media's equilibrium choice must have either $\sigma_{RR} = 1$ or $\sigma_{LL} = 1$.*

Proof: Note that at all intervals, $\Delta = bg(\alpha, \lambda, a)$ for some function $g(\alpha, \lambda, a)$, that varies across the interval. Nonetheless, the objective function can be written as

$$\frac{1}{2}\left[\kappa_1 + a\kappa_2 + b\frac{(2\lambda - 1 - a)}{(1-(2\lambda-1)a)(1+d)} g(\alpha, \lambda, a)\right]$$

For a given economy $\{\alpha, \lambda, d\}$, the first two terms κ_1 and κ_2 are independent of b. Since the third term is a product of b and some function of a, at the optimum, b should be either its highest value $1-|a|$ or the lowest 0.

If $b = 1-|a|$, we get σ on the boundary of the domain, that is either $\sigma_{RR} = 1$ or $\sigma_{LL} = 1$. On the other hand, if $b = 0$ then $\Pi(\sigma) = \frac{1}{2}[\kappa_1 + a\kappa_2]$. In that case optimum a must be, either 1 if $\kappa_2 > 0$, or -1 if $\kappa_2 < 0$. Once again the optimum is on the boundary, because $b = 1-|a|$. □

Theorem 2 reduces the optimization problem to

$$\max_{a\in[-1,1]} \quad \frac{1}{2}\left[\kappa_1 + a\kappa_2 + \frac{(2\lambda-1-a)(1-|a|)g(\alpha,\lambda,a)}{(1+d)(1-(2\lambda-1)a)}\right]$$

Our next result shows that at $d = 1$, that is when the society is unbiased on average, then the media's optimal strategy is L biased. Intuitively the media adopts L bias to separate itself from the R-biased bot.

At $d = 1$, $\kappa_1 = 1$ and $\kappa_2 = 0$. Therefore maximizing Π is equivalent to

$$\max_{a\in[-1,1]} \quad \frac{(2\lambda-1-a)(1-|a|)g(\alpha,\lambda,a)}{(1-(2\lambda-1)a)} \tag{3}$$

Proof of Theorem 3 is in the Appendix C.

Theorem 3 *If $d = 1$, then the optimal media strategy is unique. The optimal choice of a belongs to the interval $[\max\{-1, c_4\}, 0]$ when $\alpha > \lambda$ and it belongs to $[-1, \min\{c_1, 0\}]$ when $\alpha \leq \lambda$. In either case, the optimal a is implicitly given by the following equation*

$$\left[\frac{2\lambda}{(1+a)(1-(2\lambda-1)a)}\right] - \left[\frac{1}{(2\lambda-1-a)} + \frac{(\alpha+\lambda-1)}{(\lambda(2\alpha-\lambda)-(\alpha+\lambda-1)(a+1))}\right] = 0 \tag{4}$$

The optimal value of $b = (1+a)$*. In terms of* σ*, the optimum* $\sigma_{LL} = 1$ *and* $\sigma_{RR} = (1+a) \leq 1$.

An immediate consequence of Theorem 3 is that the maximum value of the objective function 3 is strictly positive. This is so because the maximum is not obtained in the interval $[c_1, \min\{c_2, c_3, 1\}]$ where the objective function takes the value 0 (because g is 0). Hence at the maximum, it has to be strictly positive. We can rewrite the objective function as $(2\lambda - 1 - a)\gamma^*$. Since at the equilibrium, $(2\lambda - 1 - a)$ is positive (because equilibrium a is negative and $2\lambda > 1$), we also have equilibrium $\gamma^* > 0$. It means that when the first stage reports are conflicting, that is first stage history is R, L, then majority of the population endorses the L report in equilibrium.

Corollary 4 *Suppose* $d = 1$*. At the media equilibrium* $\gamma^* > 0$.

It is possible to put stricter bounds on the equilibrium value of a. Although this exercise does not bring any new insight, we still include this result, because we use it later.

Theorem 5 *Take* $d = 1$*. The equilibrium value of* a *has a lower bound* $-\lambda$ *and an upper bound* $\min\{c_1, -(1-\lambda)\}$.

Note that In Theorem 3, the lower bound is $\max\{-1, c_4\}$ when $\alpha > \lambda$ and -1 when $\alpha \leq \lambda$. In Appendix *B*, we showed that $c_4 \leq -\lambda$. Hence Theorem 5 is an improvement for the lower bound. Similarly, Theorem 3 upper bound is 0 when $\alpha > \lambda$ and $\min\{c_1, 0\}$ when $\alpha \leq \lambda$. These can be jointly written as $\min\{c_1, 0\}$ because for $\alpha > \lambda$, $c_1 \geq 0$ (proved in Appendix *B*). Clearly, $\min\{c_1, -(1-\lambda)\}$ in Theorem 5 is an improvement. Proof of Theorem 5 is given in Appendix *D*.

We now check how the equilibrium changes with parameters. The next result shows that the media's left bias decreases (or equivalently, right bias increases) with an increases in the strength of PA.

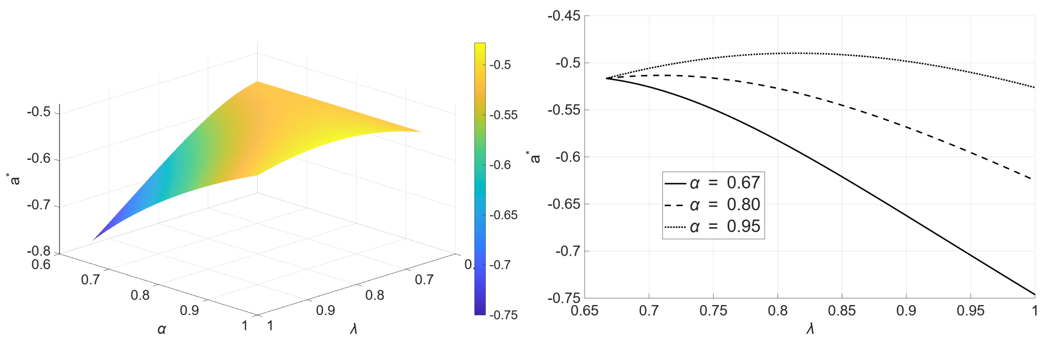
Theorem 6 *Take* $d = 1$*. Everything else remaining the same, an increase in* α *increases the equilibrium* a*. Equivalently, in terms of* σ*,* σ_{LL} *remains the same and* σ_{RR} *increases.*

Proof: This follows immediately by using Implicit Function Theorem on Equation 4. We denote the LHS of Equation 4 as ϕ.

$$\frac{\partial \phi}{\partial \alpha} = -\frac{\lambda(2-3\lambda)}{[\lambda(2\alpha - \lambda) - (\alpha + \lambda - 1)(a+1)]^2} \geq 0$$

The inequality holds because $\lambda \geq \frac{2}{3}$. $\frac{\partial \phi}{\partial a}$ is negative because a maximizes ϕ. Hence by the Implicit Function Theorem, $\frac{\partial a}{\partial \alpha} \geq 0$. □

While we would have preferred to present the comparative statics with respect to λ analytically, we were unable to determine the sign of the partial derivative $\frac{\partial a}{\partial \lambda}$ using the Implicit Function Theorem. Consequently, we resorted to numerical methods. Figure 2b illustrates how the equilibrium value of a changes with λ for three different values of α. We find that when $\alpha = \frac{2}{3}$, there is a negative relationship between the equilibrium value of a and λ. However, for higher values of α, the relationship becomes inverted U-shaped. For higher values of λ, the relationship between a^* and λ becomes negative regardless of the value of α, implying that the media becomes more left-biased as the probability with which the bot reports R rises. However, as α increases, this left bias of the media reduces.



(a) Optimal a as a function of λ and α. (b) Optimal a vs λ for different values of α.

Figure 2: Comparative Statics.

For $d \neq 1$ maximization problem may not have unique maximizer. Our next result shows that the equilibrium a is increasing in d in the following sense

Theorem 7 *Suppose $d_2 > d_1 \geq 1$. If a_1 is an equilibrium at d_1 and a_2 is an equilibrium at d_2 then $a_2 \geq a_1$. Moreover, as d goes to ∞, equilibrium value of a goes to* 1.

Proof: Since a_1 maximizes the objective function, Equation 2, its value at a_1

must be greater than that at a_2. We denote $\pi(a) = (2\lambda - 1 - a)\gamma^*(a)$. Then

$$\left[\kappa_1 + a_1\left[\frac{d_1}{1+d_1} - \frac{1}{2}\right] + \left(\frac{1}{1+d_1}\right)\pi(a_1)\right] \geq \left[\kappa_1 + a_2\left[\frac{d_1}{1+d_1} - \frac{1}{2}\right] + \left(\frac{1}{1+d_1}\right)\pi(a_2)\right]$$
$$\Rightarrow \left(\frac{d_1 - 1}{2}\right)(a_1 - a_2) \geq [\pi(a_2) - \pi(a_1)]$$

Similarly, a_2 maximizes the objective function at d_2. Thus

$$\left(\frac{d_2 - 1}{2}\right)(a_2 - a_1) \geq [\pi(a_1) - \pi(a_2)]$$

Adding above inequalities, we get

$$(a_1 - a_2)\left(\frac{d_1 - d_2}{2}\right) \geq 0$$

Since $d_2 > d_1$, we must have $a_2 \geq a_1$. Moreover, as d goes to ∞, the last term of the objective function goes to 0. The maximization problem reduces to $\left[\kappa_1 + a\left(\frac{d}{1+d} - \frac{1}{2}\right)\right]$ implying equilibrium a converges to 1.

6 Vote share

In this section, we derive the expected vote share of R and L in equilibrium and see how they respond to the parameters of our model. Throughout this part, we assume that the population is symmetrically distributed, that is $d = 1$, and is of size 2. We shall denote the expected vote share of R and L by V_R and V_L respectively. Since the population size is 2, we have $V_R + V_L = 2$.

We proceed as follows. For each second stage history, we find the size of voters, who vote for R. From Theorem 3, we know that in equilibrium $a \in [\max\{-1, c_4\}, 0]$ when $\alpha > \lambda$ and $a \in [-1, \min\{c_1, 0\}]$ when $\alpha \leq \lambda$. In both situation (see the proof of Lemma B), $\Delta\left(Z_{(R,L),l,R}\right) \geq 0$, $\Delta\left(Z_{(R,L),l,L}\right) \leq 0$, $\Delta\left(Z_{(R,L),r,R}\right) \leq 0$ and $\Delta\left(Z_{(R,L),r,L}\right) \leq 0$.

$\Delta\left(Z_{(R,L),l,R}\right) \geq 0$ implies that at the history $(R, L), l, R$ state R is more likely than state L. Hence voters at the end of this history choose R. Size of such voters is $\left(\gamma^* + 1\right)$ because those having $\gamma \leq \gamma^*$ endorse stage 1 report L (action l). Ex-ante probability of this history is $\frac{1}{2}\left[Z^R_{(R,L),l,R} + Z^L_{(R,L),l,R}\right]$ (because both state are equally likely). On the other hand, $\Delta\left(Z_{(R,L),l,L}\right) \leq 0$, $\Delta\left(Z_{(R,L),r,R}\right) \leq 0$ and $\Delta\left(Z_{(R,L),r,L}\right) \leq 0$ imply voters with these histories vote for L.

