Artificial Intelligence and Political Economy*

Daron Acemoglu† Asuman Ozdaglar‡ James Siderius§

February 23, 2025

Abstract

We consider the political consequences of the use of artificial intelligence (AI) by online platforms engaged in social media content dissemination, entertainment, or electronic commerce. We identify two distinct but complementary mechanisms, the social media channel and the digital ads channel, which together and separately contribute to the polarization of voters and consequently the polarization of parties. First, AI-driven recommendations aimed at maximizing user engagement on platforms create echo chambers (or "filter bubbles") that increase the likelihood that individuals are not confronted with counter-attitudinal content. Consequently, social media engagement makes voters more polarized, and then parties respond by becoming more polarized themselves. Second, we show that party competition can encourage platforms to rely more on targeted digital ads for monetization (as opposed to a subscription-based business model), and such ads in turn make the electorate more polarized, further contributing to the polarization of parties. These effects do not arise when one party is dominant, in which case the profit-maximizing business model of the platform is subscription-based. We discuss the impact regulations can have on the polarizing effects of AI-powered online platforms.

Keywords: political economy, artificial intelligence, social media, polarization, digital advertising

JEL Classification: L10, P40, M37

*This book chapter is inspired by the work in Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024) and Acemoglu, Ozdaglar, and Siderius (2024), with a special thanks to Daniel Huttenlocher for his contributions to the former. We would also like to thank numerous participants from the Economics of AI Conference 2024 and the Tuck School of Business for financial support.

†Massachusetts Institute of Technology, NBER, and CEPR, daron@mit.edu

‡Massachusetts Institute of Technology, asuman@mit.edu

§Tuck School of Business at Dartmouth College, james.siderius@tuck.dartmouth.edu

1 Introduction

Artificial Intelligence (AI) is reshaping almost every aspect of our social lives. Many of its far-ranging consequences are poorly understood and understudied. Its potential impacts on politics are no exception. There are myriad ways in which AI can change both autocracies and democracies. Although there is no conclusive evidence, many commentators suggest that these effects are already visible. Several works (Zeng (2020), Beraja et al. (2023) and Beraja, Yang, and Yuchtman (2023)) show how AI-driven surveillance has transformed authoritarian, and especially Chinese, politics. In the meantime, we are witnessing both high levels of misinformation and disinformation and much greater levels of polarization in democracies (Feezell, Wagner, and Conroy (2021) and Pew Research Center (2014)), as well as some suggestive evidence that online platforms are playing a major role in promoting polarization and the spread of divisive (and often misinforming) content (see, for example, Rathje et al. (2024) and Nayak et al. (2021)).

This paper is a first attempt to develop a formal model of some of AI's impacts on democratic politics. We study the effects of AI-powered business models of online platforms on the polarization of voters and parties. This choice is motivated by several considerations. An increasing fraction of the population in democracies receive their news and political information from social media. For example, 18% of Americans list social media as their main source of political and election information, while over 50% receive at least some news from social media sources (Pew Research Center (2024), Pew Research Center (2020), and Associated Press (2024)). The amount of time that people spend on online platforms for other reasons, including entertainment, shopping and other forms of communication, is even larger, and the average American is estimated to spend over two to eight hours a day on online digital platforms (Kemp (2020), Hruska and Maresova (2020), and Statista Research Department (2024a)). This makes online digital ads one of the most important channels via which politicians reach out to voters, and in the 2024 US election, 28% of political ads were online (see Statista Research Department (2024b)).

Social media and the online experience, today, are heavily shaped by complex algorithms that collect data and enable platforms to guide the information that users receive. In most cases, we also know that the incentives of online platforms are to maximize engagement from users, collect data about user types and preferences, and create custom audiences for potential advertisers as a way of monetizing these data via digital ads.

The information ecosystem created by AI-powered online platforms can impact politics via several channels. First, social media algorithms can leverage the data that they have about users in order to decide which information individuals see and who communicates with whom, with potentially major effects on people's political attitudes. One potential impact of social media is therefore on voter beliefs, for example, increasing or reducing political polarization. Voter polarization in turn affects party platforms. Second, online platforms amplify the ability and incentives of parties to reach out to specific groups of voters with custom-made messages. This channel also affects both voter and party polarization. Third, AI tools can create new ways in which non-party actors (such as civil society organizations or extremist groups) can reach users, and they also provide new means for users to check the veracity of certain claims and participate in new forms of interactions for democratic debate and

information exchange. Fourth, AI algorithms can also change internal party dynamics. Finally, AI can have broader social effects, via the education system or social networks, increasing or reducing people's interest in politics or their willingness and ability to engage in activities such as voting and political debate. In this paper, we focus on the first two channels, because they appear to have been more important so far (though the importance of the third channel may have grown over time, see Burton (2023) and Mostagir and Siderius (2024)).

Specifically, in this paper, we build on our previous work in Acemoglu, Ozdaglar, and Siderius (2024), which considered a model of information exchange on a social media platform that algorithmically determines the pattern of information sharing between users, and on Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024), which considers a model of digital ad-targeting by an online platform. These two models, therefore, bracket the two types of interactions that constitute our focus here. The first, which we refer to as the *social media* channel, is about how the AI-based news- and content-feeds affect voter polarization. The second, which we refer to as the *digital ads* channel, is about the effects of targeted digital ads. In both cases, we embed these online users in a model of political competition, comprising two parties that choose their platforms in order to attract voters and potentially buy digital ads targeting voters. In both cases, AI and digital business models play critical roles in platform incentives. In the former (social media channel which builds on Acemoglu, Ozdaglar, and Siderius (2024)), a social media platform chooses the content sharing network between users by leveraging AI tools and with the objective of maximizing engagement so as to boost its digital ad revenues. In the latter (the digital ads channel, which builds on Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024)), the social media platform uses AI tools to target digital ads to voters on behalf of political parties.

More specifically, the key decision in our model of social media is made by the platform about the structure of interactions between users. Users have a political belief (ideology), which they can update according to the information they receive. This information is contained in social media posts (articles), which come with a specific message (left-wing or right-wing) as well as a degree of reliability (observed perfectly by all players). Users decide whether to share, ignore or dislike a message, and their main source of utility is from these social media interactions. Users also vote in a two-party election according to the alignment of their beliefs with the policy positions on offer. Since each individual is infinitesimal, voting behavior does not generate an incentive to change users' social media behavior.

As in Acemoglu, Ozdaglar, and Siderius (2024), the payoff structure is such that users like to share articles, but not if others dislike and tag them as misinformation. This creates a tendency for individuals to share articles that have a message congruent with their beliefs, since they (rationally) judge those to be less likely to contain misinformation. The key decision in our model is made by the platform and concerns how the algorithm of the platform shapes the newsfeed and connections of users (observing the prior beliefs/ideology of users thanks to its AI-powered data crunching of their past behavior). In one extreme, the platform could choose a complete network, whereby posts by every user is equally likely to be seen by all other users, or an extreme filter bubble, whereby connections between users have complete "homophily", meaning that social media posts are only shared among people with the same prior beliefs/Ideology. The main result in Acemoglu, Ozdaglar, and Siderius (2024), which we leverage here, is that when articles have low reliability and the population

is already polarized, then filter bubbles are more profitable for the platform. A complete network could in principle generate engagement from a bigger base of users and generate more profits for the platform, but low-reliability articles presented to people with opposite ideology are likely to be identified and labeled as misinformation, reducing their circulation and thus the platform's revenue from engagement and digital ads. This channel implies that, when the relevant content does not have very high reliability to start with, social media algorithms increase voter polarization, and especially the polarization among the most extremist viewers. Given this type of polarization, parties choose more polarized policy positions, responding to their more extremist base. This is despite the fact that moderate voters retain identical or nearly-identical political beliefs after their social media experience. Our result consistent with evidence from Fiorina and Abrams (2008) suggesting that US polarization was for a long time characterized by a relatively stable, moderate electorate and party positions shifting towards the extremes.

Our digital ads model extends Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024) and considers two political parties reaching voters with digital ads, crafted and targeted by AI, on an online entertainment platform. Users can be naïve or sophisticated, with the key difference that naïve users are easier to convince with incorrect information, while sophisticated users have the correct model of the world. The presence of naïve voters increases the profitability of using digital ads for political purposes. When parties can use such ads, they have a greater ability to sway voters and this weakens their incentives to moderate their policies. Hence, albeit from a very different channel, this model also implies that AI and digital ad-based business models cause greater voter and party polarization.

At a high level, therefore, our analysis of both channels indicates that AI-powered platforms act as a source of polarization among voters, and political parties respond to this by choosing more polarized policy positions (even more so than the voters). The ways in which the two channels work are quite different, however. The social media channel works via the platform's ability to manufacture filter-bubbles and makes it more likely that left-leaning (right-leaning) users are exposed to more left-leaning (right-leaning) content. On the other hand, in the digital ads channel it is the parties that cater their messages for groups of voters that are more susceptible to believe the ads they receive and thus drive polarization. It is this ability to influence susceptible voters that makes digital ads more potent than other types of advertisement in informational outreach and more attractive to political parties. The overall result is again weaker incentives for parties to moderate their platforms.

In both cases, the results from the theoretical frameworks on which we are building also indicate that further competition among platforms does not ameliorate the polarization problem. We examine how certain regulations influence voter and party polarization in democracies. Specifically, we analyze the effects of a diversification standard, where platform users are exposed to more counter-attitudinal content (in the spirit of proposals such as Sunstein (2018)), and assess its potential to limit the creation of filter bubbles. In the context of digital ads, we study the implications of algorithmic regulations designed to enhance transparency and reduce the degree of ad targeting. Our analysis suggests that polarization, both among voters and in equilibrium party positions, responds to changes in digital awareness/literacy and the persuasiveness of ads.[1]

[1]Additionally, digital ad taxes that shift platforms toward a subscription-based model, discussed in detail in Acemoglu,

Related Literature. Our work is connected to three streams of literature, the first on social media, echo chambers, and filter bubbles; the second related to advertising in digital and non-digital settings; and the third about the politics of polarization. Although we build on the emerging empirical evidence in these areas, our contribution is theoretical and develops novel possible channels from the use of AI tools in social media and other platforms to voter and political party polarization.

In the context of the first literature, in a randomized-control experiment, Levy (2021) shows that Facebook's algorithm selectively chose pro-attitudinal political content to display to users based on their ideology, creating an endogenous echo chamber, or filter bubble. The negative effects of echo chambers are explored in an emerging empirical literature (Barberá (2020), Ross Arguedas et al. (2022), and Flaxman, Goel, and Rao (2016)) and has featured in a few theoretical studies as well (see Golub and Jackson (2010), Törnberg (2018), Baumann et al. (2020), and Mostagir and Siderius (2023)). Braghieri et al. (2024) and Molina (2024) study various channels by which social media interactions can lead to the emergence of echo chambers. Our contribution is distinguished from this literature because we explore mechanisms that have not yet been analyzed, in particular, the social media channel building on Acemoglu, Ozdaglar, and Siderius (2024). In this context, our analysis uncovers a novel pathway via which an AI-powered platform tends to amplify political biases and drives greater division in beliefs.

The second strand of literature studies the role of informational advertising (Butters (1977), Meurer and Stahl (1994)) and manipulative advertising (Piccolo, Tedeschi, and Ursino (2018), Gupta (2023)), where our earlier work in Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024) sits in-between. Our focus is on targeted (and manipulative) digital advertising in the political domain, where we find that party competition entices online business models to monetize via digital advertising, which also induces greater polarization in political beliefs. We refer to this as the digital ads channel, which is distinct from the social media channel, and is primarily driven by incentives of platforms to leverage AI for crafting individualized ads targeting users that are most susceptible to such ads. In turn, given this opportunity, parties pay to advertise to soften competition between each other, but in doing so, drive greater polarization in beliefs. When there is a single dominant political party, we find that the platform's business model gravitates toward subscription instead, and these polarizing forces disappear.

The last strand of literature contains both empirical articles (e.g., Fiorina and Abrams (2008), McCarty (2007), Barber et al. (2015), Callander and Carbajal (2022)) and a few theoretical works on polarization. For example, Dixit and Weibull (2007), Acemoglu, Chernozhukov, and Yildiz (2016), and Mostagir and Siderius (2022b) explore how Bayesian learning can lead to polarization under certain circumstances, while Glaeser (2005), Glaeser, Ponzetto, and Shapiro (2005), Acemoglu, Egorov, and Sonin (2013), and Druckman and Levy (2022) investigate various reasons for parties shifting away from moderate policies. Berman (2021) discusses the political science literature on the rise of radical right-wing populism, while Guriev and Papaioannou (2022) survey the broader political economy literature on populism. Tucker et al. (2018) summarize the literature on the effects of social media and polarization.

Paper Outline. The remainder of the paper proceeds as follows. Section 2 introduces our model of vot-

Huttenlocher, Ozdaglar, and Siderius (2024), can also limit political polarization, but we do not analyze them in this paper.

ing and party competition. Section 3 introduces the platform's incentives and their profit-maximization problem, both through the social media and digital ads channels. Section 4 studies and extends the model in Acemoglu, Ozdaglar, and Siderius (2024) and derives the implications for political polarization of voters and parties, whereas Section 5 examines the impact of targeted digital ads on political beliefs building on the model of Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024). We discuss the effects of various regulations in Section 6, and Section 7 concludes. The Appendix contains the proofs of the main results stated in the text.

2 Our Model of Voting

We start with our model of politics, which will be common in our explorations of the social media and digital ads channels. There are two political parties and a continuum of voters that are on online platforms (with total measure normalized 1), as well as some voters that are completely off-line.

Each online voter $i \in [0,1]$ participates both on an online platform and an election. Voter i holds prior beliefs b_i about a binary, fixed state of the world $\theta \in \{L,R\}$ (e.g., whether the left-wing or right-wing candidate is a better fit for office). The cumulative density function (cdf) of the distribution of beliefs is denoted by $H : [0,1] \to [0,1]$ and describes the political leaning of the electorate, which we assume is symmetric about the median belief of $1/2$ and non-degenerate (i.e., has density function h with $h(\alpha) = h(1-\alpha)$ for all $\alpha \in (0,1)$).