We are left with histories where first stage reports are identical. Take (R,R). All voters can only endorse R, which can be the bot or the media with equal probability. The probability of second stage history $(R,R),r,R$ given state R can be calculated as follows. $\lambda\sigma_{RR}$ is the probability of first stage report is (R,R) given state R. If one's endorsement picks the bot then the second stage signal is R with probability $(\alpha\lambda+(1-\alpha)\sigma_{RR})$, otherwise the second stage signal is R with probability $(\alpha\sigma_{RR}+(1-\alpha)\lambda)$. Hence probability of $(R,R),r,R$ given state R is

$$\lambda\sigma_{RR}\frac{1}{2}[(\alpha\lambda+(1-\alpha)\sigma_{RR})+(\alpha\sigma_{RR}+(1-\alpha)\lambda)]=\frac{1}{2}\lambda\sigma_{RR}[\lambda+\sigma_{RR}]$$

Similarly, probability of $(R,R),r,R$ given state L is $\left[\frac{1}{2}\lambda(1-\sigma_{LL})(\lambda+(1-\sigma_{LL}))\right]=0$ because at the equilibrium, $\sigma_{LL}=1$. Hence at $(R,R),r,R$, all voters vote for R. By similar argument, probability of history $(R,R),r,L$ is $\left[\frac{1}{2}\lambda\sigma_{RR}((1-\lambda)+(1-\sigma_{RR}))\right]$ given state R and $\left[\frac{1}{2}\lambda(1-\sigma_{LL})((1-\lambda)+\sigma_{LL})\right]$ given state L. Once again the latter probability is 0 and all voters vote for R.

Now take the second stage history $(L,L),l,R$. Probability of this history given state R is $\left[\frac{1}{2}(1-\lambda)(1-\sigma_{RR})(\lambda+\sigma_{RR})\right]$ and given state L is $\left[\frac{1}{2}(1-\lambda)\sigma_{LL}(\lambda+(1-\sigma_{LL}))\right]$. We show below that the latter is greater than the former and hence all voters vote for L. After replacing σ_{LL} and σ_{RR} with their equilibrium values 1 and $(1+a)$ respectively.

$$\begin{aligned}&\frac{1}{2}\lambda(1-\lambda)-\frac{1}{2}(1-\lambda)(-a)(\lambda+1+a)\\=&\frac{1}{2}(1-\lambda)(\lambda+a)(1+a)\end{aligned}$$

This is positive because $(1+a)\geq 0$ and $(\lambda+a)\geq 0$ by Theorem 5.

Finally, the history $(L,L),l,L$ occurs with probability $\left[\frac{1}{2}(1-\lambda)(1-\sigma_{RR})((1-\lambda)+(1-\sigma_{RR}))\right]$ given state R and with probability $\left[\frac{1}{2}(1-\lambda)\sigma_{LL}((1-\lambda)+\sigma_{LL})\right]$ given state L. As $\sigma_{LL}=1$ in equilibrium, state L is more likely at this history and all voters vote for L.

Therefore expected vote share of R is

$$\begin{aligned}V_R=\left(1+\gamma^*\right)\left[\frac{1}{2}\left(Z^R_{(R,L),l,R}+Z^L_{(R,L),l,R}\right)\right]+2\left[\frac{1}{4}\lambda\sigma_{RR}(\lambda+\sigma_{RR})\right]\\+2\left[\frac{1}{4}\lambda\sigma_{RR}((1-\lambda)+(1-\sigma_{RR}))\right]\end{aligned}$$

Three terms in the above expression correspond to history $(R,L),l,R$ and $(R,R),r,R$ and $(R,R),r,L$ in that order. For each term, the part inside square brackets is the probability of that history and the part outside is the size of R voters. Adding the last two terms and replacing equilibrium value of σ_{RR}, we get

$$V_R = (1+\gamma^*)\left[\frac{1}{2}\left(Z^R_{(R,L),l,R} + Z^L_{(R,L),l,R}\right)\right] + \lambda(1+a) \tag{5}$$

Note that V_R is computed at the equilibrium value of a, which in turn depends on α. Therefore to compute $\frac{dV_R}{d\alpha}$, we have to compute both $\frac{\partial V_R}{\partial a}$ and $\frac{\partial V_R}{\partial \alpha}$.

$$\begin{aligned}
\frac{dV_R}{d\alpha} &= \frac{\partial V_R}{\partial \alpha} + \frac{\partial V_R}{\partial a}\frac{da}{d\alpha} \\
&= \frac{1}{2}\left[(1+\gamma^*)\frac{\partial\left[Z^R_{(R,L),l,R} + Z^L_{(R,L),l,R}\right]}{\partial\alpha} + \left[Z^R_{(R,L),l,R} + Z^L_{(R,L),l,R}\right]\frac{\partial\gamma^*}{\partial\alpha}\right] + \\
&\quad \frac{1}{2}\left[(1+\gamma^*)\frac{\partial\left[Z^R_{(R,L),l,R} + Z^L_{(R,L),l,R}\right]}{\partial a} + \left[Z^R_{(R,L),l,R} + Z^L_{(R,L),l,R}\right]\frac{\partial\gamma^*}{\partial a} + \lambda\right]\frac{da}{d\alpha}
\end{aligned}$$

Theorem 8 *Take $d=1$. Partials, $\frac{\partial V_R}{\partial \alpha} < 0$ and $\frac{\partial V_R}{\partial a} > 0$. As λ goes to its upper bound 1, the total effect $\frac{dV_R}{d\alpha}$ is positive, whereas λ goes to its lower bound $\frac{2}{3}$ the total effect is negative.*

Proof of this result is in Appendix E. Theorem 8 has two significant implications. First, if PA increases, without any change in media response, a, vote share of R decreases. So, as such, PA alone is not responsible for a shift towards R. However, the total effect may be a shift towards R, because the media also reduces its L-bias as a response to the change in PA intensity. Secondly, even the total effect on V_R need not be always positive. When λ is relatively small, that is proportion of R-bot is relatively low, then the total effect is also negative. Therefore the much touted link between PA intensity and R vote share could hold but only in the presence of two additional factors - presence of relatively high proportion of R-bots and sufficient time for the media to adjust its policy.

7 Alternative Specifications

In this section, we explore a couple of alternative specifications of our model.

7.1 Homogeneous Voters

Suppose that the voters are homogeneous as they do not get utility from endorsing conforming reports; that is $\gamma = 0$ for all voters. We can adjust media's objective function, Equation 2, accordingly. Note that if γ^* of the main model, or equivalently Δ, is greater than 0, then homogeneous voters vote for L. On the other hand, if γ^* is less than 0, then homogeneous voters vote for R. We denote these by indicator functions $1_{\Delta\geq 0}$ and $1_{\Delta<0}$ respectively. Note that $1_{\Delta<0} = 1 - 1_{\Delta\geq 0}$. Therefore

$$\begin{aligned}\Pi(\sigma) &= \frac{1}{2}\left[(1+a)\left(\lambda\frac{1}{2}+(1-\lambda)1_{\Delta\geq 0}\right)+(1-a)\left(\lambda 1_{\Delta<0}+(1-\lambda)\frac{1}{2}\right)\right] \\ &= \frac{1}{2}\left[\frac{1}{2}+(1-\lambda)+\frac{a}{2}+(2\lambda-1-a)1_{\Delta\geq 0}\right]\end{aligned}$$

Media chooses σ_{RR} and σ_{LL} to maximize $\Pi(\sigma)$. For choice of σ such that $\Delta \geq 0$, $\Pi(\sigma) = \frac{1}{2}\left[\frac{1}{2}+\lambda-\frac{a}{2}\right]$ and otherwise $\pi(\sigma) = \frac{1}{2}\left[\frac{1}{2}+(1-\lambda)+\frac{a}{2}\right]$. This is maximized at $a = -1$ because from Lemma 1, we know that $\Delta \geq 0$. Hence at equilibrium, the media publishes L irrespective of the true state. The intuitive idea is as follows. Since voters are homogeneous, capturing the moderate R-biased voters using news quality is not a concern anymore. The only concern is signaling, which gives rise to the corner solution.

7.2 Truthful Reporting in Stage Two

Here, we relax the condition that the media has to use the same (committed) strategy σ in both the stages. Instead, let us assume that the media can commit to report truthfully in the second stage. This relaxes the quality consideration in the second stage and that should allow the media to signal more aggressively by reporting L.

Theorem 9 *Take $d = 1$. If the media can commit to truth reporting in the second stage then the its first stage equilibrium strategy is $\sigma_{RR} = 0$ and $\sigma_{LL} = 1$.*

Proof of this result is in Appendix *F*.

7.3 Cheap Talk in Stage One

Finally, we consider a situation where the media cannot commit. The first stage media report is cheap talk and in the second stage (since there is no incentive) media tells the truth. As expected, there are several partially revealing equilibrium under cheap talk. Interestingly, unlike the full commitment case, even completely truthful reporting can be an equilibrium for certain parameter values in our domain. However, in all these equilibrium (except the babbling equilibrium), exactly half of the population endorse L report in stage one. In comparison, Corollary 4 showed that majority endorse L report in the commitment game. This shows that commitment plays a crucial role in our arguments. Detailed analysis of cheap talk specification can be found in Appendix G.

8 Conclusion

While mainstream media bias has been the focus of several studies in the past, it is only recently that economists and political scientists have begun to analyze the nature of the relationship between misinformation in the form of fake news and mainstream media bias. In particular, are there conditions under which an increasing circulation of fake news leads mainstream media to adopt more biased reporting strategies? Identifying the contours of this relationship has become increasingly challenging at a time when growing numbers of individuals rely on social media platforms for news consumption. Since voters draw on the information they collect to make electoral decisions, it is also important to understand how the spread of fake news—amplified by platform algorithms—shapes voting outcomes.

Our model is the first of its kind to examine how mainstream media chooses to slant its information as it competes with fringe outlets for viewership over a social media platform using a personalized algorithm that shows content based on prior endorsements. Given that the quasi-rational voters in our setup are unbiased on average, and the media house is ideologically neutral, it would be intuitive to expect the media to report truthfully. However, we show that the media chooses to curate the truth as it tries to distinguish itself from the bot – it reports the left state truthfully and the right state with bias. We find that this left bias reduces as the polarization rises in favor of the right and as the algorithm increasingly shows content from an outlet that a user endorsed in the past. In the event of extreme polarization with $d \to \infty$, the media always reports R. On the other hand, the extent of left bias worsens in alternative settings where (i) voters are homogeneous and unbiased and (ii) the media reports the true state truthfully in the second stage, as it always reports L. We were unable to analytically establish how the frequency with which the bot reports R affects vote shares. We intend to work on these comparative statics in the future, in addition to making the extent of polarization endogenous.

A Detailed calculation of $\Delta(Z_h)$

For a second period history h, We defined

$$\Delta(Z_h) = \left(Z_h^R - Z_h^L\right) = Z_h(\sigma_{RR}) - Z_h(1-\sigma_{LL})$$

We compute $\Delta(Z_h)$ for all histories where first priod reports are (R, L).