We now introduce an order that captures the extent of polarization. An electorate described by $\tilde{H}$ is more *polarized* than an electorate described by H (denoted by $\tilde{H} \succ H$) if $\tilde{H}^{-1}(\alpha) > H^{-1}(\alpha)$ for $\alpha \in (1/2,1)$ and $\tilde{H}^{-1}(\alpha) < H^{-1}(\alpha)$ for all $\alpha \in (0,1/2)$, with $\tilde{H}^{-1}(1/2) = H^{-1}(1/2)$. In other words, our order means that $\tilde{H}$ has the same median voter as H but is more spread out to the left and the right.

There are two political parties, $p \in \{L,R\}$, which strategically choose and commit to positions x_L and x_R on the interval $[0,1]$. Party payoffs are assumed to take the following form:

$$U_p = \mathbf{1}_p \cdot \mathcal{R} - \left(x - \hat{x}_p\right)^2,$$

where $\mathbf{1}_p$ is the indicator function denoting whether party p comes to power, $\mathcal{R}$ is a rent/utility from coming to power, x is the policy that is implemented (by either this party or its rival) and $\hat{x}_p$ is party p's ideal point, so that the term $-\left(x - \hat{x}_p\right)^2$ corresponds to the disutility of the political party due to deviation of the actual policy from its ideal point. We assume that $\hat{x}_L < 1/2 < \hat{x}_R = 1 - \hat{x}_L$, so that the left-wing party is indeed left-wing and the right-wing party is indeed right-wing. Our model of politics therefore combines ideological differences between parties together with a desire to come to power, separate from policy considerations.

Agents observe the positions of the two parties (x_L, x_R), and, as in Quinn, Martin, and Whitford (1999), voter i casts her vote for party j stochastically with probability

$$\rho_j(b_i, x_L, x_R) = \frac{\exp(-(b_i - x_j)^2)}{\exp(-(b_i - x_L)^2) + \exp(-(b_i - x_R)^2)}, \tag{1}$$

with $\rho_L(b_i) + \rho_R(b_i) = 1$ for all b_i. This "soft-max" is equivalent to agent i voting for her favored party

j, given by $j \in \arg\min_{j\in\{L,R\}} |x_j - b_j|$ (i.e., picking the candidate closest to her own belief in absolute value), but trembling with exponential decaying weight in the squared-distance between b_i and x_j. This form can be micro-founded by assuming the presence of non-policy valence or idiosyncratic preference terms that affect the choice of each voter between the two parties as in the probablilistic voting models (e.g., Coughlin (1992), and Lindbeck and Weibull (1993)). An important implication of this form is that voters with more extreme policy positions are more likely to vote for their preferred party than the more moderate voters.

The soft-max in equation (1) implies that candidate j captures a total share of votes from online voters given by

$$\chi_j = \int_0^1 \rho_j(b_i, x_L, x_R)\, h(b_j)\, dj.$$

We model the voting behavior of off-line voters, which are not influenced by any online activity in our model, in a reduced-form. In particular, we assume that the total votes from this group for party j is stochastic and is given by $\zeta_j \sim$ Uniform$[0,1]$, where the uniform assumption is adopted for simplicity. This implies that the total votes of party j will be $\chi_j + \zeta_j$ and those of party $-j$ will be $\chi_{-j} + \zeta_{-j} = 1 - \chi_j + 1 - \zeta_j$. Party j then wins the election and comes to power if

$$\chi_j + \zeta_j > (1 - \chi_j) + (1 - \zeta_j).$$

We consider the game where parties $(L,\ R)$ choose positions (x_L^*, x_R^*) simultaneously, then online agents vote for candidates according to ρ and then the random variable ζ is realized.

Given the symmetry of H around the median, in what follows we focus on the *anti-symmetric* Nash equilibrium of this game, where $|x_L^* - 1/2| = |x_R^* - 1/2|$, or in words, where the two parties locate and equi-distant fashion relative to the median. Because uniqueness is not guaranteed, we will also focus on the extremal anti-symmetric Nash equilibria, whereby among all anti-symmetric Nash equilibria, we pick the one that has the largest $|x_L^* - 1/2|$. This focus is motivated by our desire to carry out simple comparative statics.

Baseline. Under the baseline Hotelling model (Hotelling (1929); Downs (1957)), we obtain the standard Downsian convergence result, where both political parties choose to cater to the median voter's preferences by setting $x_L^* = x_R^* = 1/2$ in the unique equilibrium. One of our key results will be that because of the effects of AI-powered online business strategies, we will generally have policy differentiation—$x_L^* < x_R^*$. This arises because voters near the poles are less likely to vote for a candidate that has a platform misaligned with their ideology, driving the parties' incentives to differentiate themselves (as opposed to the standard framework, where such forces are muted). Our next result characterizes this differentiation as a function of the belief distribution H.

Lemma 1.

(i) There exists at least one anti-symmetric Nash equilibrium (x_L^*, x_R^*) with $x_R^* = 1 - x_L^*$ and every equilibrium satisfies $x_L^* < \frac{1}{2} < x_R^*$;

(ii) There exists a unique extremal anti-symmetric Nash equilibrium $(\underline{x}_L^*, \bar{x}_R^*)$ *with* $\underline{x}_L^* \leq x_L^*$ *and* $\bar{x}_R^* \geq x_R^*$ *for any anti-symmetric equilibrium pair* (x_L^*, x_R^*)*;*

(iii) If $\tilde{H} \succ H$*, the extremal anti-symmetric Nash equilibrium under* H*,* $(\underline{x}_L^*, \bar{x}_R^*)$*, and under* $\tilde{H}$*,* $(\tilde{\underline{x}}_L^*, \tilde{\bar{x}}_R^*)$*, satisfy* $\tilde{\underline{x}}_L^* < \underline{x}_L^*$ *and* $\tilde{\bar{x}}_R^* > \bar{x}_R^*$*.*[2]

Lemma 1(i) establishes the existence of an anti-symmetric Nash equilibrium in our voting game where the two parties distinguish themselves from each other. In all such equilibria, we have $x_L^* < x_R^*$ and these policy positions lie on opposite sides of the median voter's most preferred policy. This is a consequence of the combination of uncertain election outcomes that depend smoothly on policy positions (due to probabilistic voting and the presence of stochastic off-line voters) and the differences in the ideological positions (ideal points) of the two parties. Intuitively, policy convergence whereby $x_L^* = x_R^* = x$ for some x could never be an equilibrium. Such a point would deviate from the ideal point of at least one of the two parties, say party R, and if this party moved towards its ideal point, this would have a second-order cost in terms of vote shares as derived in (1), but would always have a first-order gain in terms of moving closer to one's own ideal point.

While the existence of an anti-symmetric Nash equilibrium is immediate (from a straightforward application of Brouwer's fixed point theorem given the continuity of vote shares in (1)), uniqueness is not guaranteed, because the trade-off—between being close to one's ideal point and gaining more votes from (1)—is often non-convex. However, there can only be a finite number of symmetric Nash equilibria, and hence the existence of an extremal one is also guaranteed, as stated in part (ii).

The comparative static result in part (iii) of the lemma is important for the rest of our analysis. It states that a more polarized online electorate leads to more polarized policy positions among the two parties—in the sense that both parties move further away from the median voter's ideal point. There are two complementary intuitions for this result. To explain these, first note that we do not know in general whether x_L^* is to the left or to the right of $\hat{x}_L$ (and likewise for x_R^* relative to $\hat{x}_R$). Suppose first that it is to its right, meaning that party L (and symmetrically its rival) are moderating their positions in order to increase their probability of coming to power. In this scenario, a polarized electorate enables both parties to move closer to their ideal points, leading to a further polarization of party positions. Alternately, both parties could already be more polarized than their ideal points in order to strongly cater to their more extreme voters. In this case, a further polarization of the electorate strengthens this effect and again intensifies the polarization of party positions. Both of these channels are rooted in the fact that, as (1) clarifies, extremist voters are more valuable to a party because the soft-max in this expression has a steeper curvature for these voters. Also notably, both of these channels work even if only the most extremist voters in the party's base become more polarized, and this is in fact a common situation we will see in both the social media and digital ads channels.

[2] To avoid cumbersome terminology we often refer to "the equilibrium" as the extremal anti-symmetric Nash equilibrium and denote it by (x_L^*, x_R^*), avoiding the more notation-intensive $(\underline{x}_L^*, \bar{x}_R^*)$, although they refer to identical entities.

3 Platform Incentives

We next describe the online platform's profit and choice of business model, which will also apply to both of our mechanisms explored in the next two sections. For ease of exposition, we take the perspective of a social media platform, though much of our analysis applies more generally to online platforms that monetize through subscription and/or digital advertising. We assume that the platform is revenue-maximizing (meaning that we ignore costs) and its revenues come from digital ads and potentially subscription income. Namely, the objective of the platform is to maximize

$$\Pi = I^S + \int_0^1 \delta z_i \, di, \tag{2}$$

where I^S is subscription income (if any) and the second term is total digital ad revenue. Specifically, z_i denotes the number of ads (in expectation) that can be presented to user i and δ is revenue per ad.

Under social media model in the next section, we assume that there is no subscription income, δ is exogenously given, and z_i is proportional to the number of shares of posts that individual i engages in. Therefore, $\int_0^1 z_i \, di$ will be proportional to the total engagement of content on social media, incentivizing the platform to maximize engagement or shares. Under the digital ads model, the platform will choose between a subscription-based model where $I^S > 0$ and $\int_0^1 z_i \, di = 0$ or $I^S = 0$ and an ads-based business model with $\int_0^1 z_i \, di > 0$, or a mixture with both $I^S > 0$ and $\int_0^1 z_i \, di > 0$. In this case, δ will be optimally chosen by the platform as a function of the willingness of parties to pay for digital ads, and z_i will be related to the expected number of ads individual i will consume as we describe there.

4 Social Media and Content Sharing

We begin by describing our model of social media and content sharing, which is based on a simplified version of Acemoglu, Ozdaglar, and Siderius (2024). We then present Theorem 3, our main result from Acemoglu, Ozdaglar, and Siderius (2024), which characterizes the platform's profit-maximizing sharing network and shows that, when the available social media content has low reliability, this sharing network takes the form of a filter bubble, where individuals see only articles shared by like-minded users. We use this characterization to pin down how the electorate's beliefs shift through the social media channel, and derive the corresponding impact on political party positions.

4.1 Our Model of Content Sharing

Users hold prior beliefs $b_i \in [0,1]$ about $\theta = R$ and communicates stochastically with other users on social media, according to a sharing network $\mathbf{P}$ determined algorithmically by the platform. We focus on "articles" users share over this network with each other.

Each article on social media has a three-dimensional type (r, ν, m) indicating its reliability ($r \in [0,1]$), veracity ($\nu \in \{\mathcal{T}, \mathcal{M}\}$) and message ($m \in \{L, R\}$). The reliability specifies the likelihood the article is truthful instead of containing misinformation. If the article is truthful, it is more likely to argue for the true θ, whereas an article with misinformation simply provides noise about θ. Formally,

we have that $\mathbb{P}[m = \theta \,|\, \nu = \mathcal{T}] = p > \frac{1}{2}$ and $\mathbb{P}[m = \theta \,|\, \nu = \mathcal{M}] = \frac{1}{2}$. Importantly, both the reliability and the message are publicly observable, but the veracity is not—instead this must be inferred from the observable characteristics of the article, using Bayes' rule, which will also depend on agent i's prior b_i about θ.

Each agent can decide between sharing, ignoring, or disliking a given article. The payoff to ignoring is 0, whereas the payoff to disliking is equal to $\tilde{u}\mathbf{1}_{\nu=\mathcal{M}} - \tilde{c}$, which implies that the agent gets some utility $\tilde{u}$ from calling out misinformation, but must exert some effort $\tilde{c}$ in order to do so (and incurs the cost regardless of whether the article actually contains misinformation).

There are two sources of utility and costs from sharing content. The first component is network-independent; agents receive utility that only depends on the nature of the content. Agents get positive utility, $u\mathbf{1}_{\nu=\mathcal{T}}$, from sharing truthful content and pay a cost, $c\mathbf{1}_{\nu=\mathcal{M}}$, for sharing misinformation. Second, when they share, their content gets passed onto their neighbors in the sharing network, as determined by $\mathbf{P}$. For every re-share received, the agent gets an additional marginal utility of κ, whereas for every dislike received, the agent pays a reputational cost d. The parameter κ can capture the degree of "sensationalism" of the article, representing the idea that some content is more sensational and thus generates more social benefit for the sender when it spreads virally.

4.2 Platform Incentives

The platform's objective is to maximize digital advertisement revenue, as specified in equation (2). Since δ is fixed in this case, this is equivalent to maximizing $\int_0^1 z_i di$, which means maximizing the circulation of social media content by choosing the sharing network $\mathbf{P}$. We assume that this choice is without any constraints, meaning that the chosen network could be any element of the set of all possible stochastic sharing networks over nodes defined by the users on this platform. This is a simplification adopted for transparency and tractability, and the general conclusions are not affected if we assume that the sharing network is chosen to be a subset of the pre-existing network structure, such as the exogenous online social network.

4.3 Filter Bubbles

Following the the terminology in Acemoglu, Ozdaglar, and Siderius (2024), a sharing network has *maximal connectivity* if it is the complete network. It has *maximal homophily* if it can be represented by two islands with within-island link probability $p_s = 1$ and between-island link probability $p_d = 0$. Here, we first replicate Theorem 3 from Acemoglu, Ozdaglar, and Siderius (2024):

Proposition 4.1. *The platform's profit-maximizing sharing network is determined by a reliability threshold $r_P \in (0,1)$ such that:*

(i) If $r < r_P$, the platform's profit-maximizing sharing network has maximal homophily and thus corresponds to a filter bubble.

(ii) If $r > r_P$, the platform's profit-maximizing sharing network has maximal connectivity.