1.

$$\begin{aligned}
Z_{(R,L),l,R}(x) &= (1-\lambda)\,x\,[\alpha\lambda + (1-\alpha)\,x] + \lambda(1-x)[\alpha x + (1-\alpha)\lambda] \\
&= \lambda^2(1-\alpha) + x\lambda(2\alpha-\lambda) - x^2(\alpha+\lambda-1) \\
\Delta\left(Z_{(R,L),l,R}\right) &= \lambda(2\alpha-\lambda)(\sigma_{RR}+\sigma_{LL}-1) - (\alpha+\lambda-1)\left(\sigma_{RR}^2 - (1-\sigma_{LL})^2\right) \\
&= \lambda(2\alpha-\lambda)\,b - (\alpha+\lambda-1)\,b\,(a+1) \\
&= b\,[\lambda(2\alpha-\lambda) - (\alpha+\lambda-1)(a+1)]
\end{aligned}$$

Since $b \geq 0$ (by assumption),

$$\Delta\left(Z_{(R,L),l,R}\right) \geq 0 \text{ iff } a \leq \left[\frac{\lambda(2\alpha-\lambda)}{(\lambda+\alpha-1)} - 1\right]$$

2.

$$\begin{aligned}
Z_{(R,L),l,L}(x) &= (1-\lambda)\,x\,[\alpha(1-\lambda) + (1-\alpha)(1-x)] + \lambda(1-x)[\alpha(1-x) + (1-\alpha)(1-\lambda)] \\
&= \alpha(1-\lambda)^2 + (1-x)(1-\lambda)(1+\lambda-2\alpha) + (1-x)^2(\lambda+\alpha-1) \\
\Delta\left(Z_{(R,L),l,L}\right) &= (1-\lambda)(1+\lambda-2\alpha)(1-\sigma_{RR}-\sigma_{LL}) + (\lambda+\alpha-1)\left((1-\sigma_{RR})^2 - \sigma_{LL}^2\right) \\
&= b\,[(1-\lambda)(2\alpha-\lambda-1) - (\lambda+\alpha-1)(1-a)]
\end{aligned}$$

Hence

$$\Delta\left(Z_{(R,L),l,L}\right) \geq 0 \text{ iff } a \geq \left[1 - \frac{(1-\lambda)(2\alpha-\lambda-1)}{(\alpha+\lambda-1)}\right]$$

3.

$$\begin{aligned}
Z_{(R,L),r,R}(x) &= (1-\lambda)\,x\,[\alpha x + (1-\alpha)\lambda] + \lambda(1-x)[\alpha\lambda + (1-\alpha)\,x] \\
&= \lambda^2\alpha - x\lambda(2\alpha+\lambda-2) + x^2(\alpha-\lambda) \\
\Delta\left(Z_{(R,L),r,R}\right) &= -\lambda(2\alpha+\lambda-2)(\sigma_{RR}+\sigma_{LL}-1) + (\alpha-\lambda)\left(\sigma_{RR}^2 - (1-\sigma_{LL})^2\right) \\
&= b\,[-\lambda(2\alpha+\lambda-2) + (\alpha-\lambda)(1+a)]
\end{aligned}$$

Since α and λ greater than $\frac{2}{3}$ (by assumption), we have $(\lambda+2\alpha-2) \geq 0$. $(1+a) \geq 0$ because $a \in [-1,1]$. Therefore, if $\alpha \leq \lambda$ then $\Delta\left(Z_{(R,L),r,R}\right) \leq 0$. Otherwise, if $\alpha > \lambda$ then

$$\Delta\left(Z_{(R,L),r,R}\right) \geq 0 \text{ iff } a \geq \left[\frac{\lambda(2\alpha+\lambda-2)}{(\alpha-\lambda)} - 1\right]$$

4.

$$\begin{aligned}
Z_{(R,L),r,L}(x) &= (1-\lambda)\,x\,[\alpha(1-x)+(1-\alpha)(1-\lambda)] + \lambda(1-x)[\alpha(1-\lambda)+(1-\alpha)(1-x)] \\
&= (1-\lambda)^2(1-\alpha) + (1-x)(1-\lambda)(2\alpha+\lambda-1) - (1-x)^2(\alpha-\lambda) \\
\Delta\left(Z_{(R,L),r,L}\right) &= (1-\lambda)(2\alpha+\lambda-1)(1-\sigma_{RR}-\sigma_{LL}) - (\alpha-\lambda)\left((1-\sigma_{RR})^2 - \sigma_{LL}^2\right) \\
&= b\,[-(1-\lambda)(2\alpha+\lambda-1) + (\alpha-\lambda)(1-a)]
\end{aligned}$$

$(2\alpha+\lambda-1) > 0$ and $(1-a) \geq 0$ imply that, if $\alpha \leq \lambda$ then $\Delta\left(Z_{(R,L),r,L}\right) \leq 0$. Otherwise, if $\alpha > \lambda$ then

$$\Delta\left(Z_{(R,L),r,L}\right) \geq 0 \text{ iff } a \leq \left[1 - \frac{(1-\lambda)(2\alpha+\lambda-1)}{(\alpha-\lambda)}\right]$$

B Proof of Lemma 1

We start by identifying feasible partitions in the domain. The following argument shows that $c_1 < c_2$, for all values of α and λ.

For a fixed λ, c_1 is increasing in α and c_2 is decreasing in α.

$$\begin{aligned}
\frac{\partial c_1}{\partial \alpha} &= \frac{\lambda(3\lambda-2)}{(\alpha+\lambda-1)^2} > 0 \quad \text{because } \lambda > \frac{2}{3} \\
\frac{\partial c_1}{\partial \alpha} &= -\frac{(1-\lambda)(3\lambda-1)}{(\alpha+\lambda-1)^2} < 0
\end{aligned}$$

The highest value of c_1 (at $\alpha = 1$) is $(1-\lambda)$ and the lowest value of c_2 (at $\alpha = 1$) is $\left[1-\frac{(1-\lambda)^2}{\lambda}\right]$. Since $(1-\lambda) < \left[1-\frac{(1-\lambda)^2}{\lambda}\right]$, for all $\lambda > \frac{2}{3}$; we have $c_1 < c_2$. This additionally shows that $c_1 < 1$. c_1 is greater than -1 because $(2\alpha-\lambda) > (1-\lambda) > 0$ and hence $\frac{\lambda(2\alpha-\lambda)}{(\lambda+\alpha-1)}$ is positive. Hence $-1 < c_1 < 1$ for all values of α and λ.

When $\alpha \leq \lambda$, only two cut-off are relevant, c_1 and c_2. However, $\alpha \leq \lambda$ implies $(2\alpha-\lambda-1) \leq 0$ and hence $c_2 \geq 1$. Therefore, for $\alpha < \lambda$, only two intervals are feasible $[-1, c_1]$ and $[c_1, 1]$.

Let us now take $\alpha > \lambda$. Here all four cut-off are relevant. We already know that $c_1 < c_2$. The following argument establishes $c_4 < c_1 \leq c_3$.

$$\frac{\partial c_3}{\partial \alpha} = \frac{\lambda(2-3\lambda)}{(\alpha-\lambda)^2} < 0 \quad \text{because } \lambda > \frac{2}{3}$$
$$\frac{\partial c_4}{\partial \alpha} = -\frac{(1-\lambda)(1-3\lambda)}{(\alpha-\lambda)^2} > 0$$

c_3 is decreasing in α and is unbounded above as $\alpha \downarrow \lambda$. Its lowest value (at $\alpha = 1$) is $\left[\frac{\lambda^2}{(1-\lambda)} - 1\right]$. For $\lambda \geq \frac{2}{3}$, it is weakly greater than the highest value of c_1, which is $(1-\lambda)$. Hence $c_1 \leq c_3$. On the other hand c_4 is increasing in α and its highest value (at $\alpha = 1$) is $-\lambda$, which is strictly negative. The lowest value of c_1 (at $\alpha = \lambda$) is $\left[\frac{(1-\lambda)^2}{(2\lambda-1)}\right]$, which is (weakly) positive. Hence $c_4 < c_1$. Combining these observations, we can conclude that for $\alpha > \lambda$,

$$c_4 < 0 \leq c_1 \leq c_2 \text{ and } c_3$$

It is already clear that c_2 and c_3 can be greater than 1. c_4 can be smaller than -1 as it unbounded below as $\alpha \downarrow \lambda$. We can further identify the range of values of α and λ such that $c_2 \geq c_3$ (or the opposite) but that is redundant for what follows. It is only important to note that either is possible.

Now, we can proceed to computation of Δ.

1. $\alpha > \lambda$: First, take the interval $[c_1, \min\{c_2, c_3, 1\}]$. For the sake of easy reference, we call this the central interval. Recall that $c_4 < c_1$. For any $a \in [c_1, \min\{c_2, c_3, 1\}]$; c_1 and $c_4 \leq a \leq c_2$ and c_3. This imply $\Delta\left(Z_{(R,L),l,R}\right) \leq 0$, $\Delta\left(Z_{(R,L),l,L}\right) \leq 0$, $\Delta\left(Z_{(R,L),r,R}\right) \leq 0$ and $\Delta\left(Z_{(R,L),r,L}\right) \leq 0$. Hence,

   $$\Delta = \left[Z^L_{(R,L),l,R} + Z^L_{(R,L),l,L}\right] - \left[Z^L_{(R,L),r,R} + Z^L_{(R,L),r,L}\right] = 0$$

   because both the terms are equal to $P((R,L) \mid \text{state} = L)$.

   The neighbouring interval on the right is either $[c_2, \min\{c_3, 1\}]$ or $[c_3, \min\{c_2, 1\}]$. We calculate them separately. For $a \in [c_2, \min\{c_3, 1\}]$; $c_1, c_2, c_4 \leq a \leq c_3$. Compared to the central interval, only $\Delta\left(Z_{(R,L),l,L}\right)$ has changed from negative positive. Hence

$$\begin{aligned}\Delta &= \left[Z^L_{(R,L),l,R} + Z^R_{(R,L),l,L}\right] - \left[Z^L_{(R,L),r,R} + Z^L_{(R,L),r,L}\right] \\ &= \left[Z^L_{(R,L),l,R} + Z^L_{(R,L),l,L}\right] - \left[Z^L_{(R,L),r,R} + Z^L_{(R,L),r,L}\right] + \left[Z^R_{(R,L),l,L} - Z^L_{(R,L),l,L}\right] \\ &= 0 + \Delta\left(Z_{(R,L),l,L}\right) \\ &= b\left[(1-\lambda)(2\alpha-\lambda-1) - (\lambda+\alpha-1)(1-a)\right] \\ &= b\left[(\alpha-\lambda) - \lambda(2\alpha-\lambda) + (\alpha+\lambda-1)a\right]\end{aligned}$$

Similarly, for $a \in [c_3, \min\{c_2, 1\}]$

$$\begin{aligned}\Delta &= 0 - \Delta\left(Z_{(R,L),r,R}\right) \\ &= b\left[\lambda(2\alpha+\lambda-2) - (\alpha-\lambda)(1+a)\right]\end{aligned}$$

If $a \in [\max\{c_2, c_3\}, 1]$, then $a \geq c_1, c_2, c_3, c_4$. Once again, in reference to the central interval,

$$\begin{aligned}\Delta &= 0 + \Delta\left(Z_{(R,L),l,L}\right) - \Delta\left(Z_{(R,L),r,R}\right) \\ &= b\left[-2\lambda(1-\lambda) + (2\lambda-1)a\right]\end{aligned}$$

On the left side of the central interval, its neighbour is $[\max\{-1, c_4\}, c_1]$. For any a in this interval, $c_4 \leq a \leq c_1, c_2, c_3$. Therefore,

$$\begin{aligned}\Delta &= 0 + \Delta\left(Z_{(R,L),l,R}\right) \\ &= b\left[\lambda(2\alpha-\lambda) - (\alpha+\lambda-1)(a+1)\right]\end{aligned}$$

Finally, $a \in [-1, c_4]$ implies $a \leq c_1, c_2, c_3, c_4$. This interval is the mirror image of $a \in [\max\{c_2, c_3\}, 1]$. Hence,

$$\Delta = b\left[2\lambda(1-\lambda) - (2\lambda-1)a\right]$$

2. $\alpha \leq \lambda$: There are only two intervals. For $a \in [c_1, 1]$, $c_1 \leq a \leq c_2$. Thus $\Delta\left(Z_{(R,L),l,R}\right) \leq 0$, $\Delta\left(Z_{(R,L),l,L}\right) \leq 0$. We also know that for all a, $\Delta\left(Z_{(R,L),r,R}\right) \leq 0$ and $\Delta\left(Z_{(R,L),r,L}\right) \leq 0$. This case is identical to the central interval and $\Delta = 0$. For $a \in [-1, c_1]$, $a \leq c_1$ and c_2. This case is identical to $a \in [\max\{-1, c_4\}, c_1]$ under $\alpha > \lambda$. □

C Proof of Theorem 3

At $d = 1$, $\kappa_1 = 1$ and $\kappa_2 = 0$. Therefore maximizing Π is equivalent to

$$\max_{a\in[-1,1]} \quad \frac{(2\lambda-1-a)(1-|a|)\,g(\alpha,\lambda,a)}{(1-(2\lambda-1)\,a)}$$

To reduce notational burden we write $g(a)$, instead of $g(\alpha,\lambda,a)$, as long as α and λ are fixed.

• The maximum Π is strictly positive. We can show it by choosing a close to -1, say $a = -1+\epsilon$ where ϵ is a small positive number. Since a is negative, both $(2\lambda-1-a)$ and $(1-(2\lambda-1)a)$ are positive. $(1-|a|)$ is also positive $a \neq -1, 1$. As for g, there are three possibilities. If $\alpha > \lambda$, then either $a \in [-1, c_4]$ or $a \in [\max\{-1, c_4\}, c_1]$. If $\alpha \leq \lambda$ then $a \in [-1, c_1]$. These possibilities are evaluated below

$$g(a) = [2\lambda(1-\lambda)+(2\lambda-1)(1-\epsilon)] > 0 \quad \text{if } a \in [-1, c_4]$$
$$g(a) = \lambda(2\alpha-\lambda)-(\alpha+\lambda-1)\epsilon > 0 \quad \text{if } a \in [\max\{-1, c_4\}, c_1] \text{ or if } a \in [-1, c_1]$$

The first inequality holds because $\frac{1}{2} < \lambda \leq 1$. The second inequality holds because $2\alpha > 1 \geq \lambda$ and $\alpha+\lambda > 1$ and because ϵ can be chosen sufficiently small.

We shall now take one interval at a time and argue whether a maximum is possible in that interval or not. We first check it for $\alpha > \lambda$.

• Clearly maximum does not occur at the central interval, $[c_1, \min\{c_2, c_3, 1\}]$, because $g(a) = 0$.

• Take the interval $[c_2, \min\{c_3, 1\}]$. At the lower boundary of this interval $g(c_2) = 0$. g is increasing in a because its slope $(\alpha+\lambda-1) > 0$ (because $\alpha, \lambda \geq \frac{2}{3}$). Hence for all $a \in [c_2, \min\{c_3, 1\}]$, $g(a) \geq 0$. Among the other terms of the objective function, $(1-|a|)$ and $(1-(2\lambda-1)a)$ are positive (the latter is positive because both $a \leq 1$ and $(2\lambda-1) \leq 1$). However, $(2\lambda-1-a)$ is negative, as shown below. This makes the objective function negative in this interval and hence the maximum cannot occur here.

Recall that the lowest value of c_2 is $\left[1-\frac{(1-\lambda)^2}{\lambda}\right]$ (shown in Appendix B). Hence for any $a \in [c_2, \min\{c_3, 1\}]$, $a \geq \left[1-\frac{(1-\lambda)^2}{\lambda}\right]$. But,

$$(2\lambda-1)-\left(1-\frac{(1-\lambda)^2}{\lambda}\right) = \frac{(3\lambda-1)(\lambda-1)}{\lambda} \leq 0$$

The last inequality above, holds because $\frac{2}{3} \le \lambda \le 1$. Therefore $a \ge (2\lambda - 1)$.

• The same logic applies for the interval $[\max\{c_2, c_3\}, 1]$ when the $\max\{c_2, c_3\} = c_3$. In this case, the term $(2\lambda - 1 - a)$ is still negative as $a \ge c_3 \ge c_2$. $g(a)$ is also positive because at the lower boundary $g(c_3) > 0$ (we have previously argued that g is positive in $[c_2, c_3]$) and within the interval g is increasing in a. Hence the maximum does not occur in this interval.

• Next up is the interval $[c_3, \min\{c_2, 1\}]$. This case is slightly more involved as $g(a)$ is negative. This is so because at the lower boundary of this interval $g(a)$ is 0 and within the interval g is decreasing in a. Moreover, $(2\lambda - 1 - a)$ is also negative, because

$$a \ge c_3 \ge \left[\frac{\lambda^2}{(1-\lambda)} - 1\right] \ge (2\lambda - 1)$$

The second inequality was shown in Appendix B and the last inequality holds as

$$\left(\frac{\lambda^2}{1-\lambda} - 1\right) - (2\lambda - 1) = \frac{\lambda(3\lambda - 2)}{(1-\lambda)} \ge 0$$

Thus the objective function is now positive. However, we can still show that the maximum does not belong to this interval. To this end, we compare the value of our objective function at any $a \in [c_3, \min\{c_2, 1\}]$ with that at $-a$. We show below that the latter is greater.

The objective function has $(1 - |a|)$ in the numerator, which is the same at a and at $-a$. It remains to show that

$$\frac{(2\lambda - 1 - a)\, g(a)}{(1 - (2\lambda - 1)a)} < \frac{(2\lambda - 1 + a)\, g(-a)}{(1 + (2\lambda - 1)a)}$$

Alternatively,

$$\left[\frac{(a - 2\lambda + 1)}{(1 - (2\lambda - 1)a)}\right]\big(-g(a)\big) < \left[\frac{(2\lambda - 1 + a)}{(1 + (2\lambda - 1)a)}\right] g(-a)$$

Note that all four terms in the above inequality is positive. In particular, $g(-a)$ is positive because $g(c_1) = 0$ and g is decreasing in a for all $a < c_1$. Below, we show that the first term (resp. second term) on RHS is greater than the first term (second term) on LHS.

$$\left[\frac{(2\lambda - 1 + a)}{(1 + (2\lambda - 1)a)}\right] - \left[\frac{(a - 2\lambda + 1)}{(1 - (2\lambda - 1)a)}\right] = \frac{2(2\lambda - 1)\left(1 - a^2\right)}{(1 + (2\lambda - 1)a)(1 - (2\lambda - 1)a)} > 0$$

Next, we show that the first second term on RHS is greater than the second term on LHS. $g(-a)$ depends on whether $-a \in [\max\{-1, c_4\}, c_1]$ or $-a \in [-1, c_4]$. We can compute the former with reference to $g(c_1)$

$$\begin{aligned} &g(-a) - g(c_1) = (\alpha + \lambda - 1)(a + c_1) \\ \Rightarrow &g(-a) = (\alpha + \lambda - 1)(a + c_1) \quad \text{because } g(c_1) = 0 \end{aligned}$$

and the later with respect to $g(c_4)$

$$\begin{aligned} &g(-a) - g(c_4) = (2\lambda - 1)(a + c_4) \\ \Rightarrow &g(-a) = (2\lambda - 1)(a + c_4) + g(c_4) \\ \Rightarrow &g(-a) = (2\lambda - 1)(a + c_4) + (\alpha + \lambda - 1)(-c_4 + c_1) + g(c_1) \\ \Rightarrow &g(-a) = (2\lambda - 1)(a + c_4) + (\alpha + \lambda - 1)(-c_4 + c_1) \quad \text{because } g(c_1) = 0 \end{aligned}$$

Since $\alpha \geq \lambda$, under both the cases, $g(-a) \geq (2\lambda - 1)(c_1 + a)$. On the other hand $-g(a)$ can be computed with reference to $g(c_3)$.

$$\begin{aligned} &g(c_3) - g(a) = (\alpha - \lambda)(a - c_3) \\ \Rightarrow &- g(a) = (\alpha - \lambda)(a - c_3) \quad \text{because } g(c_3) = 0 \end{aligned}$$

Now $(\alpha - \lambda) \leq (2\lambda - 1)$ because $3\lambda \geq 2 \geq 1 + \alpha$ and $(a - c_3) < (c_1 + a)$ because both c_1 and c_3 are positive (see Appendix B). Hence $g(-a) > -g(a)$.

• Next, take the interval $[\max\{c_2, c_3\}, 1]$ when the $\max\{c_2, c_3\} = c_2$. In case $g(a) > 0$ for some a in this interval, it cannot give the maximum because the value of the objective function at this point is negative (similar to the case $\max\{c_2, c_3\} = c_3$). On the other hand, if $g(a) \leq 0$, then we can replicate the proof used for $[c_3, \min\{c_2, 1\}]$.

• To sum up, then the maximum must belong to either $[\max\{-1, c_4\}, c_1]$ or $[-1, c_4]$. We can further rule out a part of the former, namely, $(0, c_1]$ (recall that, $c_1 \geq 0$, when $\alpha > \lambda$). The argument is as follows. In this interval $g(a)$ decreasing in a. It is also positive, because at the upper bound of the interval, c_1, $g(c_1) = 0$. Lets check that other terms are also positive and deceasing in a. $(1 - |a|) = (1 - a)$ is positive and decreasing in a. Finally, $\left[\frac{(2\lambda - 1 - a)}{(1 - (2\lambda - 1)a)}\right]$ is positive because $a \leq c_1 \leq (1 - \lambda) \leq (2\lambda - 1)$. In Appendix B, we showed that $(1 - \lambda)$ is the highest value of c_1 and the last inequality holds as $\lambda \geq \frac{2}{3}$. $\left[\frac{(2\lambda - 1 - a)}{(1 - (2\lambda - 1)a)}\right]$ is also decreasing in a because

$$\frac{\partial}{\partial a}\left[\frac{(2\lambda - 1 - a)}{(1 - (2\lambda - 1)a)}\right] = \left[\frac{(2\lambda - 1)^2 - 1}{(1 - (2\lambda - 1)a)^2}\right] \leq 0$$

The inequality holds because $(2\lambda-1)\leq 1$. Therefore the objective function is positive and decreasing in a, in the interval $(0,c_1]$. This rules out maximum occurring in $(0,c_1]$.