In a nutshell, when the relevant articles have low-reliability in the sense of $r < r_P$ (where $r_P \in (0,1)$), the platform's profit-maximizing sharing network is a filter bubble, whereby the article is recommended only to some set of agents, $\mathcal{A}'$, which made up of agents with the same ideological position and is consequently fully disconnected from the set of other agents, $\mathcal{A}''$. When $r > r_P$, the platform chooses the complete network. The intuition, as already hinted in Section 1, is that with low-reliability articles, the platform is worried about agents with different priors/political views calling out misinformation and thus slowing down or preventing the viral spread of articles, which is costly for the platform because it reduces engagement. One other noteworthy feature of this theorem is that it makes it natural for us to focus on extreme filter bubbles and complete networks, as the two types of sharing network structures that the profit-maximizing platform will choose.

In what follows, we focus on the case where $r < r_P$ and also assume that the benchmark for comparison without AI is the complete network, since without AI and extensive data collection about users, the platform could not ensure that an article is only seen by those who have similar views to those it advocates. Moreover, to deliver our insights in the most transparent manner, we consider two articles (r_1, ν_1, m_1) and (r_2, ν_2, m_2), and suppose that $r_1 = r_2 = r$ and $m_1 = L$ but $m_2 = R$.[3]

We next show that the reliability parameter r both determines the size and nature of the filter bubble and acts as a proxy for the quantity of misinformation in circulation on social media (with higher reliability indicating less pervasive misinformation).

Proposition 4.2. *Suppose $r < r_P$. As r decreases further, the size (and the ideological homogeneity) of the filter bubble also decreases. As $r \to 0$, only the online voters with the most extreme views aligned with the article are included in the filter bubble.*

The proof and the main idea of this proposition rely on Proposition 4.1. As we have seen, when content is not too reliable ($r < r_P$), the profit-maximizing sharing network $\mathbf{P}^*$ induces an extreme filter bubble, sharing a given article only within an ideologically-aligned community. Note, however, that even within this regime, the value of r determines the size of the filter-bubble. When $r \approx r_P$, then the filter bubble is in fact quite similar to the complete network and includes a large fraction of all users. But as r declines further, the filter bubble gets smaller and more extreme, and in the limit, it includes a tiny and very extremist fraction of the population. What this implies, in particular, is that for sufficiently low r, moderate voters are not part of the filter bubble. This result, which is new relative to Acemoglu, Ozdaglar, and Siderius (2024), and is proved in the Appendix.

4.4 Polarization of the (Online) Electorate

Once agents receive articles (regardless of the sharing network $\mathbf{P}$), they update their priors b_i according to Bayes' rule to form posterior beliefs $\hat{b}_i$ about θ depending on the message of the article they observe.

[3]The case where $r_1 \neq r_2$ will lead to fundamentally similar insights, but the distribution of beliefs after the articles are recommended, $\tilde{H}$, will depend on the difference between r_1 and r_2. Since we are interested in the role AI plays in its recommendation algorithm, we simplify along this dimension and focus on articles that have similar likelihood to contain misinformation. Likewise, the case of $m_1 = m_2 = R$ or $m_1 = m_2 = L$ is less interesting because one political party is "handicapped" by a lack of information it can push on the platform. If parties produced articles themselves, this would naturally ensure that $m_1 = L$ and $m_2 = R$.

Importantly, we assume agents do not update their beliefs about θ for articles that may be hidden from them.[4] Political parties then choose their positions $(\tilde{x}_L^*, \tilde{x}_R^*)$, given the distribution of $\hat{b}_i$'s in the population, denoted by $\tilde{H}$. We say that $\tilde{H}$ is more polarized than an alternative H' according to the polarization definition above, and denote the distribution of beliefs under a filter bubble by $\tilde{H}_{FB}$ and under the complete network by $\tilde{H}_C$.

We also define $(\tilde{x}_L^{FB}, \tilde{x}_R^{FB})$ as the party positions under a filter bubble, and $(\tilde{x}_L^C, \tilde{x}_R^C)$ as party positions under a complete network. We say that parties are more polarized under filter bubbles if $\tilde{x}_L^{FB} < \tilde{x}_L^C$ and $\tilde{x}_R^{FB} > \tilde{x}_R^C$. Existence and uniqueness of the extremal anti-symmetric equilibrium is guaranteed by Lemma 1 in Section 2.

Our next result establishes the impact of filter bubbles on voter polarization. We refer to the individuals who are at the median among the left-leaning and the right-leaning voters, respectively (those that the midpoints between 0 and 1/2, and between 1/2 and 1, or mathematically, $\tilde{H}^{-1}(1/4)$ and $\tilde{H}^{-1}(3/4)$) as the *left median* and *the right median*.

Proposition 4.3. *There exist* $0 < r_1 < r_2 < r_P < 1$ *such that an article with reliability* $r \in (r_1, r_2)$ *leads to voter polarization in the sense that* $\tilde{H}_{FB} \succ \tilde{H}_C$*, but the left and the right medians remain identical under both belief distributions, i.e.,* $\tilde{H}_{FB}^{-1}(1/4) = \tilde{H}_C^{-1}(1/4)$ *and* $\tilde{H}_{FB}^{-1}(3/4) = \tilde{H}_C^{-1}(3/4)$.

This is a critical result for our analysis. It establishes not only the presence of filter bubbles, but it also indicates that with sufficiently low reliability ($r < r_2$), social media content has no impact on moderate voters, including those between the left (respectively the right) median and the overall median at $1/2$. The intuition is given by Proposition 4.2. When $r < r_2$, the profit-maximizing filter bubble on the left does not include the 25th quantile of the belief distribution, and thus only the beliefs of voters that are more extreme than this could be impacted by social media content. Put simply, low-reliability content on social media affects only extremist voters, and not moderate ones. This finding is congruent with claims made in Fiorina and Abrams (2008), who have argued that the moderate electorate did not become more polarized, and US political polarization was driven by parties. Our theory suggests that this pattern may be the result of the polarization by extreme voters, to which party positions responded. This story is also in-line with the work of Guess, Nyhan, and Reifler (2020) and Lazer et al. (2018), who argue that most of the misinformation in the American media diet came from a small percentage of social media users who were the most extreme on both sides of the political spectrum. To this result, we also add the possibility that AI-powered filter bubbles are making this polarization of the extremes even more intensive. The intuition in our case is exactly the same: while $\tilde{H}_{FB}$ is more polarized than $\tilde{H}_C$, the additional polarization is accounted for the most extremist social media participants.

It is also interesting to see why similar dynamics do not transpire under complete networks. In this case, all agents will see both articles, which have the same reliability but argue for opposite perspectives. In consequence, their overall effects will cancel each other and the posterior belief distribution $\tilde{H}$ will continue to coincide with the prior distribution H, exhibiting no voter polarization. In more general settings, left-leaning and right-leaning articles may have different credibilities, and it is natural

[4] Appendix C in Acemoglu, Ozdaglar, and Siderius (2024) provides a microfoundations for this feature.

to assume that the right-leaning articles will have greater credibility when $\theta = R$ and vice versa. If so, then with a complete network, people will update their priors and posterior distribution will shift in the direction of the truth.

4.5 Polarization of Parties

Using Lemma 1, we can establish the following consequence of Proposition 4.3.

Corollary 4.1. *An article with reliability $r \in (r_1, r_2)$ induces parties to become more extreme in equilibrium, i.e., $\tilde{x}_L^{FB} < x_L^*$ and $\tilde{x}_R^{FB} > x_R^*$.*

When the reliability of social media content is not too low, then the most extreme right-wing agents will update their priors b_i to more extreme posteriors $\hat{b}_i > b_i$ after seeing message $m_2 = R$. Likewise, the most extreme left-wing agents will update their priors b_i to more extreme beliefs $\hat{b}_i < b_i$ after seeing message $m_1 = L$. This is because, in such a strong filter bubble, agents from these groups will not see the opposing messages. Then, from Lemma 1(ii), this type of voter polarization will induce party polarization, with each party choosing more extreme policy positions.

To further highlight the role of content reliability, in the next proposition we show that the results of Proposition 4.3 do not hold when content reliability is very high or very low.

Proposition 4.4. *Fix $\varepsilon > 0$. There exist $0 < \underline{r}^* < \bar{r}^* < 1$ such that if $r < \underline{r}^*$ or $r > \bar{r}^*$, then*

$$|\tilde{x}_L^{FB} - x_L^*| < \varepsilon \text{ and } |\tilde{x}_R^{FB} - x_R^*| < \varepsilon.$$

This proposition therefore highlights that AI-powered social media is most likely to generate voter and party polarization when the reliability of news content is middling (or when there is sufficient uncertainty about the veracity of online content). Nevertheless, our reading of the evidence is that the middling reliability regime is the best approximation to the current political climate, which has sometimes been described as a "post-truth" era, where there is little agreement in society what is an objective truth (see Lewandowsky, Ecker, and Cook (2017) and Rochlin (2017), as well as Suarez-Lledo and Alvarez-Galvez (2021), Nyhan (2020), and Pennycook and Rand (2019)).

Several interesting comparative statics follow straightforwardly from this result. First, if we hold constant reliability r and increase the sensationalism of the article, κ, or decrease the reputational concerns, d, from sharing the article, then the platform's sharing network is less likely to induce a filter bubble. This will lead to less polarized political positions in equilibrium, which may at first appear paradoxical. However, in reality, highly sensational articles are also often unreliable (see Grinberg et al. (2019)). Thus, if we consider a simultaneous increase in κ and a decrease in r is smaller, then filter bubbles can become more likely, and we can see more extreme political party positions, as stated in Corollary 4.1.

We can also show, with similar arguments to those in Acemoglu, Ozdaglar, and Siderius (2024), that more divisive content (with higher p and lower q) is more likely to induce polarization among voters and thus between parties. This is for two related reasons. First, with higher p and lower q, the article leads to stronger belief updates, which then generates more polarized posterior beliefs. Second,

when content is more divisive, it becomes more profitable for the platform to opt for a filter bubble, as shown in Proposition 3 of Acemoglu, Ozdaglar, and Siderius (2024), and via this choice of platform architecture, more divisive content again promotes polarization.

Finally, we can also study the implications of the prior distribution of the electorate becoming more polarized. According to Proposition 3 in Acemoglu, Ozdaglar, and Siderius (2024), when prior polarization becomes stronger, filter bubbles become more likely, which further drives belief polarization by Proposition 4.3, and in turn, party polarization by Corollary 4.1. In this sense, polarization is like "Pandora's box": once the electorate becomes polarized, social media fuels the flames of polarization through the AI's choice of algorithm, which further reinforces the polarization. This perpetuating cycle working via social media can lead to a positive feedback loop that exacerbates any initial polarization.

5 Digital Political Advertising

Next, we extend a simplified version of the model of Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024) of online business models and digital ads to study our second channel for political polarization

5.1 Sophisticated and Naïve Users

We let λ fraction of the users be *sophisticated* and $1-\lambda$ fraction of the users be *naïve*. There is a state of the world $\theta \in \{L, R\}$ as before, and we suppose that users hold prior beliefs $b_i \sim H$. Sophisticated users have a correctly-specified subjective model of digital ads, which is that the signals s_i contain no information about the underlying state θ. On the other hand, naïve agents believe that political ads contain information about θ, and that false positives are relatively rare. In particular, we assume that naïfs hold the subjective model that $\mathbb{P}[s_i \neq \theta] = \phi \in [0, 1/2]$. Both ϕ and λ encode the degree of targeting and manipulation possible from digital ads on the platform, as discussed in detail in Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024). Lower values of ϕ indicate a larger misspecification on the part of naïve agents who believe that false positives are relatively rarer than they actually are. This allows political parties to send a positive signal, independent of how fit they are for office, to persuade naïfs to vote for them. Similarly, lower values of λ represent a larger fraction of naïfs in the population, who are more susceptible to digital ad targeting.[5] This structure implies that as ϕ and λ increase, we approach the standard model with fully rational agents (which is in fact the limit of our setting with $\lambda = 1$ or $\phi = 1/2$).

5.2 Digital Ads and Platform Choices

When there are multiple parties, we assume there are two different ads the platform can show (one for the left-wing candidate and one for the right-wing candidate) with ad loads $\alpha^{(L)}$ and $\alpha^{(R)}$ and total ad load $\alpha \equiv \alpha^{(L)} + \alpha^{(R)}$ (with the probability of the left-wing candidate's ad being shown with probability

[5] In this context, "targeting" refers to the ability of AI-crafted messages to be more persuasive for naïve agents, rather than targeting and reaching naïve agents or those with a particular ideology more frequently. Such additional dimensions of targeting can also be incorporated, but we do not pursue them in our current work, since even without this feature, we show that digital ads polarize the electorate and induce further polarization between political parties.

$\alpha^{(L)}/(\alpha^{(L)}+\alpha^{(R)})$). Existence and generic uniqueness of an equilibrium under this generalization with multiple parties follows immediately from Proposition 8 in Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024). The parameter α here represents the overall frequency at which a political ad is displayed relative to organic content, such as entertaining videos. U'ssers view the ad a stochastic number of times $k \sim \text{Poisson}(\alpha T)$. Each ad provides the same signal $s_i \in \{L, R\}$, depending from which party it comes. We assume that the left-wing party's ad always argues for the left-wing candidate being better and the right-wing party's ad always argues for the right-wing candidate. This implies that ad content is not truly informative about the underlying state θ. Sophisticated voters recognize this and do not update their beliefs b_i about θ in response to viewing ads.[6] Naïve users, on the other hand, do not fully recognize this and, consequently, update their beliefs using Bayes' rule under the misspecified subjective model that $\mathbb{P}[s_i \neq \theta] = \phi \leq \frac{1}{2}$.

A user receives no inherent value from the ads, whereas she collects an expected entertainment utility of $(1-\alpha)T$ from viewing organic content.[7] This implies that more frequent ads have a cost in terms of lower entertainment utility of users. The platform makes a take-it-or-leave-it offer to the political parties in return for the digital ads that will be shown to users. These offers determine δ and the z_i's above.

5.3 Two-Party Competition

Our main result in this section characterizes the profit-maximizing platform business model as a function of the digital nature of advertising and the ad technology when there are two political parties.

Proposition 5.1. *There exists $\phi^*(\lambda) \in [0,1]$ such that*

(i) If $\phi > \phi^*(\lambda)$, the platform offers a subscription-based plan with $P^* = T - v$ and no advertising $\alpha^* = 0$;

(ii) If $\phi < \phi^*(\lambda)$, the platform offers an ad-based plan with $P^* \in [0, T-v)$ and symmetric advertising loads $\alpha^{(L)*} = \alpha^{(R)*} = \alpha^*/2 > 0$.

Moreover, $\phi^*(\lambda)$ is decreasing in λ.

The intuition for Proposition 5.1 can be seen by varying the fraction of naïve agents (lower λ) and/or the amount of false positive information in digital ads (lower ϕ). Under Proposition 5.1(i), the fraction of naïve agents is small and digital ads do not send many false positives or equivalently have little persuasiveness for naïve agents. In this regime, the platform is better off monetizing via subscription fees because ads do little to sway the beliefs in the population, making parties unwilling to pay much to

[6] This assumption is similar to the result of Lemma 1 in Acemoglu et al. (2024), which shows that no surplus is generated from sophisticated voters who have linear demand over the advertised product. In our model with political digital ads, we impose a similar finding, which is that rational voters do not update their beliefs (because ads contain no information θ) and thus, similarly, no surplus is generated from advertising to only rational agents.

[7] The fact that users are atomless in voting and do not directly benefit from being better informed about the candidates is one key difference between our model of political digital ads and Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024), where agents recognize that the information in the ads will allow them to make better purchasing decisions. Another key difference is that candidates are vertically differentiated in this model, whereas products are horizontally differentiated in Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024).

advertise on the platform. Consequently, the cost of showing ads is outweighed by the loss of revenue due to lower entertainment utility for users. In contrast, in Proposition 5.1(ii), the population is more susceptible to digital ads and there are more naïfs in the population. In this case, advertising occurs in equilibrium, as both parties are willing to pay for digital ads to try differentiate themselves from their rival and this compensates the platform for the loss in subscription revenue resulting from lower entertainment utility for users. The next proposition characterizes the implications of digital ads for the belief distribution of voters.

Proposition 5.2. *Let density h be convex and continuous. Then:*

(i) If $\phi > \phi^*(\lambda)$, then the electorate has the same belief distribution H and political parties choose the same positions (x_L^*, x_R^*) in the extremal anti-symmetric Nash equilibrium.

(ii) If $\phi < \phi^*(\lambda)$, then the electorate has a belief distribution $\tilde{H} \succ H$ and political parties choose $(\tilde{x}_L^*, \tilde{x}_R^*)$ in the extremal anti-symmetric Nash equilibrium with $\tilde{x}_L^* > x_L^*$ and $\tilde{x}_R^* > x_R^*$.

Proposition 5.2 shows that political advertising in the regime where ads are targeted at naïfs tends to generate greater political polarization among the electorate, which in turn affects the polarization of political parties. Polarization of the electorate in regime (ii) is driven by the fact that agents receive more signals about the political parties, which naturally induces greater movements in their beliefs via Bayesian updating. While some agents become more moderate as a result of political advertising, the average movement of beliefs is toward the extremes. This can be seen from the fact that if all agents were sophisticated, advertising would induce a mean-preserving spread of beliefs (which follows from the Martingale property of Bayesian beliefs). However, with naïve agents, belief movements are more pronounced and have a positive drift toward more extreme positions given the misspecification about the informativeness of political ads. Then from Lemma 1, such transitions to more extreme ideologies by the electorate also leads the parties to anchor to more polarizing positions as a result.

In Proposition 5.2, as opposed to in Proposition 4.3, we do not know where among the existing belief distribution polarization kicks in. The next corollary shows that this takes a form similar to Proposition 4.3, with the extremes becoming more polarized, provided that the naïve voters were already more extreme than sophisticates to start with. To state this result, we introduce the additional notation of H_N and H_S that designate the belief distributions of naïve voters and sophisticated voters, respectively, with $H = \lambda H_S + (1-\lambda)H_N$.

Corollary 5.1. *Let H_N have support $[0, b_N^*] \cup [1 - b_N^*, 1]$ and H_S have support $[b_N^*, 1 - b_N^*]$. If $\lambda > 1/2$ and $\phi < \phi^*(\lambda)$, then voters at the medians on the left and the right ($\tilde{H}^{-1}(1/4)$ and $\tilde{H}^{-1}(3/4)$) continue to have the same b_i and $\tilde{b}_i$ after the introduction of digital ads, but the overall distribution becomes more polarized, i.e., $\tilde{H} \succ H$, which results in polarization in political party positions, $\tilde{x}_L^* < x_L^*$ and $\tilde{x}_R^* > x_R^*$.*

Corollary 5.1 establishes an analogous result to Proposition 4.3, whereby relatively extreme voters become even more extreme in response to digital ads, while there is no similar polarization among moderate voters. The movements in the two extremes still lead to greater polarization on the part of the parties, however, with the same logic as in Lemma 1(i). While we do not formally study it in

this paper, Corollary 5.1 opens up questions about the potential welfare implications of digital ads in driving parties to cater more to their most extreme supporters, at the cost of neglecting more moderate agents whose beliefs are less impacted via the digital ads channel.

5.4 Online Business Models with a Dominant Party

In this subsection, we consider a dominant party (e.g., an incumbent) that can advertise on the platform (which, without loss of generality, we suppose is the right-wing party). To model political dominance, we relax symmetry of the underlying belief distribution H, as well as symmetry of the ideal points $\hat{x}_L$ and $\hat{x}_R$ so that the dominant party wins probability $\chi_R^* > \frac{1}{2}$ (without advertising).[8] We also assume, for simplicity, that under belief distribution H, the extremal equilibrium position x_R^* satisfies $x_R^* \geq \hat{x}_R$, so the equilibrium position the dominant party adopts is more extreme than its ideal point.

Our next result shows that, in this case, the platform's profit-maximizing business model is fully subscription-based, rather than monetized via digital ads, independent of λ and ϕ.

Proposition 5.3. *There exists $\bar{\chi} > 0$ such that if $\chi_R^* > \bar{\chi}$, the platform offers a subscription-based plan with $P^* = T - v$ and no advertising $\alpha^* = 0$. Moreover, the electorate has the same belief distribution H and political parties choose the same positions (x_L^*, x_R^*) in the extremal Nash equilibrium.*

The intuition for Proposition 5.3 comes from considering the platform's incentives in setting its business model. If the dominant party decided to advertise, it would increase its vote share, as digital ads make the electorate more favorable toward it. However, the platform must weigh the revenue it earns from digital ads against the potential loss of subscription fees. When there is a dominant party, the incremental gain in vote share the party receives is small in comparison to the subscription revenue the platform forfeits by showing ads (given the party already has captured a substantial fraction χ_R^* of the vote). Importantly, this conclusion holds regardless of whether we are in part (i) or (ii) of Proposition 5.1. Put differently, the equilibrium business model may be subscription-based even when there are many naïve voters susceptible to digital ads (low λ and low ϕ), and the ad compensation from the dominant party may be insufficient to outweigh lost subscription revenue for the platform.

In contrast, as emphasized in Proposition 5.1(ii), with sufficiently naïve voters, competition between the two parties induces an equilibrium with digital ads. In this case, each party has an incentive to use digital ads to improve its position relative to its rival. The juxtaposition of these two results implies that, starting from a position in which there is a dominant party, as political competition intensifies, at some point we may switch away from subscription-based business model towards intensive use of digital political ads.

6 Effects of Regulations

In this section, we examine how certain online regulations might influence the polarizing effects of AI-powered platform economics. We analyze two types of interventionsâ€”one targeting the social media

[8] We prove in Appendix A that an extremal Nash equilibrium, this time defined only with reference to the dominant party, exists even when H is not symmetric (though the equilibrium will typically not be anti-symmetric).

mechanism and another focusing on the digital ads mechanismâ€"to study their potential impact on polarization dynamics.

6.1 Diversification Standards for Social Media

One tool that can be useful for countering the polarizing effects of AI-powered social media is a diversification standard, as advocated by Sunstein (2018). Let us consider the following type of diversification standard, denoted by γ: If the platform chooses a sharing network $\mathbf{P}$ that belongs to the class of island networks with parameters (k, p_s, p_d) (where k is the number of islands, p_s is the within-island link probability, and p_d is the between-island link probability), then the chosen network has to satisfy $p_s/p_d \equiv p^* \leq \gamma$. Notice that for $\gamma < \infty$ such a constraint will bind when $r < r_P$ because, as we have seen, the platform's profit-maximizing sharing network takes the form of an island network with $p_s = 1$ and $p_d = 0$. We think of γ as a *diversification* standard, requiring that agents be exposed to a certain fraction of cross-cutting perspectives.

Proposition 6.1. *There exists $\bar{r} < r_2$ such that for an article of reliability $r \in (0, \bar{r})$ and diversification standard $\gamma < \infty$, the platform chooses $\mathbf{P}$ from the class of island models (satisfying $p^* \leq \gamma$), and $\tilde{x}_L^* > x_L^{(FB)}$ and $\tilde{x}_R^* < x_R^{(FB)}$.*

Note that Proposition 6.1 follows almost immediately from the proof of Proposition 5 in Acemoglu, Ozdaglar, and Siderius (2024). The key insight is that for a sufficiently light diversification standard γ, the platform is still incentivized to choose a sharing network that follows the (p_s, p_d) island structure from before. The only difference is that now some voters from the left will be exposed to counter-attitudinal right-wing articles, and vice-versa, and this can make the electorate less polarized and moderate party positions.[9]

6.2 Digital Ad Transparency

Another regulation, especially relevant in the context of the digital ads channel, is increasing the transparency of digital ads, which can be thought to reduce the fraction of naïve agents in the population (increase λ) or reduce the effectiveness of ad targeting (increase ϕ), or both. We let λ' and ϕ' be the fraction of sophisticated agents and the power of the ad targeting post-intervention, whereas λ and ϕ are these corresponding quantities before the intervention.

Proposition 6.2. *Suppose $\phi < \phi^*(\lambda)$. There exist $\bar{\lambda} > 0$ and $\bar{\phi} > 0$ such that if $\lambda' > \bar{\lambda}$ or $\phi' > \bar{\phi}$, then digital ad transparency moderates electorate beliefs moderate to $\bar{H} \prec \tilde{H}$. Consequently, political parties choose more moderate positions $(\bar{x}_L^*, \bar{x}_R^*)$ with $\bar{x}_L^* > \tilde{x}_L^*$ and $\bar{x}_R^* < \tilde{x}_R^*$ in the extremal anti-symmetric Nash equilibrium.*

Proposition 6.2 highlights the potential benefits of reducing the influence of targeted digital ad on naïve voters. Note that Proposition 6.2 does not follow directly from Proposition 5.1, which requires

[9] Bail et al. (2018) show that over-exposure to counter-attitudinal news can have "backfire", leading to greater polarization. This suggests that too tight diversification standards may not achieve their desired intent.

both λ and ϕ to lie in a particular orthant of the parameter space, whereas Proposition 6.2 consider a situation in which either λ *or* ϕ is increased.

We also note that a complementary tool to digital ad transparency is improving digital literacy among users. This might act in a similar fashion by making targeted digital ads less effective in influencing naïve voters. Relatedly, Hudders, Van Reijmersdal, and Poels (2019) and Hobbs, Kanižaj, and Pereira (2019) discuss digital literacy and how it can make users more aware of misleading advertisements. Another complementary policy tool would be regulations that reduce the ability of the advertisement technology to micro-target very specific user groups and/or regulations on how generative AI can be used in such advertisements (see Simchon, Edwards, and Lewandowsky (2024) and Golab-Andrzejak (2023)).

6.3 Additional Interventions

Several insights from Acemoglu, Ozdaglar, and Siderius (2024) and Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024) can also be adapted to this environment, but we do not do so in order to avoid repetition. In particular, as in Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024), we examine how digital ad taxes might influence subscription-based business models and, through this channel, alter the incentives for polarizing voter beliefs and consequently party positions. Our analysis suggests that digital ad taxes could lead to lower λ^* and ϕ^* in Proposition 5.1 and Proposition 5.2. The proof follows directly from Proposition 15 in Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024): a tax on digital ad revenues would require that parties compensate the platform further for the corresponding loss in subscription revenues. For this to be profitable, a greater fraction of the population would need to be susceptible to ad targeting (lower λ) or digital ads would need to be more persuasive (lower ϕ), or some combination of the two. Additionally, we consider how watermarking schemes that tag reliable content on social media, along with other regulations, might also affect political polarization. However, as emphasized in Acemoglu, Ozdaglar, and Siderius (2024), these interventions can also generate offsetting effects, and if not carefully designed, they may have negative unintended consequences.

7 Concluding Remarks

It is now widely recognized that AI will transmute not just our economic and social lives, but also our politics. Yet, there is relatively little analysis of how, and through what channels, this political impact will work out in different kinds of societies. In this paper, we develop models to understand the effects of AI-powered social media and entertainment platforms on polarization and the political economy.

We focused on two types of influences, which we explored separately. In the first, which we refer to as the *social media channel,* AI's impacts work through the algorithms that a social media platform utilizes for determining the news- and content-feed of users, with the aim of maximizing engagement and the resulting digital ad revenue. The collection of data from users and powerful AI tools enable the platform to create "filter bubbles", whereby users only engage with the content passed on by like-minded other users. We established, following our previous work in Acemoglu, Ozdaglar, and Siderius (2024), that with low-reliability content, manufacturing filter bubbles is a profit-maximizing strategy

for the platform, because it encourages the viral spread of news items, which would have otherwise been stopped and tagged as misinformation by users with different beliefs. We showed that such filter bubbles lead to the polarization of voters that already hold extreme views. Notably, even though moderate voters on the left and the right may not become much more polarized, political parties respond to the polarization at the tails by choosing more extreme policy positions.

In the second, which we call the *digital ads channel*, we consider an entertainment platform that intermixes content valued by users and digital ads from political parties. Following Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024), we supposed that there are both naïve and sophisticated users, which have different susceptibilities to individualized ads, prepared and targeted by means of AI tools. We also allowed the platform to choose a business model based on subscriptions or one monetized via digital ads. We showed that with two parties from the two ends of the spectrum locked in neck-and-neck competition, monetization via digital ads is more profitable, and these digital ads sway naïve voters to be more attached to their party. As a result, the electorate once again becomes more polarized, and parties similarly respond to such voter polarization by choosing more extreme policy positions. We show, in contrast, that the same outcome does not occur when platforms use a subscription-based model, and we also establish that it is the nature of political competition that makes digital ads more profitable; in fact, with one dominant party, the profit-maximizing business model would be the subscription-based one without this type of polarization. Hence, it is the interplay of AI-powered digital ads targeting naïve users and intense political competition between the two parties that undergirds voter polarization, which then paves the way to political party polarization.