• To proceed further, we need to differentiate the objective function with respect to a. Note that we have ruled out $a>0$, hence we shall take $a\leq 0$. To save on notations, denote

$$\begin{aligned}
\nu_1(a) &= (2\lambda-1-a)\\
\nu_2(a) &= (1-|a|) = (1+a)\\
\nu_3(a) &= g(a)\\
\nu_4(a) &= (1-(2\lambda-1)a)
\end{aligned}$$

Differentiating with respect to a, we get

$$\left[\frac{\nu_1(a)\,\nu_2(a)\,\nu_3(a)}{\nu_4(a)}\right]\left[\frac{\frac{\partial \nu_1(a)}{\partial a}}{\nu_1(a)}+\frac{\frac{\partial \nu_2(a)}{\partial a}}{\nu_2(a)}+\frac{\frac{\partial \nu_3(a)}{\partial a}}{\nu_3(a)}-\frac{\frac{\partial \nu_4(a)}{\partial a}}{\nu_4(a)}\right]$$

The first term is simply the value of the objective function. This is weakly positive because for $-1\leq a\leq 0$, because each of $\nu_1(a), \nu_2(a), \nu_3(a), \nu_4(a)$ is weakly positive. Thus the sign of first derivative of objective function is the same as that of

$$\left[\frac{\frac{\partial \nu_1(a)}{\partial a}}{\nu_1(a)}+\frac{\frac{\partial \nu_2(a)}{\partial a}}{\nu_2(a)}+\frac{\frac{\partial \nu_3(a)}{\partial a}}{\nu_3(a)}-\frac{\frac{\partial \nu_4(a)}{\partial a}}{\nu_4(a)}\right] \tag{6}$$

In fact, at the maximum, the above expression should be equal to 0, because the maximum value of the objective function is strictly positive.

• Now we argue that the maximum does not occur in the interval $[-1,c_4]$. To this end, we show that Expression 6 is strictly positive for all a in the interval.

$$\begin{aligned}
&\left[\frac{\frac{\partial \nu_1(a)}{\partial a}}{\nu_1(a)}+\frac{\frac{\partial \nu_2(a)}{\partial a}}{\nu_2(a)}+\frac{\frac{\partial \nu_3(a)}{\partial a}}{\nu_3(a)}-\frac{\frac{\partial \nu_4(a)}{\partial a}}{\nu_4(a)}\right]\\
&\quad=\left[-\frac{1}{(2\lambda-1-a)}+\frac{1}{(1+a)}-\frac{(2\lambda-1)}{(2\lambda(1-\lambda)-(2\lambda-1)a)}+\frac{(2\lambda-1)}{(1-(2\lambda-1)a)}\right]\\
&\quad=\left[\frac{2\lambda}{(1+a)(1-(2\lambda-1)a)}\right]-\left[\frac{1}{(2\lambda-1-a)}+\frac{(2\lambda-1)}{(2\lambda(1-\lambda)-(2\lambda-1)a)}\right]
\end{aligned}$$

The denominator of the first term above is increasing in a, because

$$\frac{\partial\left[(1+a)(1-(2\lambda-1)a)\right]}{\partial a}=2(1-\lambda)-2(2\lambda-1)a>0$$

Thus the first term is decreasing in a, while the second and the last term are increasing in a. Hence Expression 6 is decreasing in a. To complete our argument that Expression 6 is strictly positive, we now show that it is positive at $a=-\frac{1}{2}$. This is sufficient as the highest value of c_4 is $-\lambda$ (see Appendix *B*), which in turn, is smaller than $-\frac{1}{2}$. Expression 6 at $a=-\frac{1}{2}$ is

$$\begin{aligned}
&\left[\frac{2\lambda}{\frac{1}{2}\left(1+\frac{(2\lambda-1)}{2}\right)}\right]-\left[\frac{1}{\left(2\lambda-\frac{1}{2}\right)}+\frac{(2\lambda-1)}{\left(2\lambda(1-\lambda)+\frac{(2\lambda-1)}{2}\right)}\right]\\
&=\left[\frac{8\lambda}{(2\lambda+1)}-\frac{2}{(4\lambda-1)}\right]-\left[\frac{2(2\lambda-1)}{(4\lambda(1-\lambda)+(2\lambda-1))}\right]\\
&=\left[\frac{2(2\lambda-1)(8\lambda+1)}{(2\lambda+1)(4\lambda-1)}\right]-\left[\frac{2(2\lambda-1)}{(4\lambda(1-\lambda)+(2\lambda-1))}\right]\\
&=\frac{2(2\lambda-1)}{(2\lambda+1)(4\lambda-1)[4\lambda(1-\lambda)+(2\lambda-1)]}\Big[(8\lambda+1)[4\lambda(1-\lambda)+(2\lambda-1)]-(2\lambda+1)(4\lambda-1)\Big]\\
&=\frac{2(2\lambda-1)}{(2\lambda+1)(4\lambda-1)[4\lambda(1-\lambda)+(2\lambda-1)]}\Big[-32\lambda^3+36\lambda^2-4\lambda\Big]\\
&=\frac{8(2\lambda-1)\lambda(1-\lambda)(8\lambda-1)}{(2\lambda+1)(4\lambda-1)[4\lambda(1-\lambda)+(2\lambda-1)]}
\end{aligned}$$

This is positive because every term of the above expression is positive.

• Since we have ruled out everything else, the maximum is obtained in the interval $[\max\{-1,c_4\},0]$. The maximum is given implicitly by the FOC

$$\begin{aligned}
&\left[-\frac{1}{(2\lambda-1-a)}+\frac{1}{(1+a)}-\frac{(\alpha+\lambda-1)}{(\lambda(2\alpha-\lambda)-(\alpha+\lambda-1)(a+1))}+\frac{(2\lambda-1)}{(1-(2\lambda-1)a)}\right]=0\\
\Rightarrow&\left[\frac{2\lambda}{(1+a)(1-(2\lambda-1)a)}\right]-\left[\frac{1}{(2\lambda-1-a)}+\frac{(\alpha+\lambda-1)}{(\lambda(2\alpha-\lambda)-(\alpha+\lambda-1)(a+1))}\right]=0
\end{aligned}$$

We have already shown that the first term is decreasing in a. the second and the last term are increasing in a. Therefore the LHS is decreasing in a and the maximum is unique. This completes the case $\alpha>\lambda$.

• For $\alpha\le\lambda$ there are only two intervals. The value of the objective function is 0 for $[c_1,1]$. Hence the maximum must belong to the interval $[-1,c_1]$. In case

$c_1 > 0$, we can also rule out a maximum in $(0, c_1]$ by following exactly the same argument as in the case of $\alpha > \lambda$. Hence the maximum belongs to the interval $[-1, \min\{c_1, 0\}]$. The FOC is the same as in the case of $\alpha > \lambda$. □

D Proof of Theorem 5

We first check the lower bound: in equilibrium $a \geq -\lambda$. Recall that the First order condition, LHS of Equation 4 is decreasing in a (proved in Appendix C). Hence, if we can show that the LHS of Equation 4 is positive at $a = -\lambda$ then we are done. Value of the LHS at $a = -\lambda$ is

$$\left[\frac{1}{(1-\lambda)} + \frac{(2\lambda-1)}{(1+(2\lambda-1)\lambda)}\right] - \left[\frac{1}{(3\lambda-1)} + \frac{(\alpha+\lambda-1)}{(\lambda(2\alpha-\lambda)-(\alpha+\lambda-1)(1-\lambda))}\right]$$

The sum of the first two terms is greater than 3, because $\frac{1}{(1-\lambda)} \geq 3$ (as $\lambda \geq \frac{2}{3}$) and the second term is positive. Let us now put upper bound on the last two terms. Since $\lambda \geq \frac{2}{3}$; the first of these, $\left[\frac{1}{(3\lambda-1)}\right] \leq 1$. We show that the last term has upper bound 2, which will complete the proof. If the denominator of this term is negative, then it is trivially true, because the numerator is positive. Hence we assume that the denominator is positive.

$$2 - \left[\frac{(\alpha+\lambda-1)}{\lambda(2\alpha-\lambda)-(\alpha+\lambda-1)(1-\lambda)}\right] = \frac{[2\lambda(2\alpha-\lambda)-(\alpha+\lambda-1)(3-2\lambda)]}{[\lambda(2\alpha-\lambda)-(\alpha+\lambda-1)(1-\lambda)]}$$

Rearranging the numerator, we obtain,

$$\left[3\alpha(2\lambda-1)+(1-\lambda)(3-2\lambda)-2\lambda^2\right]$$

which is increasing in α. It is sufficient to show that this is positive for smallest α, that is at $\alpha = \frac{2}{3}$.

$$\left[2(2\lambda-1)+(1-\lambda)(3-2\lambda)-2\lambda^2\right] = (1-\lambda) \geq 0$$

This completes the proof for lower bound. Next, we verify the new upper bound. It has two sub-cases

- $\alpha > 2\lambda - 1$: Here we show that LHS of Equation 4 is negative at $a = -(1-\lambda)$, which will imply that the equilibrium value of a must be smaller than $-(1-\lambda)$ (because LHS of Equation 4 is decreasing in a). Value of the LHS at $a = -(1-\lambda)$ is

$$\left[\frac{2}{\lambda(3-2\lambda)}\right] - \left[\frac{1}{\lambda} + \frac{(\alpha+\lambda-1)}{\lambda(\alpha-2\lambda+1)}\right] = \left[\frac{2}{\lambda(3-2\lambda)}\right] - \left[\frac{(2\alpha-\lambda)}{\lambda(\alpha-2\lambda+1)}\right]$$
$$= \left[\frac{\left(2-\lambda-2\lambda^2\right)-4\alpha(1-\lambda)}{\lambda(3-2\lambda)(\alpha-2\lambda+1)}\right]$$

Since $\alpha > 2\lambda - 1$, the denominator is positive. The numerator is decreasing in α. Therefore

$$\left(2-\lambda-2\lambda^2\right)-4\alpha(1-\lambda) \leq \left(2-\lambda-2\lambda^2\right)-4(2\lambda-1)(1-\lambda)$$
$$= \left(6\lambda^2-13\lambda+6\right) = 6(1-\lambda)^2-\lambda$$

Since $6(1-\lambda)^2$ is decreasing in λ, its upper bound is obtained at $\lambda = \frac{2}{3}$. Thus

$$6(1-\lambda)^2 \leq 6\left(1-\frac{2}{3}\right)^2 = \frac{2}{3} \leq \lambda$$

Hence he numerator is negative and consequently LHS of Equation 4 is negative.

- $\alpha \leq 2\lambda - 1$: Since $2\lambda - 1 \leq \lambda$, we also have $\alpha \leq \lambda$. Theorem 3, states that equilibrium a is less than $\min\{c_1, 0\}$. If we can show that, for this sub-case, $c_1 \leq -(1-\lambda)$, then we are done because it would mean $\min\{c_1, 0\} = \min\{c_1, -(1-\lambda)\} = c_1$. We show below that indeed, $c_1 \leq -(1-\lambda)$. Recall that $c_1 = \frac{\lambda(2\alpha-\lambda)}{(\lambda+\alpha-1)} - 1$ and it is increasing in α (Proved in Appendix *B*). Hence for this sub-case, an upper bound of c_1 is obtained by computing c_1 at $\alpha = (2\lambda - 1)$.

$$c_1 \leq \frac{\lambda(2(2\lambda-1)-\lambda)}{(\lambda+(2\lambda-1)-1)} - 1 = -(1-\lambda)$$

□

E Proof of Theorem 8

In Appendix *A* we have shown

$$Z_{(R,L),l,R}(x) = \lambda^2(1-\alpha) + x\lambda(2\alpha-\lambda) - x^2(\alpha+\lambda-1)$$

To compute $Z^R_{(R,L),l,R}$ and $Z^L_{(R,L),l,R}$, we evaluate $Z_{(R,L),l,R}(x)$ at $x=\sigma_{RR}$ and $x=1-\sigma_{LL}$ respectively and replace equilibrium value of σ; $\sigma_{LL}=1$ and $\sigma_{RR}=(1+a)$.

$$\begin{aligned}
Z^R_{(R,L),l,R}+Z^L_{(R,L),l,R} &= 2\lambda^2(1-\alpha)+\lambda(2\alpha-\lambda)(\sigma_{RR}+1-\sigma_{LL})-(\alpha+\lambda-1)\left(\sigma_{RR}^2+(1-\sigma_{LL})^2\right)\\
&= 2\lambda^2(1-\alpha)+\lambda(2\alpha-\lambda)(1+a)-(\alpha+\lambda-1)(1+a)^2
\end{aligned}$$

We can also rearrange the above as

$$\begin{aligned}
\left[Z^R_{(R,L),l,R}+Z^L_{(R,L),l,R}\right] &= \lambda^2(1-a)-\alpha\lambda^2-\alpha\left(\lambda^2-2\lambda(1+a)+(1+a)^2\right)+(1-\lambda)(1+a)^2\\
&= \lambda^2(1-a)-\alpha\lambda^2-\alpha(\lambda-(1+a))^2+(1-\lambda)(1+a)^2
\end{aligned}$$

Partials are

$$\begin{aligned}
\frac{\partial\left[Z^R_{(R,L),l,R}+Z^L_{(R,L),l,R}\right]}{\partial\alpha} &= -2\lambda^2+2\lambda(1+a)-(1+a)^2=-\left[\lambda^2+(\lambda-(1+a))^2\right]\\
\frac{\partial\left[Z^R_{(R,L),l,R}+Z^L_{(R,L),l,R}\right]}{\partial a} &= \lambda(2\alpha-\lambda)-2(\alpha+\lambda-1)(1+a)
\end{aligned}$$

Now we take partials of γ^*. Note that γ^* can be written as $\frac{\pi}{(2\lambda-1-a)}$, where π is the objective function in Equation 3. In Theorem 3, we have also shown that the maximization problem is solved when either $a\in[\max\{-1,c_4\},0]$ or $a\in[-1,\min\{c_1,0\}]$. In both case, the corresponding $g(\alpha,\lambda,a)=[\lambda(2\alpha-\lambda)-(\alpha+\lambda-1)(a+1)]$. Thus

$$\pi=\frac{(2\lambda-1-a)(1+a)[\lambda(2\alpha-\lambda)-(\alpha+\lambda-1)(a+1)]}{(1-(2\lambda-1)a)}$$

and

$$\begin{aligned}
\frac{\partial\left[\frac{\pi}{(2\lambda-1-a)}\right]}{\partial\alpha} &= \frac{1}{(2\lambda-1-a)}\frac{\partial\pi}{\partial\alpha}=\frac{(2\lambda-1-a)(1+a)}{(1-(2\lambda-1)a)}\\
\frac{\partial\left[\frac{\pi}{(2\lambda-1-a)}\right]}{\partial a} &= \frac{(2\lambda-1-a)\frac{\partial\pi}{\partial a}+\pi}{(2\lambda-1-a)^2}=\frac{\pi}{(2\lambda-1-a)^2}=\frac{\gamma^*}{(2\lambda-1-a)}
\end{aligned}$$

The last equality uses the fact that $\frac{\partial\pi}{\partial a}=0$ at the equilibrium. We show below $\frac{\partial\gamma^*}{\partial\alpha}$ is positive but smaller than $(1+a)$. It is positive because at the equilibrium

$-1 < a < 0$ (see Theorem 5) and $2\lambda > 1$. It is less than $(1+a)$ because

$$(2\lambda-1-a)-(1-(2\lambda-1)a)=2(\lambda-1)(1+a)\leq 0 \quad \text{as } \lambda\leq 1$$
$$\Rightarrow \frac{(2\lambda-1-a)}{(1-(2\lambda-1)a)}\leq 1$$

Therefore

$$\begin{aligned}\left[Z^R_{(R,L),l,R}+Z^L_{(R,L),l,R}\right]\frac{\partial\gamma^*}{\partial\alpha} &= \left[\lambda^2(1-a)-\alpha\lambda^2-\alpha(\lambda-(1+a))^2+(1-\lambda)(1+a)^2\right]\frac{\partial\gamma^*}{\partial\alpha}\\ &= \left[\lambda^2(1-a)\right]\frac{\partial\gamma^*}{\partial\alpha}+\left[(1-\lambda)(1+a)^2-\alpha\lambda^2-\alpha(\lambda-(1+a))^2\right]\frac{\partial\gamma^*}{\partial\alpha}\end{aligned}$$

Since $\frac{\partial\gamma^*}{\partial\alpha}$ is less than $(1+a)$, the first term is smaller than $\lambda^2\left(1-a^2\right)$ which is less than λ^2. The second term is negative because $\frac{\partial\gamma^*}{\partial\alpha}$ is positive and

$$\left[(1-\lambda)(1+a)^2-\alpha\lambda^2-\alpha(\lambda-(1+a))^2\right]\leq(1-\lambda)\lambda^2-\alpha\lambda^2-\alpha(\lambda-(1+a))^2<0$$

The first inequality follows because $a\leq-(1-\lambda)$ by Theorem 5. The second inequality is true because $\alpha+\lambda>1$. Hence

$$\left[Z^R_{(R,L),l,R}+Z^L_{(R,L),l,R}\right]\frac{\partial\gamma^*}{\partial\alpha}\leq\lambda^2 \tag{7}$$

On the other hand

$$\left(1+\gamma^*\right)\frac{\partial\left[Z^R_{(R,L),l,R}+Z^L_{(R,L),l,R}\right]}{\partial\alpha}=-\left(\lambda^2+(\lambda-(1+a))^2\right)\left(1+\gamma*\right)\leq-\lambda^2 \tag{8}$$

because γ^* is positive by Corollary 4. Combining Equations 7 and 8,

$$\frac{\partial V_R}{\partial\alpha}=\frac{1}{2}\left[\left(1+\gamma^*\right)\frac{\partial\left[Z^R_{(R,L),l,R}+Z^L_{(R,L),l,R}\right]}{\partial\alpha}+\left[Z^R_{(R,L),l,R}+Z^L_{(R,L),l,R}\right]\frac{\partial\gamma^*}{\partial\alpha}\right]\leq 0$$

Let's check $\frac{\partial V_R}{\partial a}$,

$$\begin{aligned}\frac{\partial V_R}{\partial a} &= \left(1+\gamma^*\right)\frac{\partial\left[Z^R_{(R,L),l,R}+Z^L_{(R,L),l,R}\right]}{\partial a}+\lambda\\ &= \left(1+\gamma^*\right)[\lambda(2\alpha-\lambda)-2(\alpha+\lambda-1)(1+a)]+\lambda\\ &\geq [\lambda(2\alpha-\lambda)-2(\alpha+\lambda-1)(1+a)]+\lambda \quad [\text{because } \gamma^*>0]\\ &\geq \lambda(2\alpha-\lambda)-2(\alpha+\lambda-1)\lambda+\lambda \quad [\text{because } a\leq-(1-\lambda)]\\ &= \lambda(3-3\lambda)>0\end{aligned}$$

and

$$\left[Z^R_{(R,L),l,R} + Z^L_{(R,L),l,R}\right] \frac{\partial \gamma^*}{\partial a} = \left[Z^R_{(R,L),l,R} + Z^L_{(R,L),l,R}\right] \frac{\gamma^*}{(2\lambda - 1 - a)} > 0$$

because $\left[Z^R_{(R,L),l,R} + Z^L_{(R,L),l,R}\right]$ is a probability measure and $\gamma^* > 0$. Therefore

$$\frac{\partial V_R}{\partial a} = \frac{1}{2}\left[(1+\gamma^*)\frac{\partial\left[Z^R_{(R,L),l,R} + Z^L_{(R,L),l,R}\right]}{\partial a} + \left[Z^R_{(R,L),l,R} + Z^L_{(R,L),l,R}\right]\frac{\partial \gamma^*}{\partial a} + \lambda\right] > 0$$

The total effect can be both positive and negative. In Theorem 6, we have shown that

$$\frac{da}{d\alpha} = -\frac{\frac{\lambda(3\lambda-2)}{[\lambda(2\alpha-\lambda)-(\alpha+\lambda-1)(a+1)]^2}}{\frac{\partial \phi}{\partial a}} \geq 0$$

where ϕ is the LHS of Equation 4. At $\lambda = \frac{2}{3}$, the numerator of $\frac{da}{d\alpha}$ is 0. Hence the total effect on V_R,

$$\frac{dV_R}{d\alpha} = \frac{\partial V_R}{\partial \alpha} + \frac{\partial V_R}{\partial a}\frac{da}{d\alpha} = \frac{\partial V_R}{\partial \alpha} < 0$$

Hence for sufficiently small α, $\frac{dV_R}{d\alpha}$ is negative. We show that the total effect is strictly positive at $\lambda = 1$. At $\lambda = 1$, the equilibrium value of a can be explicitly computed from Equation 4

$$\left[\frac{2}{(1+a)(1-a)}\right] - \left[\frac{1}{(1-a)} + \frac{\alpha}{((2\alpha-1)-\alpha(a+1))}\right] = 0 \Leftrightarrow a = -\frac{1}{2\alpha}$$

Substituting a

$$\begin{aligned}
Z^R_{(R,L),l,R} + Z^L_{(R,L),l,R} &= 2(1-\alpha) + (2\alpha-1)\left(1 - \frac{1}{2\alpha}\right) - \alpha\left(1 - \frac{1}{2\alpha}\right)^2 \\
&= 2(1-\alpha) + \frac{(2\alpha-1)^2}{4\alpha} \\
&= (1-\alpha) + \frac{1}{4\alpha}
\end{aligned}$$

and

$$\begin{aligned}\gamma^* &= \frac{(1+a)[\lambda(2\alpha-\lambda)-(\alpha+\lambda-1)(a+1)]}{(1-(2\lambda-1)a)} \\ &= \frac{\left(1-\frac{1}{2\alpha}\right)\left[(2\alpha-1)-\alpha\left(1-\frac{1}{2\alpha}\right)\right]}{\left(1+\frac{1}{2\alpha}\right)} \\ &= \frac{(2\alpha-1)^2}{2(2\alpha+1)}\end{aligned}$$