Overall, our analysis of both channels indicates that AI-powered platforms act as a source of polarization among voters, and political parties respond to this by choosing more polarized policy positions. We focused on the two channels separately, because they work in distinct ways and are impacted by different regulations differentially. In the social media channel, the AI's role is to enable sufficient data collection and processing in order to determine each user's ideological leanings and then create custom-made filter-bubbles that make left-leaning (right-leaning) users more exposed to more left-leaning (right-leaning) content being transmitted by other left-leaning (right-leaning) users. For this reason, we also indicated how diversification standards for social media (e.g., as suggested by Sunstein (2018)) could be an effective deterrent to further polarization. In the digital ads channel, AI is used for designing and targeting persuasive digital ads for naïve users (with high susceptibility to targeted ads) from the two ends of the political spectrum. This suggests different tools, in particular, digital ad transparency measures could be effective by reducing the susceptibility of naïve users to such ads (see Simchon, Edwards, and Lewandowsky (2024) and Golab-Andrzejak (2023)). Additionally, digital ad taxes that push platforms towards a subscription-based model, as considered in Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024), could be effective in this case.

We view our paper as a first step in a vast landscape that involves multi-faceted interactions between AI and politics. To start with, much more empirical work is necessary to delineate how online articles and messages impact people's political beliefs and how these translate into voting and other political actions in the off-line world. Additionally, we have only touched upon a few of the ways in which AI impacts politics. It is useful to briefly discuss some of the other pathways, which are in

principle equally important and also equally under-researched.

First, AI tools can create new ways in which both non-party actors (such as civil society organizations or extremist groups) and political parties can reach users in democracies. A few reports already discuss how platform algorithms created "rabbit holes" towards extremist groups in the 2010s, and more powerful AI models can significantly amplify these capabilities (see Brown et al. (2022) and O'Callaghan et al. (2015)). There are also understandable concerns about deepfakes enabling more manipulative outreach by both established and fringe actors (e.g., Veerasamy and Pieterse (2022) and Langguth et al. (2021)).

Second, AI can likewise provide new capabilities to nondemocratic regimes for indoctrination and spreading false information. There is relatively little work on what the limits of such indoctrination may be (for example, will the population ultimately understand the new capabilities in this domain and become immune to such indoctrination?). There is both some theoretical work (e.g., Mostagir, Ozdaglar, and Siderius (2022)) and empirical work (e.g., Hristakieva et al. (2022)) that scratch the surface of such questions, while also raising important directions for future work in this area.

Third, existing research (e.g., Zeng (2020), Beraja et al. (2023) and Beraja, Yang, and Yuchtman (2023)) shows that AI capabilities have already been used effectively for surveillance and quelling dissent, which can transform politics both in nondemocratic and democratic systems. Nevertheless, there are relatively few theoretical and empirical studies of how intensified surveillance affects politics.

Fourth, a wealth of AI-generated content can reduce the population's interest and engagement in politics, with fairly uncertain but wide-ranging social and political implications. A version of this effect was anticipated by the philosopher Hannah Arendt when she wrote "If everybody lies to you, the consequence is not that you believe the lies, but rather that nobody believes anything any longer"(Berkowitz (2024)). In practice, this and many other responses are possible in a world with more targeted, potentially more manipulative messages online (see Carroll et al. (2023), Chen and Papanastasiou (2021), and Mostagir and Siderius (2022a)).

Fifth, AI tools can conceivably enable users to check the veracity and reliability of claims and messages in social media, or can be deployed for creating new platforms for more reliable political communication and even new ways for consensus and compromise-building (see some discussions of this issue in Chapter 10 of Acemoglu and Johnson (2023)). These AI-driven systems could assess the credibility of sources, detect inconsistencies in narratives, and provide users with context or alternative perspectives to improve informed decision-making (e.g., see Gabriel et al. (2024)). Additionally, AI could facilitate real-time content moderation by flagging or debunking misleading claims before they gain traction (see in Lu et al. (2022)). Beyond fact-checking, AI-powered platforms could foster more constructive political discourse by structuring debates, summarizing key arguments, and even suggesting areas of compromise based on common ground between opposing viewpoints (see Huerta et al. (2023)).

Finally, the myriad other effects of AI working through automation, job loss, wealth and income inequality, and productivity-enhancement can also have downstream effects on politics. Automation increases productivity but also displaces jobs, particularly in middle-skill occupations, worsening income and wealth inequality (Gallego and Kurer (2022), Chen, Li, and Tang (2022), Acemoglu and

Restrepo (2020)). These shifts can fuel populist movements, reshape labor policies, and drive debates on AI regulation and social safety nets. Moreover, AI-driven economic concentration may alter political alignments and policymaking, as wealth and power become increasingly centralized in tech-centric regions. AI's economic disruptions have significant political consequences, influencing governance, policy debates, and public sentiment.

We view all of these areas as fruitful directions for future research.

A Equilibria with a Dominant Party

Suppose there is a dominant party in power, which implies that the distribution of prior ideological beliefs, H, may not necessarily symmetric (i.e., it is skewed toward the dominant party). This yields a slightly different version of the result in Lemma 1. We will leverage this result in Section 5.4.

Lemma A.1.

(i) There exists a Nash equilibrium (x_L^*, x_R^*) in potentially asymmetric strategies with $x_L^* < x_R^*$;

(ii) There exists an extremal Nash equilibrium $(\bar{x}_L^*, \bar{x}_R^*)$ with $\bar{x}_R^* \geq x_R^*$ for all Nash equilibria (x_L^*, x_R^*);

(iii) If $\tilde{H} \succ_{FOSD} H$, then for every Nash equilibrium (x_L^*, x_R^*) under belief distribution H, there exists a Nash equilibrium $(\tilde{x}_L^*, \tilde{x}_R^*)$ such that $\tilde{x}_L^* > x_L^*$ and $\tilde{x}_R^* > x_R^*$.

Proof of Lemma A.1. For part (i), consider the function $\varphi : [0,1]^2 \to [0,1]^2$ that maps an arbitrary position (x_L, x_R) into the best response positions (x_L^{BR}, x_R^{BR}), where the right-wing party takes the left-wing's party position as given, and vice-versa. Similar to the proof of Lemma 1, we have a continuous mapping from a compact, convex set onto itself. By Brouwer's theorem we are guaranteed a fixed point of φ such that $\varphi(x_L^*, x_R^*) = (x_L^*, x_R^*)$, which is the definition of a Nash equilibrium.

Then it just remains to show we cannot have $x_L^* = x_R^*$. We argue by contradiction, that $x_L^* = x_R^*$, which means at least one party is not at its ideal point, let's say party R without loss of generality. Then, when $x_L^* = x_R^*$, one can show similar to in Lemma 1 that

$$\frac{\partial}{\partial x_R}\mathcal{R} \cdot \int_0^1 \frac{\exp(-(\beta - x_R)^2)}{\exp(-(\beta - x_R)^2) + \exp(-(\beta - x_L^*)^2)} h(\beta)\, d\beta \Big|_{x_R = x_R^* = x_L^*} = 0,$$

whereas $\frac{\partial}{\partial x_R}(x_R - \hat{x}_R)^2 = 2(x_R - \hat{x}_R) \neq 0$. Thus, there is a profitable deviation of party R closer to its ideal policy point, contradicting that this is an equilibrium. This establishes part (i).

To see part (ii), it is sufficient to argue that there exist finitely many Nash equilibria. Suppose that (x_L^*, x_R^*) is a Nash equilibrium and take a neighborhood $B_\epsilon(x_L^*, x_R^*)$ of radius $\epsilon > 0$. Now consider $\varphi' = \varphi - I$ where $I(\cdot)$ is the identity function. Then $\varphi'(x_L^*, x_R^*) = 0$ and given that 1 is not an eigenvalue of $d\varphi$ (which holds generically), we have that $\det(d\varphi') \neq 0$, and hence φ' is a local diffeomorphism by the inverse function theorem, implying that φ' is bijective over $B_\epsilon(x_L, x_R)$. This prevents another point $(x_L', x_R') \in B_\epsilon(x_L, x_R)$ such that $\varphi'(x_L', x_R') = 0$ (i.e., in other words, it bars the existence of another fixed point in the open-neighborhood of (x_L, x_R)). Finally we leverage that $[0,1]^2$ is compact: Take an open cover around each fixed points of φ and the complement of the set of fixed points. There exists a finite subcover, so there must exist finitely many fixed points of φ.

To see part (iii), we define the function $\varphi : [x_L^*, 1] \times [x_R^*, 1] \to [x_L^*, 1] \times [x_R^*, 1]$ as before but defined on a subspace of $[0,1]^2$, where we map arbitrary cutoffs in $(x_L, x_R) \in [x_L^*, 1] \times [x_R^*, 1]$ to best-response cutoffs (x_L^{BR}, x_R^{BR}) under distribution $\tilde{H}$. To see this is a well-defined mapping, note that for any $x_L \in [x_L^*, 1]$ it is a best response for party R to choose a position at least at x_R^*, and similarly for any $x_R \in [x_R^*, 1]$ it is a

best response for party L to choose a position at least at x_L^*, given the first-order stochastic dominance shift in H to $\tilde{H}$. We can see this by noting that

$$\frac{\partial}{\partial x_R}\mathcal{R}\cdot\int_0^1 \frac{\exp(-(\beta-x_R)^2)}{\exp(-(\beta-x_R)^2)+\exp(-(\beta-x_L^*)^2)}h(\beta)\,d\beta - 2(x_R^*-\hat{x}_R)\Big|_{x_R=x_R^*} = 0,$$

by optimality of x_R^* for party R, which implies that

$$\frac{\partial}{\partial x_R}\mathcal{R}\cdot\int_0^1 \frac{\exp(-(\beta-x_R)^2)}{\exp(-(\beta-x_R)^2)+\exp(-(\beta-x_L^*)^2)}\tilde{h}(\beta)\,d\beta - 2(x_R^*-\hat{x}_R)\Big|_{x_R=x_R^*} > 0,$$

and moreover $\partial x_R^*/\partial x_L^* > 0$ everywhere. Thus, φ has a fixed point in this subspace of $[0,1]^2$, exactly as in claim (iii). ∎

B Proofs

Proof of Lemma 1. To prove (i), we establish the existence of an anti-symmetric Nash equilibrium. The utility for party j is equal to $U_p = R\cdot\mathbb{P}[\theta \geq 1-\chi_j] - (x-\hat{x}_i)^2$, which is equivalent to

$$U_j = \mathcal{R}\cdot\int_0^1 \frac{\exp(-(\beta-x_i)^2)}{\exp(-(\beta-x_i)^2)+\exp(-(\beta-x_j)^2)}h(\beta)\,d\beta - (x_i-\hat{x}_i)^2.$$

First, We argue that $x_L^* \leq \frac{1}{2} \leq x_R^*$. Suppose not, and without loss generality, consider $x_R^* < \frac{1}{2}$. Then

$$\frac{\partial U_R(x_R^*)}{\partial x_R^*} = \mathcal{R}\cdot\int_0^1 \frac{2(\beta-x_R^*)e^{-(\beta-x_R^*)^2-(\beta-x_L^*)^2}}{(e^{-(\beta-x_L^*)^2}+e^{-(\beta-x_R^*)^2})^2}h(\beta)\,d\beta - 2(x_R^*-\hat{x}_R) > 0,$$

given that $x_L', x_R' < \frac{1}{2}$, and H is symmetric. To see this, we just consider:

$$\begin{aligned}
&\int_0^1 \frac{2(\beta-x_R^*)e^{-(\beta-x_R^*)^2-(\beta-x_L^*)^2}}{(e^{-(\beta-x_L^*)^2}+e^{-(\beta-x_R^*)^2})^2}h(\beta)\,d\beta \\
&= \int_0^{x_R^*} \frac{2(\beta-x_R^*)e^{-(\beta-x_R^*)^2-(\beta-x_L^*)^2}}{(e^{-(\beta-x_L^*)^2}+e^{-(\beta-x_R^*)^2})^2}h(\beta)\,d\beta + \int_{x_R^*}^{1/2} \frac{2(\beta-x_R^*)e^{-(\beta-x_R^*)^2-(\beta-x_L^*)^2}}{(e^{-(\beta-x_L^*)^2}+e^{-(\beta-x_R^*)^2})^2}h(\beta)\,d\beta \\
&+ \int_{1/2}^{1/2+x_R^*} \frac{2(\beta-x_R^*)e^{-(\beta-x_R^*)^2-(\beta-x_L^*)^2}}{(e^{-(\beta-x_L^*)^2}+e^{-(\beta-x_R^*)^2})^2}h(\beta)\,d\beta + \int_{1/2+x_R^*}^1 \frac{2(\beta-x_R^*)e^{-(\beta-x_R^*)^2-(\beta-x_L^*)^2}}{(e^{-(\beta-x_L^*)^2}+e^{-(\beta-x_R^*)^2})^2}h(\beta)\,d\beta
\end{aligned}$$

And note by symmetry of H (and by virtue of $x_L^*, x_R^* < 1/2$), the first integral is smaller in absolute value than the third integral, while the second and fourth integral are strictly positive. At the same time $2(x_R^*-\hat{x}_R) < 0$ given $x_R^* < 1/2 < \hat{x}_R$. Thus, it is a profitable deviation for party R to increase its x_R from any policy point $x_R^* < 1/2$, a contradiction of x_R^* being an equilibrium. An analogous argument shows that $x_L^* \leq \frac{1}{2}$, so we must have $x_L^* \leq \frac{1}{2} \leq x_R^*$.

Next, we consider the map $\varphi : [0,1] \to [0,1]$ that maps an arbitrary position x_L into the comple-

mentary best-response position $1 - x_R^{\text{BR}}$, where x_R^{BR} solves the following (with x_L taken as given):

$$\max_{x_R \in [0,1]} U_R = \mathcal{R} \cdot \int_0^1 \frac{\exp(-(\beta - x_R)^2)}{\exp(-(\beta - x_L)^2) + \exp(-(\beta - x_R)^2)} h(\beta)\, d\beta - (x_R - \hat{x}_R)^2.$$

Utility U_R is quasi-concave when H is symmetric, so it must be the case that this admits a unique $x_R^*(x_L)$ which is a global maximum for U_R over all of $[0,1]$ (see Mas-Colell et al. (1995)). Moreover, by the inverse function theorem, the solution $x_R^*(x_L)$ is continuous in x_L. By Brouwer's theorem, there exists a fixed point of φ such that $x_R^* = x_R^{\text{BR}}(1 - x_R^*)$, which by symmetry implies that $x_L^* = x_L^{\text{BR}}(x_R^*)$, and thus (x_L^*, x_R^*) is an anti-symmetric Nash equilibrium by construction.

For part (ii), it is sufficient to argue that there exist finitely many anti-symmetric Nash equilibria, and then we can let $(\underline{x}_L^*, \bar{x}_R^*)$ denote the most extremal anti-symmetric Nash equilibrium. We use the Lefschetz fixed-point theorem to establish this (see Dold (2012)). We prove that each fixed point of φ is a Lefschetz fixed point (i.e., it is isolated); consider a fixed point $(x_L^*, 1 - x_L^*)$. Note that there exists an open neighborhood $B_\epsilon(x_L^*)$ where $x_R^*(x_L)$ is bijective over $B_\epsilon(x_L^*)$ for small enough ϵ given that $\frac{\partial x_R^*}{\partial x_L}\Big|_{x_L = x_L^*, x_R = 1 - x_L^*} \neq 1$, which holds for generic $\mathcal{R}$ by the inverse function theorem.

Thus, there exists a unique fixed point $x_L^* \in B_\epsilon(x_L^*)$ for small enough $\epsilon > 0$. Finally, it just remains to consider the following open cover of $[0,1]$: Take $B_\epsilon(x_L^*)$ around each of the fixed points x_L^*, and the complement of the set of fixed points (which is open because the set of fixed points is always closed). Since $[0,1]$ is compact, there must be finitely many subcovers, which in particular, implies there exist finitely many Lefschetz fixed points.

By contradiction, suppose there exist infinitely many fixed points, so for any $\varepsilon > 0$, there exists $|\tilde{x}_L^* - \tilde{\tilde{x}}_L^*| < \varepsilon$ such that

$$\frac{\partial U_R(x_R^*)}{\partial x_R^*} = \mathcal{R} \cdot \int_0^1 \frac{2(\beta - x_R^*) e^{-(\beta - x_R^*)^2 - (\beta - \underline{x}_L^*)^2}}{(e^{-(\beta - \underline{x}_L^*)^2} + e^{-(\beta - x_R^*)^2})^2} h(\beta)\, d\beta + 2(x_R^* - \hat{x}_R)$$

Consequently, fixed points are isolated As such, we consider open neighborhoods around the fixed points of $(x_L^*, 1 - x_L^*)$ and take an open cover around the fixed points of φ and the complement of the set of fixed points. Existence of a finite subcover (given that $[0,1]$ is compact) guarantees there exist finitely many fixed points of φ.

Finally, it remains to show that $x_L' = x_R' = \frac{1}{2}$ is not an equilibrium. Note that in this case we have that for party R, their payoff is given by

$$U_i = \mathcal{R} \cdot \int_0^1 \frac{\exp(-(\beta - x_R)^2)}{\exp(-\beta - x_R)^2 + \exp(-\beta - \frac{1}{2})^2} h(\beta)\, d\beta + (x_R - \hat{x}_R)^2$$

Taking a first-order condition with respect to x_R evaluated at proposed equilibrium $x_R' = \frac{1}{2}$, we see that

$$\begin{aligned}\frac{\partial U_i}{\partial x_R}\left(x_R = \frac{1}{2}\right) =& \mathcal{R} \cdot \int_0^1 \frac{4(\beta - \frac{1}{2})\exp(-2(\beta - \frac{1}{2})^2) - 4(\beta - \frac{1}{2})\exp(-2(\beta - \frac{1}{2})^2)}{4\exp(-2(\beta - \frac{1}{2})^2)} h(\beta)\, d\beta + 2(x_R - \hat{x}_R) \\ =& 2(x_R - \hat{x}_R)\end{aligned}$$

Thus, party R has an incentive to move some small ε to the right of $x'_R = \frac{1}{2}$ (closer to its policy ideal point), while losing only a second-order fraction of voters by doing so. This is contradiction of $x'_L = x'_R = \frac{1}{2}$ being an equilibrium, as we wanted to show.

For part (ii), suppose for distribution H we have an extremal anti-symmetric Nash equilibrium $(\underline{x}^*_L, \bar{x}^*_R)$ and let us consider distribution $\tilde{H}$ and extremal symmetric Nash equilibrium $(\tilde{\underline{x}}^*_L, \tilde{\bar{x}}^*_R)$. Under $\tilde{H}$, we know that the best-response of x_R given $\underline{x}^*_L$ is some $x^*_R > \bar{x}^*_R$ (and conversely, the new best-response position for party L satisfies $x^*_L < \underline{x}^*_L$ given $\bar{x}^*_R$). To see this, first notice that we have the following first-order condition

$$\frac{\partial U_R(x^*_R)}{\partial x^*_R} = \mathcal{R} \cdot \int_0^1 \frac{2(\beta - x^*_R)e^{-(\beta - x^*_R)^2 - (\beta - \underline{x}^*_L)^2}}{(e^{-(\beta-\underline{x}^*_L)^2} + e^{-(\beta - x^*_R)^2})^2} h(\beta)\, d\beta - 2(x^*_R - \hat{x}_R) = 0,$$

whereas under $\tilde{H}$, we leverage Lemma A.3 in Acemoglu, Ozdaglar, and Siderius (2024), and compute the gain in voter share if an agent were originally at belief b_i under H (but now has belief $b_i + \epsilon_i$ under $\tilde{H}$) and consider a deviation slightly to the right:

$$\begin{aligned}\frac{\partial U_R}{\partial x^*_R} &= \mathcal{R} \cdot \int_0^{1/2} \frac{2(\beta - \epsilon(\beta) - x^*_R)e^{-(\beta-\epsilon(\beta)-x^*_R)^2 - (\beta-\epsilon(\beta)-\underline{x}^*_L)^2}}{(e^{-(\beta-\epsilon(\beta)\underline{x}^*_L)^2} + e^{-(\beta-\epsilon(\beta)-x^*_R)^2})^2} h(\beta)\, d\beta \\ &+ \mathcal{R} \cdot \int_{1/2}^1 \frac{2(\beta + \epsilon(\beta) - x^*_R)e^{-(\beta+\epsilon(\beta)-x^*_R)^2 - (\beta+\epsilon(\beta)-\underline{x}^*_L)^2}}{(e^{-(\beta+\epsilon(\beta)-\underline{x}^*_L)^2} + e^{-(\beta+\epsilon(\beta)-x^*_R)^2})^2} h(\beta)\, d\beta - \mathcal{R} \cdot \int_0^1 \frac{2(\beta - x^*_R)e^{-(\beta-x^*_R)^2 - (\beta-\underline{x}^*_L)^2}}{(e^{-(\beta-\underline{x}^*_L)^2} + e^{-(\beta-x^*_R)^2})^2} h(\beta)\, d\beta.\end{aligned}$$

By the fact that $\epsilon(\beta) = \epsilon(1-\beta)$ and via symmetry of H, we will argue that this expression is positive. To see this, observe that letting $f(\beta) = 2(\beta - x^*_R)\frac{e^{-(\beta-x^*_R)^2-(\beta-\underline{x}^*_L)^2}}{(e^{-(\beta-\underline{x}^*_L)^2}+e^{(-\beta-x^*_R)^2})^2}$, we obtain a relationship between the integrals $\int_0^{1/2} f(\beta-\epsilon(\beta), x^*_R, x^*_L)h(\beta)\, d\beta$, $\int_{1/2}^1 f(\beta+\epsilon(\beta), x^*_R, x^*_L)h(\beta)\, d(\beta)$ and $\int_0^1 f(\beta, x^*_R, x^*_L)h(\beta)\, d(\beta)$. Via second-order Taylor series approximation, we have that for a symmetric distribution H,

$$\int_0^{1/2} f(\beta - \epsilon(\beta))h(\beta)\, d(\beta) + \int_{1/2}^1 f(\beta + \epsilon(b))h(\beta)\, d\beta \approx \int_0^1 \left(f(\beta) + \frac{1}{2}\epsilon(\beta)^2 f''(\beta)\right) h(\beta)\, d(\beta)$$

and correspondingly we can see that our original expression can be approximated by $\frac{1}{2}\int_0^1 \epsilon(\beta)^2 f''(\beta)h(\beta)\, d\beta$, which determines the sign of $\partial U_R/\partial x^*_R$. It just suffices then to look at $\text{sgn}(\int_0^1 \epsilon(\beta)^2 f''(\beta)h(\beta)\, d\beta)$, which requires that $\text{sgn}(f''(1-\beta) + f''(\beta)) > 0$ for all β given that H is symmetric, which can be straightforwardly verified when evaluated at $\bar{x}^*_R > 1/2$, given that $\underline{x}^*_L = 1 - \bar{x}^*_R < 1/2$, as follows:

$$\begin{aligned}f'(\beta) &= -\frac{2e^{(\beta-x^*_R)^2+(\beta-\underline{x}^*_L)^2}}{\left(e^{(\beta-x^*_R)^2} + e^{(\beta-\underline{x}^*_L)^2}\right)^3} \cdot \left(e^{(\beta-x^*_R)^2}\left(-1 - 2\beta x^*_R + 2(x^*_R)^2 + 2\beta\underline{x}^*_L - 2x^*_R\underline{x}^*_L\right)\right. \\ &\left. + e^{(\beta-\underline{x}^*_L)^2}\left(-1 - 2(x^*_R)^2 + 2\beta(x^*_R - \underline{x}^*_L) + 2x^*_R\underline{x}^*_L\right)\right) \\ f''(\beta) &= -\frac{8(x^*_R - \underline{x}^*_L)e^{(\beta-x^*_R)^2+(\beta-\underline{x}^*_L)^2}}{\left(e^{(\beta-x^*_R)^2} + e^{(\beta-\underline{x}^*_L)^2}\right)^4} \cdot \left(e^{2(\beta-x^*_R)^2}\left(\beta(\underline{x}^*_L - x^*_R) + (x^*_R)^2 - x^*_R\underline{x}^*_L - 1\right)\right. \\ &\left. + e^{2(\beta-\underline{x}^*_L)^2}\left(\beta(\underline{x}^*_L - x^*_R) + (x^*_R)^2 - x^*_R\underline{x}^*_L + 1\right) + 4(\beta - x^*_R)(x^*_R - \underline{x}^*_L)e^{(\beta-x^*_R)^2+(\beta-\underline{x}^*_L)^2}\right).\end{aligned}$$

Plugging in β and $1-\beta$ to $f''(\cdot)$ and summing yields the desired claim of positivity. Analogously, we see that the best-response of x_L given $\bar{x}_R^*$ satisfies $x_L^* < \underline{x}_L^*$. The monotonicity of the equilibrium cutoffs $(\tilde{x}_R, \tilde{x}_L)$ under transformations of $\tilde{H}$ then follows from Milgrom and Shannon (1994).

Finally, it just remains to show that $(\underline{x}_L^*, \bar{x}_R^*)$ is not an equilibrium under $\tilde{H}$. But this has already been done in the first step above, by showing that party R has a profitable deviation to increase x_R above $\bar{x}_R^*$ (when party L plays $\underline{x}_L^*$) and party L has a profitable deviation to decrease x_L below $\underline{x}_L^*$ (when party R plays $\bar{x}_R^*$), given a shift in the underlying belief distribution to $\tilde{H}$. ∎

Proof of Proposition 4.1. This follows immediately from the proof of Theorem 3 in Acemoglu, Ozdaglar, and Siderius (2024). ∎

Proof of Proposition 4.2. First, we claim that if $m = R$, the filter bubble consists of agents with prior beliefs $b_i \in [\underline{b}, 1]$ for some $\underline{b} < 1$ given that $r < r_P$. By the proof of Theorem 3 in Acemoglu, Ozdaglar, and Siderius (2024), the platform will choose the smallest $\underline{b} \in \{0, \varepsilon, \ldots, m\varepsilon, \ldots, 1\}$ such that the fully connected network of beliefs $[\underline{b}, 1]$ induce all these agents to either share or ignore (known as the "filter bubble"). If there were a disconnected region, say $b_1 < b_2 < b_3$ with b_1 and b_3 in the filter bubble, but b_2 not, then the sharing network of the platform is not profit-maximizing, because by adding the agent with b_2 to the network, the payoff to agents with b_1 and b_3 increases or remains constant if agent b_2 shares, whereas b_2 is necessarily willing to share in some equilibrium if agent b_1 is willing to, given π_2 (agent i with belief b_2's posterior belief about article veracity) is higher than π_1. The result then follows from Topkis (1998) given our social media game is supermodular.

Second, we show that $\underline{b}$ increases as r decreases. Once again this is a direct consequence of Topkis (1998) and noting that the payoff to sharing uniformly decreases for all agents in the population as r decreases. Formally, the payoff to sharing is given by $U_i = u\pi_i - c(1-\pi_i) + \kappa S_i - dD_i$, and π_i is decreasing in r, and this has a supermodular form. We thus get monotone comparative statics on the network of agents with priors on $[\bar{b}, 1]$, with some agents potentially now disliking given the decrease in r. Consequently, the profit-maximizing sharing network per Proposition 4.1 (and Theorem 3 in Acemoglu, Ozdaglar, and Siderius (2024)) consists of agents with prior beliefs $[\underline{b}', 1]$ with $\underline{b}' > \underline{b}$.