By Equation 5

$$V_R = \left[1+\frac{(2\alpha-1)^2}{2(2\alpha+1)}\right]\left[\frac{1}{2}\left((1-\alpha)+\frac{1}{4\alpha}\right)\right]+\left(1-\frac{1}{2\alpha}\right)$$

Then

$$\begin{aligned}\frac{dV_R}{d\alpha} &= \frac{1}{2}\left((1-\alpha)+\frac{1}{4\alpha}\right)\left(\frac{(2\alpha+3)(2\alpha-1)}{(2\alpha+1)^2}\right)-\frac{1}{2}\left(\frac{1}{4\alpha^2}+1\right)\left(1+\frac{(2\alpha-1)^2}{2(2\alpha+1)}\right)+\frac{1}{2\alpha^2} \\ &= \underbrace{\frac{1}{2}\left((1-\alpha)+\frac{1}{4\alpha}\right)\left(\frac{(2\alpha+3)(2\alpha-1)}{(2\alpha+1)^2}\right)}_{>0}+\underbrace{\frac{1}{2}\left[\frac{1}{\alpha^2}-\frac{1}{2}\left(\frac{1}{4\alpha^2}+1\right)\left(1+\frac{(2\alpha-1)^2}{2(2\alpha+1)}\right)\right]}\end{aligned}$$

The first term is positive because $\frac{2}{3} \le \alpha \le 1$. We show that the second under-braced term is also positive

$$\left[\frac{1}{\alpha^2}-\frac{1}{2}\left(\frac{1}{4\alpha^2}+1\right)\left(1+\frac{(2\alpha-1)^2}{2(2\alpha+1)}\right)\right] = \frac{\left[16(2\alpha+1)-\left(1+4\alpha^2\right)\left(4\alpha^2+3\right)\right]}{16\alpha^2(2\alpha+1)}$$

The numerator is positive because $16(2\alpha+1) \ge \frac{112}{3}$ (as $\alpha \ge \frac{2}{3}$) and $\left(1+4\alpha^2\right)\left(4\alpha^2+3\right) \le 35$ (as $\alpha \le 1$). Hence $\frac{dV_R}{d\alpha} > 0$ when $\lambda = 1$ or by continuity for λ sufficiently close to 1. □

F Proof of Theorem 9

Take the history where first stage reports are (R, L), voter i has endorsed L, and second stage personalized report to i is R. Probability of this history given true state R is

$$Z^R_{(R,L),l,R} = (1-\lambda)\sigma_{RR}[\alpha\lambda+(1-\alpha)]+\lambda(1-\sigma_{RR})[\alpha+(1-\alpha)\lambda]$$

Here $Z^R_{(R,L),l,R}$ has been adjusted for truth telling in the second stage. Similar expression can be computed for all other histories.

$$\begin{aligned}
\Delta\left(Z_{(R,L),l,R}\right) &= b\lambda(\alpha-\lambda)+[(1-\alpha)(1-\lambda)\sigma_{RR}+\alpha\lambda(1-\sigma_{RR})] \\
\Delta\left(Z_{(R,L),l,L}\right) &= b(1-\lambda)(\alpha-\lambda)-[(1-\alpha)(1-\lambda)(1-\sigma_{LL})+\alpha\lambda\sigma_{LL}] \\
\Delta\left(Z_{(R,L),r,R}\right) &= b\lambda(1-\alpha-\lambda)+[\alpha(1-\lambda)\sigma_{RR}+(1-\alpha)\lambda(1-\sigma_{RR})] \\
\Delta\left(Z_{(R,L),r,L}\right) &= b(1-\lambda)(1-\alpha-\lambda)-[\alpha(1-\lambda)(1-\sigma_{LL})+(1-\alpha)\lambda\sigma_{LL}]
\end{aligned}$$

We can immediately see that $\Delta\left(Z_{(R,L),r,L}\right)$ is negative because $b \geq 0$ and $(1-\alpha-\lambda) < 0$. If $\alpha \leq \lambda$ then $\Delta\left(Z_{(R,L),l,L}\right)$ is also trivially negative. We show below that $\Delta\left(Z_{(R,L),l,L}\right)$ is negative even for $\alpha > \lambda$. Since $b = \sigma_{RR}+\sigma_{LL}-1 \leq \sigma_{LL}$, we have

$$\begin{aligned}
\Delta\left(Z_{(R,L),l,L}\right) &\leq (1-\lambda)(\alpha-\lambda)\sigma_{LL}-[(1-\alpha)(1-\lambda)(1-\sigma_{LL})+\alpha\lambda\sigma_{LL}] \\
&= \sigma_{LL}[(1-\lambda)(\alpha-\lambda)+(1-\alpha)(1-\lambda)-\alpha\lambda]-(1-\alpha)(1-\lambda).
\end{aligned}$$

If co-efficient of σ_{LL} is negative then we are done. So lets take the the co-efficient to be positive. Then

$$\begin{aligned}
\Delta\left(Z_{(R,L),l,L}\right) &\leq [(1-\lambda)(\alpha-\lambda)+(1-\alpha)(1-\lambda)-\alpha\lambda]-(1-\alpha)(1-\lambda) \\
&= (1-\lambda)(\alpha-\lambda)-\alpha\lambda \leq 0.
\end{aligned}$$

The first inequality above follows from $\sigma_{LL} \leq 1$ and the second inequality follows from $(1-\lambda) < \lambda$ and $(\alpha-\lambda) < \alpha$.
On the other hand, it is immediate that $\Delta\left(Z_{(R,L),l,R}\right) \geq 0$ for $\alpha \geq \lambda$. Next we show that $\Delta\left(Z_{(R,L),r,R}\right) \geq 0 \Rightarrow \Delta\left(Z_{(R,L),l,R}\right) \geq 0$ when $\lambda > \alpha$. This will be easier to see if we rewrite $\Delta\left(Z_{(R,L),r,R}\right)$ and $\Delta\left(Z_{(R,L),l,R}\right)$ in terms of σ_{RR} and σ_{LL}.

$$\begin{aligned}
\Delta\left(Z_{(R,L),l,R}\right) &= \sigma_{LL}\lambda(\alpha-\lambda)+\sigma_{RR}\left[(1-\alpha)(1-\lambda)-\lambda^2\right]+\lambda^2 \\
\Delta\left(Z_{(R,L),r,R}\right) &= \sigma_{LL}\lambda(1-\alpha-\lambda)+\sigma_{RR}\left[\alpha(1-\lambda)-\lambda^2\right]+\lambda^2.
\end{aligned}$$

Note that coefficients of σ_{LL} are negative in the above expressions. Coefficients of σ_{RR} are also negative because $\lambda > \alpha > (1-\alpha)$ and $\lambda > (1-\lambda)$.

Therefore for a fixed σ_{RR},

$$\begin{aligned}\Delta\left(Z_{(R,L),r,R}\right)\geq 0 \Rightarrow \sigma_{LL} &\leq \frac{\lambda^2(1-\sigma_{RR})}{\lambda(\lambda+\alpha-1)}+\frac{\alpha(1-\lambda)\sigma_{RR}}{\lambda(\lambda+\alpha-1)}\\ &\leq \frac{\lambda^2(1-\sigma_{RR})}{\lambda(\lambda-\alpha)}+\frac{\alpha(1-\lambda)\sigma_{RR}}{\lambda(\lambda+\alpha-1)} \text{ because } \lambda-\alpha\leq\lambda+\alpha-1\\ &\leq \frac{\lambda^2(1-\sigma_{RR})}{\lambda(\lambda-\alpha)}+\frac{(1-\alpha)(1-\lambda)\sigma_{RR}}{\lambda(\lambda-\alpha)} \Rightarrow \Delta\left(Z_{(R,L),l,R}\right)\geq 0.\end{aligned}$$

The last inequality holds because

$$\frac{\alpha}{\lambda+\alpha-1}-\frac{1-\alpha}{\lambda-\alpha}=\frac{(2\alpha-1)(\lambda-1)}{(\lambda+\alpha-1)(\lambda-\alpha)}\leq 0.$$

Let us now compile these inequalities into the following possibilities:

1. All four $\Delta(.)$ are negative. In this case

$$\Delta=\left[Z^L_{(R,L),l,R}+Z^L_{(R,L),l,L}\right]-\left[Z^L_{(R,L),r,R}+Z^L_{(R,L),r,L}\right]=0$$

   because both the terms are equal to $P((R,L)\mid \text{state}=L)$.

2. $\Delta\left(Z_{(R,L),l,L}\right)$, $\Delta\left(Z_{(R,L),r,L}\right)$, and $\Delta\left(Z_{(R,L),r,R}\right)$ are negative but $\Delta\left(Z_{(R,L),l,R}\right)$ is positive. Here

$$\begin{aligned}\Delta &=\left[Z^R_{(R,L),l,R}+Z^L_{(R,L),l,L}\right]-\left[Z^L_{(R,L),r,R}+Z^L_{(R,L),r,L}\right]\\ &=\left[Z^L_{(R,L),l,R}+Z^L_{(R,L),l,L}\right]-\left[Z^L_{(R,L),r,R}+Z^L_{(R,L),r,L}\right]+\left[Z^R_{(R,L),l,R}-Z^L_{(R,L),l,R}\right]\\ &=0+\Delta\left(Z_{(R,L),l,R}\right)=\sigma_{LL}\lambda(\alpha-\lambda)+\sigma_{RR}\left[(1-\alpha)(1-\lambda)-\lambda^2\right]+\lambda^2.\end{aligned}$$

3. $\Delta\left(Z_{(R,L),l,L}\right)$ and $\Delta\left(Z_{(R,L),r,L}\right)$ are negative but $\Delta\left(Z_{(R,L),r,R}\right)$ and $\Delta\left(Z_{(R,L),l,R}\right)$ are positive. By similar calculations,

$$\begin{aligned}\Delta &=\Delta\left(Z_{(R,L),l,R}\right)-\Delta\left(Z_{(R,L),r,R}\right)\\ &=\sigma_{LL}\lambda(2\alpha-1)-\sigma_{RR}(2\alpha-1)(1-\lambda)\\ &=(2\alpha-1)[\lambda\sigma_{LL}-(1-\lambda)\sigma_{RR}].\end{aligned}$$

Media maximizes

$$\max_{\substack{\sigma_{RR},\sigma_{LL}\in[0,1]\\ a=(\sigma_{RR}-\sigma_{LL})}}\frac{(2\lambda-1-a)\Delta}{(1-(2\lambda-1)a)}$$

We argue below that the maximum is obtained at $\sigma_{RR}=0$ and $\sigma_{LL}=1$. In the proof of Theorem 3 we have shown that $\frac{(2\lambda-1-a)}{(1-(2\lambda-1)a)}$ is decreasing in a. So this term is maximized at $a=-1$, that is at $\sigma_{RR}=0$ and $\sigma_{LL}=1$. We now show that Δ is also maximized $\sigma_{RR}=0$ and $\sigma_{LL}=1$.