Third, we show that for any agent i on the platform with prior $b_i < 1$, there exists $\underline{r}$ such that if $r < \underline{r}$, ignoring is a better response than sharing. The payoff to sharing is upper bounded by $u\pi_i - c(1-\pi_i) + \kappa N$ and the payoff to ignoring is 0. Whenever $\pi_i < \frac{c-\kappa N}{u-c}$, the payoff to ignoring is higher, and since $\pi_i \to 0$ as $r \to 0$ and $\kappa < c/N$ (see Section 2.6 of Acemoglu, Ozdaglar, and Siderius (2024)), there exists such a value of r sufficiently small such that the agent of type b_i will not share. Moreover, we show there exists $\underline{r}'$ such that if $r < \underline{r}'$, disliking is a better response to ignoring, which makes disliking the best response. To see this, notice that the payoff to disliking is $\tilde{u}(1-\pi_i) - \tilde{c}$, and $1 - \pi_i \to 1$ as $r \to 0$ and $\tilde{u} > \tilde{c}$. This implies that agent i will not be included in the filter bubble in the profit-maximizing sharing network. ∎

Proof of Proposition 4.3. First, notice that if an agent receives two articles of opposing messages with the same reliability, her posterior belief $\tilde{b}_i$ will remain unchanged. This is due to the symmetry of the two messages and the Martingale property of Bayesian beliefs; if their posterior changed after seeing both $m = L$ and $m = R$ of the same reliability r, then their prior belief must be miscalibrated. Thus,

$\tilde{H}_C = H$, and the beliefs of the agents before and after the news sharing remains identical, so it just suffices to prove that $\tilde{H}_{FB} \succ H$.

It is clear from Proposition 4.1 and Proposition 4.2 that, for any $\underline{b}$, we can pick a value for r such that the agent's with priors $[\underline{b}, 1]$ are the ones in the platform's profit-maximizing filter bubble (again, assuming message $m = R$). Let us pick r_2 such that $\underline{b} > H^{-1}(3/4)$, so then agents who receive article $m = L$ will be those with prior beliefs at most $b_i \in [0, 1 - \underline{b}]$ and the agents who receive article $m = R$ will be those with prior beliefs at most $b_i \in [\underline{b}, 1]$. Consequently, agents with beliefs $b_i \in [H^{-1}(1/4), H^{-1}(3/4)]$ do not view any article, and thus do not update their beliefs, so $\tilde{H}$ remains consistent with H over all beliefs $b_i \in [H^{-1}(1/4), H^{-1}(3/4)]$.

It finally remains to prove that agents with $b_i \in [\underline{b}, 1]$ who view an article $m = R$ update their beliefs to some $b_i + \epsilon_i$ (by Lemma A.3 from Acemoglu, Ozdaglar, and Siderius (2024)). We find that

$$\tilde{b}_i = \frac{pb_i\pi_i}{pb_i + (1-p)(1-b_i)} + (1-\pi_i)b_i,$$

where

$$\pi_i = \frac{(pb_i + (1-p)(1-b_i))r}{(b_i/2 + (1-b_i)/2)(1-r) + (pb_i + (1-p)(1-b_i))r}.$$

We get the simplification that

$$\tilde{b}_i - b_i = \pi_i \left(\frac{pb_i}{pb_i + (1-p)(1-b_i)} - b_i \right)$$

where clearly $\frac{pb_i}{pb_i+(1-p)(1-b_i)} > b_i$ given $p > 1/2$ and $b_i > 1/2$, so $\tilde{b}_i - b_i > 0$. It follows from an analogous argument that agents with $b_i \in [0, 1 - \underline{b}]$ who view an article $m = L$ update their beliefs to some $b_i - \epsilon_i$. The claim follows by choosing $r_1 < r_2$ but sufficiently large that the filter bubble is non-empty. ∎

Proof of Corollary 4.1. This follows immediately from the results of Proposition 4.3 and Lemma 1(ii). ∎

Proof of Proposition 4.4. Choose $\bar{r}^* = r_P$; then the platform's profit-maximizing sharing network is fully connected. By the same arguments as in Proposition 4.3, posterior beliefs do not change for any agents in the population because they view both articles, so in particular the inequality holds for $\varepsilon = 0$.

Next, note that $\pi_i \to 0$ as $r \to 0$, so it is sufficient to work with π_i. From the proof of Proposition 4.3, for the right-wing agents who receive the $m = R$ article when reliability is low, we have the relationship that

$$\tilde{b}_i - b_i = \pi_i \left(\frac{pb_i}{pb_i + (1-p)(1-b_i)} - b_i \right) \leq \pi_i$$

By continuity of the voting game in Section 2, there exists a small enough ζ such that if $\pi_i \leq \zeta$, then the extremal anti-symmetric Nash equilibrium changes by at most $\varepsilon > 0$ in both $\tilde{x}_L^{FB}$ and $\tilde{x}_R^{FB}$. ∎

Proof of Proposition 5.1. Notice that naïve and sophisticated agents always choose the same online plan because they have an atomless contribution to the outcome of the election and hence do not value the additional information that they will get via digital ads (even when they think such ads are informative, as is the case for naïve agents). Hence, they get no inherent value from advertising, and

they would simply both choose the plan that maximizes $(1-\alpha_\ell)T - P_\ell$. Offering two plans is, therefore, redundant and without loss of generality we can suppose the platform offers a single plan (this is in stark contrast to Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024)).

First, for advertising intensities $(\alpha^{(L)}, \alpha^{(R)})$, the platform makes a maximal advertising revenue from naïfs of $(1-\lambda)m^*(\alpha^{(L)}, \alpha^{(R)})$ and can charge a subscription fee $P^*(\sum_{j\in\{L,R\}} \alpha^{(j)})$, which is decreasing in the total ad load; in particular, the optimal price the platform charges conditional on ad intensity $(\alpha^{(L)}, \alpha^{(R)})$ is just $P^* = (1-\alpha^{(L)} - \alpha^{(R)})T - v$. The platform then solves

$$\max_{(\alpha^{(L)},\alpha^{(R)})} (1-\lambda)m^*(\alpha^{(L)}, \alpha^{(R)}) + P^*\left(\sum_{j\in\{L,R\}} \alpha^{(j)}\right) \tag{3}$$

Let us denote $(\alpha^{(L)*}(\lambda,\phi), \alpha^{(R)*}(\lambda,\phi))$ as the solution to equation (3). Notably, $m^*(\alpha^{(L)*}(\lambda,\phi), \alpha^{(R)*}(\lambda,\phi))$ is decreasing in ϕ by the arguments in the proof of Lemma 2 in Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024). This implies that equation (3) is decreasing in ϕ, as we have for $\phi' > \phi$,

$$\begin{aligned}
&(1-\lambda)m^*_\phi(\alpha^{(L)*}(\lambda,\phi), \alpha^{(R)*}(\lambda,\phi)) + P^*\left(\sum_{j\in\{L,R\}} \alpha^{(j)}(\lambda,\phi)\right) \\
\geq &(1-\lambda)m^*_\phi(\alpha^{(L)*}(\lambda, \phi'^{(R)*}(\lambda, \phi'^* \left(\sum_{j\in\{L,R\}} \alpha^{(j)}(\lambda,\phi')\right) \\
> &(1-\lambda)m^*_{\phi'}(\alpha^{(L)*}(\lambda, \phi'^{(R)*}(\lambda, \phi'^* \left(\sum_{j\in\{L,R\}} \alpha^{(j)}(\lambda,\phi')\right),
\end{aligned}$$

where m^*_ϕ is the ad revenue under parameter ϕ. Fixing λ, we can first check whether setting $\phi = 0$ yields $(\alpha^{(L)*}, \alpha^{(R)*}) = (0,0)$, in which case we set $\phi^*(\lambda) = 0$. Otherwise one can find the supremal value $\phi^* > 0$ such that

$$m^*(\alpha^{(L)*}(\lambda,\phi^*), \alpha^{(R)*}(\lambda,\phi^*)) + P^*\left(\sum_{j\in\{L,R\}} \alpha^{(j)^*}(\lambda,\phi^*)\right) > T - v \tag{4}$$

which has a unique solution $\phi^*(\lambda)$ because the LHS is decreasing in ϕ, the RHS is constant, the function $m^*((\alpha^{(L)*}(\lambda,0), \alpha^{(R)*}(\lambda,0))) + P^*\left(\sum_{j\in\{L,R\}} \alpha^{(j)^*}(\lambda,0)\right) > T - v$ by assumption, and when $\phi = 1/2$, we have $m^*((\alpha^{(L)*}(\lambda,\frac{1}{2}), \alpha^{(R)*}(\lambda,\frac{1}{2}))) + P^*\left(\sum_{j\in\{L,R\}} \alpha^{(j)^*}(\lambda,\frac{1}{2})\right) < T - v$ given that the maximal transfer from the firm satisfies $m^*((\alpha^{(L)*}(\lambda,\frac{1}{2}), \alpha^{(R)*}(\lambda,\frac{1}{2}))) = 0$ and $P^*\left(\sum_{j\in\{L,R\}} \alpha^{(j)^*}(\lambda,\frac{1}{2})\right) < T - v$ unless $\alpha^{(L*)} = \alpha^{(R*)} = 0$. We thus have the desired result, that for $\phi > \phi^*(\lambda)$, the platform offers a fully subscription-based model whereas when $\phi < \phi^*(\lambda)$, the platform offers at least some advertising $\alpha^* = \alpha^{(L)*} + \alpha^{(R)*} > 0$, by construction of ϕ^*.

Second, it remains to show that $\phi^*(\lambda)$ is decreasing in λ. As in the proof of Proposition 7 of Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024), we know that $m^*((\alpha^{(L)*}(\lambda,\phi^*), \alpha^{(R)*}(\lambda,\phi^*)))$ is decreasing in λ, so the expression $(1-\lambda)m^*((\alpha^{(L)*}(\lambda,\phi^*), \alpha^{(R)*}(\lambda,\phi^*)))$ is decreasing in λ, and the LHS

of equation (4) is decreasing in λ (by the same reasoning as the chain of inequalities for ϕ). Thus for higher values of λ, the subscription model becomes more viable for the platform, which in turn implies that $\phi^*(\lambda)$ decreases in λ.

Finally, we argue that the advertising loads are symmetric in equilibrium. To see this, consider the map $\psi : [0,1] \to [0,1]$ that maps left-wing advertiser's load, $\alpha^{(L)}$, to advertising load of the right-wing party, $\alpha^{(R)}$, that would maximize the gains for the right-wing party from advertising. Generically, this map is well-defined (there is a single global maximum, see Ott and Yorke (2005)), continuous and maps a convex set onto itself, and thus by Brouwer's theorem has a fixed point $\psi(\alpha^*/2) = \alpha^*/2$ for some total advertising load $\alpha^* > 0$. By symmetry, when the right-wing party is advertising at $\alpha^*/2$, it is also a best response for the left-wing party to advertise at $\alpha^*/2$. Thus, the proposed contract from the platform to the firm in equilibrium involves $(\alpha^{(L)*}, \alpha^{(R)*}) = (\alpha^*/2, \alpha^*/2)$. ∎

Proof of Proposition 5.2. Part (i) follows immediately from noting that the electorate has $\tilde{H} = H$ under a fully subscription-based model, and so the extremal anti-symmetric Nash equilibrium does not change from (x_L^*, x_R^*).

For part (ii), we use Lemma A.3 in Acemoglu, Ozdaglar, and Siderius (2024) to show that for every $b_i > 1/2$, there exists $\epsilon_i > 0$ such that $\tilde{h}(b_i) \geq h(b_i - \epsilon_i)$, and similarly for every $b_i < 1/2$, there exists $\epsilon_i > 0$ such that $\tilde{h}(b_i) \geq h(b_i + \epsilon_i)$. For the former, this corresponds to some $\frac{1}{2} < b_i' < b_i$ where $b_i' + \epsilon_i' = b_i$ under $\tilde{H}$, and note that there may exist other b_i'''s such that $b_i'' + \epsilon_i'' = b_i$ (hence, the inequality, $\tilde{h}(b) \geq h(b_i - \epsilon_i)$). For the latter, analogously we have that this corresponds to $b_i < b_i' < \frac{1}{2}$, where $b_i = b_i' - \epsilon_i'$ under $\tilde{H}$, and note that there may exist other b_i'''s such that $b_i = b_i'' - \epsilon_i''$ (hence, the inequality $\tilde{h}(b) \geq h(b + \epsilon_i)$). Under the advertising-based model, with probability $1 - e^{-\alpha^* T}$ agents receive a signal $s_i \in \{L, R\}$ with equal probability and update their beliefs as follows:

$$\tilde{b}_i = \begin{cases} \frac{(1-\phi)b_i}{(1-\phi)b_i + \phi(1-b_i)}, \text{ if } m = R \\ \frac{\phi b_i}{\phi b_i + (1-\phi)(1-b_i)}, \text{ if } m = L, \end{cases}$$

whereas with probability $e^{-\alpha^* T}$ agents receive no signal and leave their beliefs unchanged, $\tilde{b}_i = b_i$. Note then that for $\tilde{b}_i > 1/2$, we have that

$$\tilde{h}(\tilde{b}_i) = (1 - e^{-\alpha^* T}) \left(\frac{1}{2} h \left(\frac{\phi \tilde{b}_i}{(1-\tilde{b}_i) + \phi(2\tilde{b}_i - 1)} \right) + \frac{1}{2} h \left(\frac{(1-\phi)\tilde{b}_i}{\phi + \tilde{b}_i - 2\tilde{b}_i \phi} \right) \right) + e^{-\alpha^* T} h(\tilde{b}_i) \geq h(\tilde{b}_i - \tilde{\epsilon}_i)$$

for $\tilde{\epsilon}_i < \tilde{b}_i - \frac{1}{2}$, by virtue of $\phi \in [0, 1/2]$ and convexity of h. To see this, notice that

$$\frac{(1 - e^{-\alpha^* T})}{2} \frac{\phi \tilde{b}_i}{(1-\tilde{b}_i) + \phi(2\tilde{b}_i - 1)} + \frac{(1 - e^{-\alpha^* T})}{2} \frac{(1-\phi)\tilde{b}_i}{\phi + \tilde{b}_i - 2\tilde{b}_i \phi} + e^{-\alpha^* T} \tilde{b}_i < \tilde{b}_i,$$

by the same reasoning as that of Lemma 2 in Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024), where the naïve users drift more toward the extremes, so tend to come from less extreme positions to

begin with. Similarly for agents with $\tilde{b}_i > 1/2$, we can leverage the related fact that

$$\tilde{h}(\tilde{b}_i) = (1-e^{-\alpha^* T})\left(\frac{1}{2}h\left(\frac{\phi\tilde{b}_i}{(1-\tilde{b}_i)+\phi(2\tilde{b}_i-1)}\right)+\frac{1}{2}h\left(\frac{(1-\phi)\tilde{b}_i}{\phi+\tilde{b}_i-2\tilde{b}_i\phi}\right)\right)+e^{-\alpha^* T}h(\tilde{b}_i) \geq h(\tilde{b}_i+\tilde{\epsilon}_i)$$

for $\tilde{\epsilon}_i > \frac{1}{2}-\tilde{b}_i$, by virtue of $\phi \in [0,1/2]$ and convexity of h. This establishes that $\tilde{H} \succ H$, and the resulting anti-symmetric Nash equilibrium shift follows exactly from Lemma 1(ii). ∎