First note that at this σ; $\Delta\left(Z_{(R,L),l,L}\right)=\lambda\alpha>0$ and $\Delta\left(Z_{(R,L),r,L}\right)=\lambda(1-\alpha)>0$. Therefore it belongs to zone 3 above and $\Delta=(2\alpha-1)\lambda$. This is the highest possible value of Δ in zone 3 as Δ is increasing in σ_{LL} and decreasing in σ_{RR}. In zone 2 the highest value of Δ is less than $\lambda^2+\lambda(\alpha-\lambda)=\lambda\alpha$ because $\sigma_{LL}\leq 1$ and the coefficient of σ_{RR} is negative. Moreover, since $\alpha\geq\frac{2}{3}$, $(2\alpha-1)\lambda\geq\lambda\alpha$. Hence the highest Δ is also obtained at $\sigma_{RR}=0$ and $\sigma_{LL}=1$. □

G Analysis of the Cheap Talk Game

Lets check media's incentive when the true state is R. If the media reports R, then its payoff is $\left[\lambda\frac{1}{2}+(1-\lambda)\frac{1-\gamma^*}{2}\right]$, where γ^* represents the individual who is indifferent between endorsing report R and report L. Explanation for this expression is as follows. With probability λ the bot reports R which leads to the first stage reports (R,R), in which case the media is endorsed with probability $\frac{1}{2}$ and with probability $(1-\lambda)$ the bot reports L, in which case the media is endorsed by all agents whose $\gamma\geq\gamma^*$. Instead, if the media reports L then its payoff is $\left[\lambda\frac{1+\gamma^*}{2}+(1-\lambda)\frac{1}{2}\right]$. Notice that media's payoff only depends on the endorsement and not on the actual state. Hence the payoffs from report L and R do not change when the true state is L. Comparing these payoffs, we get,

$$\gamma^*>0 \Leftrightarrow \Delta>0 \Rightarrow \text{media chooses } L,\ \text{that is } \sigma_{RR}=0, \sigma_{LL}=1$$
$$\gamma^*<0 \Leftrightarrow \Delta<0 \Rightarrow \text{media chooses } R,\ \text{that is } \sigma_{RR}=1, \sigma_{LL}=0$$
$$\gamma^*=0 \Leftrightarrow \Delta=0 \Rightarrow \text{media is indifferent },\ \text{that is } 0\leq\sigma_{RR}\leq 1, 0\leq\sigma_{LL}=1\leq 1.$$

The first two are babbling equilibrium of our cheap talk game. We check the third possibility now. In any non-trivial equilibrium of this kind, however, $\gamma^*=0$, that is endorsement is equally divided between R report and L report.

Media's ability to commit does not affect receiver's updated belief and best response to an anticipated strategy. Since media reports the truth in the second stage, voters' choices (both endorsement and voting) to media's anticipated strategy σ is the same as in Theorem 9. In the proof of Theorem 9, Appendix F, we have shown that Δ can take three possible values. Out of these, case (2) cannot occur in equilibrium, because here $\Delta=\Delta\left(Z_{(R,L),l,R}\right)$, which is positive. Either of case (1) or case (3) is possible.

For an equilibrium that satisfies case (1), σ_{RR}, σ_{LL}, must satisfy the following conditions:

(i) $b=\sigma_{RR}+\sigma_{LL}-1\geq 0$

(ii) $\Delta\left(Z_{(R,L),l,R}\right)=b\lambda(\alpha-\lambda)+[(1-\alpha)(1-\lambda)\sigma_{RR}+\alpha\lambda(1-\sigma_{RR})]\leq 0$

(iii) $\Delta\left(Z_{(R,L),r,R}\right)=b\lambda(1-\alpha-\lambda)+[\alpha(1-\lambda)\sigma_{RR}+(1-\alpha)\lambda(1-\sigma_{RR})]\leq 0$

Truth telling satisfies these conditions and hence is a cheap talk equilibrium for certain values of α and λ. Lets take $\frac{2}{3}\leq\alpha\leq 1$ and $\lambda=1$. Since $\sigma_{RR}=1$ and $\sigma_{LL}=1$, we have $b=1$, so condition (i) is satisfied. Condition (ii) is satisfied because $\Delta\left(Z_{(R,L),l,R}\right)=(\alpha-1)\leq 0$ and condition (iii) is satisfied because $\Delta\left(Z_{(R,L),r,R}\right)=-\alpha<0$.

Finally, for an equilibrium that satisfies case (3), σ_{RR}, σ_{LL}, must satisfy the following conditions:

(i) $b=\sigma_{RR}+\sigma_{LL}-1\geq 0$

(ii) $\Delta\left(Z_{(R,L),l,R}\right)=b\lambda(\alpha-\lambda)+[(1-\alpha)(1-\lambda)\sigma_{RR}+\alpha\lambda(1-\sigma_{RR})]\geq 0$

(iii) $\Delta\left(Z_{(R,L),r,R}\right)=b\lambda(1-\alpha-\lambda)+[\alpha(1-\lambda)\sigma_{RR}+(1-\alpha)\lambda(1-\sigma_{RR})]\geq 0$

(iv) $\Delta=(2\alpha-1)[\lambda\sigma_{LL}-(1-\lambda)\sigma_{RR}]=0$

In the proof of Theorem 9, Appendix F, we have shown that $\Delta\left(Z_{(R,L),r,R}\right)\geq 0\Rightarrow\Delta\left(Z_{(R,L),l,R}\right)\geq 0$, hence, condition (ii) stands redundant. Condition (iv) gives $\sigma_{LL}=\frac{1-\lambda}{\lambda}\sigma_{RR}$ and putting it in condition (i) , we obtain $\sigma_{RR}\geq\lambda$. Condition (iii) also reduces to $\left[\lambda^2-\sigma_{RR}(2\lambda-1)\right]\geq 0$, which holds true for all $\lambda\leq\sigma_{RR}\leq 1$. Therefore for any α and λ in our domain, we have cheap talk equilibrium where $\lambda\leq\sigma_{RR}\leq 1$ and $\sigma_{LL}=\frac{1-\lambda}{\lambda}\sigma_{RR}$.

References

Acemoglu, D., Ozdaglar, A., and Siderius, J. (2024). A model of online misinformation. *Review of Economic Studies*, 91:3117–3150.

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *The Journal of Economic Perspectives*, 31(2):211–235.

Anderson, S. P. and McLaren, J. (2012). Media mergers and media bias with rational consumers. *Journal of the European Economic Association*, 10:831–859.

Aziz, A. I. and Bischoff, I. (2025). Social media campaigning and voter behavior–evidence for the german federal election 2021. *European Journal of Political Economy*, 88(102685).

Azzimonti, M. and Fernandes, M. (2023). Social media networks, fake news, and polarization. *European Journal of Political Economy*, 76:102256.

Baron, D. (2006). Persistent media bias. *Journal of Public Economics*, 90:1–36.

Battaglini, M. (2002). Multiple referrals and multidimensional cheap talk. *Econometrica*, 70(4):1379–1401.

Battisti, M., Kauppinen, I., and Rude, B. (2024). Breaking the silence: The effects of online social movements on gender-based violence. *European Journal of Political Economy*, 85(102598).

Besley, T. and Prat, A. (2006). Handcuffs for the grabbing hand? Media capture and government accountability. *The American Economic Review*, 96(3):720–736.

Brocas, I., Carrillo, J. D., and Palfrey, T. R. (2012). Information gatekeepers: Theory and experimental evidence. *Economic Theory*, 51:649–676.

Chowdhury, P. R. (2024). Persuasion in social media: Smoke and mirror. *SSRN Working Paper*.

Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50(6):1431–1451.

DellaVigna, S. and Kaplan, E. (2007). The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, 122:1187–1234.

Djankov, S., McLiesh, C., Nenova, Tatiana, and Shleifer, A. (2003). Who owns the media? *Journal of Law and Economics*, 46(2):341–381.

Flückiger, M. and Ludwig, M. (2025). The structure of online social networks and social movements: Evidence from the black lives matter protests. *Journal of Public Economics*, 246(105373).

Foerster, M. (2023). A theory of media bias and disinformation. *SSRN Working Paper*.

Gentzkow, M. and Kamenica, E. (2017). Bayesian persuasion with multiple senders and rich signal spaces. *Games and Economic Behavior*, 104:411–429.

Gentzkow, M. and Shapiro, J. M. (2006). Media bias and reputation. *Journal of Political Economy*, 114(2):280–316.

Gentzkow, M. and Shapiro, J. M. (2010). What drives media slant? Evidence from u.s. daily newspapers. *Econometrica*, 78:35–71.

Gentzkow, M., Shapiro, J. M., and Stone, D. F. (2015). Chapter 14 - media bias in the marketplace: Theory. In Anderson, S. P., Waldfogel, J., and Strömberg, D., editors, *Handbook of Media Economics*, volume 1 of *Handbook of Media Economics*, pages 623–645. North-Holland.

Gerber, A. S., Karlan, D., and Bergan, D. (2009). Does the media matter? A field experiment measuring the effect of newspapers on voting behavior and political opinions. *American Economic Journal: Applied Economics*, 1(2):35–52.

Groseclose, T. and Milyo, J. (2005). A measure of media bias. *The Quarterly Journal of Economics*, 120(4):1191–1237.

Gul, F. and Pesendorfer, W. (2012). The war of information. *The Review of Economic Studies*, 79:707–734.

Guo, W.-C. and Lai, F.-C. (2014). Media bias, slant regulation, and the public-interest media. *Journal of Economics*, 114:291–308.

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *The American Economic Review*, 101(6).

Krishna, V. and Morgan, J. (2001). A model of expertise. *The Quarterly Journal of Economics*, 116(2):747–775.

Lewis, P. (2018). 'Fiction is outperforming reality': How youtube's algorithm distorts truth. *The Guardian*.

Lewis, P. and McCormick, E. (2018). How an ex-youtube insider investigated its secret algorithm. *The Guardian*.

Liberini, F., Redoano, M., Russo, A., Cuevas, A., and Cuevas, R. (2025). Politics in the facebook era. evidence from the 2016 us presidential elections. *European Journal of Political Economy*, 87(102641).

Milgrom, P. and Roberts, J. (1986). Relying on information of interested parties. *The RAND Journal of Economics*, 17(1):18–32.

Morris, S. (2001). Political correctness. *Journal of Political Economy*, 109(2):231–265.

Mullainathan, S. and Shleifer, A. (2005). The market for news. *The American Economic Review*, 95(4):1031–1053.

Rayo, L. and Segal, I. (2010). Optimal information disclosure. *Journal of Political Economy*, 118(5):949–987.

Xiang, Y. and Sarvary, M. (2007). News consumption and media bias. *Marketing Science*, 26(5):611–628.

Yea, S. (2018). Persuasion under the influence of fake news. *SSRN Working Paper*.

Zimdars, M. and McLeod, K. (2020). *Fake News: Understanding Media and Misinformation in the Digital Age*. The MIT Press.