Proof of Corollary 5.1. Given $\lambda > 1/2$, and H_N and H_S do not have overlapping support, we see that the sophisticates always make up the distribution $H^{-1}(\alpha) = \tilde{H}^{-1}(\alpha)$ for all $\alpha \in (1/4, 3/4)$, and do not change their beliefs in response to digital political ads. Conversely, by Proposition 5.1(ii), we see that $\tilde{H}_N \succ H_N$ when $\phi < \phi^*(\lambda)$ and the platform adopts an advertising-based business model. Consequently, $\tilde{H} \succ H$, so applying Lemma 1(ii) we see that the political parties become more polarized in response to increased polarization of the naïfs. ∎

Proof of Proposition 5.3. Recall the utility of the dominant party R is given by

$$U_R = \mathcal{R}\cdot\int_0^1 \frac{\exp(-(\beta-x_R^*)^2)}{\exp(-(\beta-x_R^*)^2)+\exp(-(\beta-x_L^*)^2)}h(\beta;\alpha)\,d\beta-(x_R^*-\hat{x}_R)^2,$$

where the belief distribution $h(\beta;\alpha)$ depends on the advertising α from the dominant party. From advertising, the party improves the first part of its payoff by at most $\frac{\partial}{\partial\alpha}\mathcal{R}(1-\chi_R^*)(1-\lambda)(1-e^{-\alpha T}) = \mathcal{R}\alpha(1-\lambda)(1-\chi_R^*)e^{-\alpha T}$, which comes from assuming the remaining $1-\chi_R$ share who would not vote for party R, would necessarily then vote for party R conditional on viewing a political ad in favor of R (which is a stronger outcome than the political ad would have). Thus, we get the upper bound

$$\frac{\partial}{\partial\alpha}\mathcal{R}\int_0^1 \frac{\exp(-(\beta-x_R^*)^2)}{\exp(-(\beta-x_R^*)^2)+\exp(-(\beta-x_L^*)^2)}h(\beta;\alpha)\,d\beta \leq \mathcal{R}\alpha(1-\lambda)(1-\chi_R^*)e^{-\alpha T} \leq \mathcal{R}\alpha(1-\lambda)(1-\chi_R^*)$$

Moreover, it is easy to see that $\partial x_R^*/\partial\alpha > 0$ from Lemma A.1(ii), which decreases the second part of the payoff to party j under the assumption that $\hat{x}_R < x_R^*$.

Thus, for any positive amount of advertising $\alpha^* > 0$, the party is willing to compensate the firm at most $\mathcal{R}\alpha^*(1-\lambda)(1-\chi_R^*)$, but the platform takes a loss in subscription revenue of $\alpha^* T$, for which it must be compensated for it to be willing to offer an advertising-based plan. Setting $\bar{\chi} > 1-\frac{T}{\mathcal{R}}$ achieves the desired result for all values of $\lambda \in [0,1]$ (and $\phi \in [0,1/2]$) where the platform opts for a subscription-based business model. Consequently, $\tilde{H} = H$ and the set of Nash equilibria (which exist by Lemma A.1(i)) are the same both with and without the potential for political ads, given a dominant party. ∎

Proof of Proposition 6.1. Before the diversification standard, an agent with prior belief $[0,\underline{b}]$ receives an article $m = L$ of reliability $r \in (r_1, r_2)$ with probability 1 and receives an article $m = R$ of the same reliability with probability 0. Similarly, an agent with prior belief $[1-\underline{b}, 1]$ receives an article $m = L$ of reliability $r \in (r_1, r_2)$ with probability 0 and receives an article $m = R$ of the same reliability with probability 1.

By the proof of Proposition 7 in Acemoglu, Ozdaglar, and Siderius (2024), the platform chooses a $\mathbf{P}$ that is still an island network structure but satisfies the diversification standard $p_s/p_d \leq \gamma$. In particular, this implies that an agent with prior belief $[0, \underline{b}]$ receives an article $m = L$ of reliability $r \in (r_1, r_2)$ with probability $\zeta < 1$ and receives an article $m = L$ of the same reliability with probability $\xi > 0$. At the same time an agent with prior belief $[1 - \underline{b}, 1]$ receives an article $m = L$ of reliability r with probability $\xi > 0$ and receives an article $m = R$ of the same reliability with probability $\zeta < 1$.

Notice that agent with prior belief $b_i < \underline{b}$ updates their posterior under the new diversification standard according to:

$$\tilde{b}_i = \begin{cases} \tilde{\underline{b}}_i, \text{with probability } \zeta(1-\xi) \\ b_i, \text{ with probability } (1-\zeta)(1-\xi) + \zeta\xi \\ \tilde{\bar{b}}_i, \text{with probability } (1-\zeta)\xi \end{cases}$$

where $\tilde{\underline{b}}_i$ is the belief from Proposition 4.3 for a left-wing agent and $\tilde{\bar{b}}_i < \frac{1}{2}$ for $\bar{r}$ sufficiently small. Similarly, notice that for an agent with prior belief $b_i > \bar{b}$, she updates her posterior under the new diversification standard according to:

$$\tilde{b}_i = \begin{cases} \tilde{\bar{b}}_i, \text{with probability } \zeta(1-\xi) \\ b_i, \text{ with probability } (1-\zeta)(1-\xi) + \zeta\xi \\ \tilde{\underline{b}}_i, \text{ with probability } (1-\zeta)\xi \end{cases}$$

where $\tilde{\bar{b}}_i$ is the belief from Proposition 4.3 for a right-wing agent and $\tilde{\underline{b}}_i > \frac{1}{2}$ for $\bar{r}$ sufficiently small. It is clear that $\tilde{H} \prec H$ as a result of the diversification standard, so by Lemma 1(ii), we know that both political party positions moderate in response to the standard. ∎

Proof of Proposition 6.2. Similar to the proof of Proposition 14 in Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024), the maximal revenue the platform can collect from the firms (and specifically from the naïfs) is given by $m^*(\alpha^{(L)}, \alpha^{(R)}) \leq L_\lambda(\alpha^{(L)} + \alpha^{(R)})$ for some constant $L_\lambda > 0$. The total advertising revenue is thus bounded above $(1-\lambda)L_\lambda(\alpha^{(L)} + \alpha^{(R)})$, but the loss in subscription revenue from the platform from advertising is $(\alpha^{(L)} + \alpha^{(R)})T$. For $\bar{\lambda} = 1 - T/L_\lambda$, if $\lambda' > \bar{\lambda}$, then the platform opts for a fully subscription-based business model instead, with the sophisticated users protecting the naïve ones from an ad-based plan that fails to generate sufficient revenue to compensate for the loss in subscription. This leads posterior beliefs to equal $\bar{H} = H \prec \tilde{H}$, which by Lemma 1(ii) moderates the positions of the political parties (undoing the effect identified in Proposition 5.2(ii)).

Likewise, the maximal revenue the platform can collect from the firms (and specifically from the naïfs) is given by $m^*(\alpha^{(L)}, \alpha^{(R)}) \leq L_\phi(\frac{1}{2} - \phi)(\alpha^{(L)} + \alpha^{(R)})$ for some constant $L_\phi > 0$, as we did in the proof of Proposition 14 in Acemoglu, Huttenlocher, Ozdaglar, and Siderius (2024). To see this, one can just let L_ϕ to be defined as the maximal change in advertising revenue following a decrease in the technology ϕ, i.e., $L_\phi = \max_{\phi \in [0,1/2]} |\partial m^*(\alpha^{(L)*}(\phi, \lambda), \alpha^{(R)*}(\phi, \lambda))/\partial\phi| < \mathcal{R}$. For $\bar{\phi} = \frac{1}{2} - T/L_\phi$, if $\phi' > \bar{\phi}$, then the platform opts for a fully subscription-based business model instead, leading to the same effect as with $\lambda' > \bar{\lambda}$, following from Lemma 1(ii). ∎

References

Acemoglu, Daron, Victor Chernozhukov, and Muhamet Yildiz (2016), "Fragility of asymptotic agreement under bayesian learning." *Theoretical Economics*, 11, 187–225.

Acemoglu, Daron, Georgy Egorov, and Konstantin Sonin (2013), "A political theory of populism." *The quarterly journal of economics*, 128, 771–805.

Acemoglu, Daron, Daniel Huttenlocher, Asuman Ozdaglar, and James Siderius (2024), "Online business models, digital ads, and user welfare." Technical report, National Bureau of Economic Research.

Acemoglu, Daron and Simon Johnson (2023), *Power and progress: Our thousand-year struggle over technology and prosperity*. Hachette UK.

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius (2024), "A model of online misinformation." *Review of Economic Studies*, 91, 3117–3150.

Acemoglu, Daron and Pascual Restrepo (2020), "Robots and jobs: Evidence from US labor markets." *Journal of political economy*, 128, 2188–2244.

Associated Press (2024), "Influencers play a growing role in election news consumption." URL https://apnews.com/article/influencers-election-trump-harris-news-eacd42bce73d6e11cbc760caf28c993a. Accessed: 2025-01-18.

Bail, Christopher A, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky (2018), "Exposure to opposing views on social media can increase political polarization." *Proceedings of the National Academy of Sciences*, 115, 9216–9221.

Barber, Michael, Nolan McCarty, Jane Mansbridge, and Cathie Jo Martin (2015), "Causes and consequences of polarization." *Political negotiation: A handbook*, 37, 39–43.

Barberá, Pablo (2020), "Social media, echo chambers, and political polarization." *Social media and democracy: The state of the field, prospects for reform*, 34–55.

Baumann, Fabian, Philipp Lorenz-Spreen, Igor M Sokolov, and Michele Starnini (2020), "Modeling echo chambers and polarization dynamics in social networks." *Physical Review Letters*, 124, 048301.

Beraja, Martin, Andrew Kao, David Y Yang, and Noam Yuchtman (2023), "Ai-tocracy." *The Quarterly Journal of Economics*, 138, 1349–1402.

Beraja, Martin, David Y Yang, and Noam Yuchtman (2023), "Data-intensive innovation and the state: evidence from ai firms in china." *The Review of Economic Studies*, 90, 1701–1723.

Berkowitz, Roger (2024), "On fake hannah arendt quotations." *Amor Mundi*.

Berman, Sheri (2021), "The causes of populism in the west." *Annual Review of Political Science*, 24, 71–88.

Braghieri, Luca, Sarah Eichmeyer, Ro'ee Levy, Markus M Mobius, Jacob Steinhardt, and Ruiqi Zhong (2024), "Level slant and polarization of news consumption on social media." *Available at SSRN 4932600*.

Brown, Megan A, James Bisbee, Angela Lai, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker (2022), "Echo chambers, rabbit holes, and algorithmic bias: How youtube recommends content to real users." *Available at SSRN 4114905*.

Burton, Joe (2023), "Algorithmic extremism? the securitization of artificial intelligence (ai) and its impact on radicalism, polarization and political violence." *Technology in society*, 75, 102262.

Butters, Gerard R. (1977), "Equilibrium distributions of sales and advertising prices." *Review of Economic Studies*, 44, 465–491.

Callander, Steven and Juan Carlos Carbajal (2022), "Cause and effect in political polarization: A dynamic analysis." *Journal of Political Economy*, 130, 825–880.

Carroll, Micah, Alan Chan, Henry Ashton, and David Krueger (2023), "Characterizing manipulation from ai systems." In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–13.

Chen, Li and Yiangos Papanastasiou (2021), "Seeding the herd: Pricing and welfare effects of social learning manipulation." *Management Science*, 67, 6734–6750.

Chen, Ni, Zhi Li, and Bo Tang (2022), "Can digital skill protect against job displacement risk caused by artificial intelligence? empirical evidence from 701 detailed occupations." *PLoS One*, 17, e0277280.

Coughlin, Peter J (1992), *Probabilistic voting theory*. Cambridge University Press.

Dixit, Avinash K and Jörgen W Weibull (2007), "Political polarization." *Proceedings of the National Academy of sciences*, 104, 7351–7356.

Dold, Albrecht (2012), *Lectures on algebraic topology*, volume 200. Springer Science & Business Media.

Downs, Anthony (1957), "An economic theory of political action in a democracy." *Journal of political economy*, 65, 135–150.

Druckman, James N and Jeremy Levy (2022), "Affective polarization in the american public." In *Handbook on politics and public opinion*, 257–270, Edward Elgar Publishing.

Feezell, Jessica T, John K Wagner, and Meredith Conroy (2021), "Exploring the effects of algorithm-driven news sources on political behavior and polarization." *Computers in human behavior*, 116, 106626.

Fiorina, Morris P and Samuel J Abrams (2008), "Political polarization in the american public." *Annu. Rev. Polit. Sci.*, 11, 563–588.

Flaxman, Seth, Sharad Goel, and Justin M Rao (2016), "Filter bubbles, echo chambers, and online news consumption." *Public opinion quarterly*, 80, 298–320.

Gabriel, Saadia, Liang Lyu, James Siderius, Marzyeh Ghassemi, Jacob Andreas, and Asu Ozdaglar (2024), "Misinfoeval: Generative ai in the era of" alternative facts"." *arXiv preprint arXiv:2410.09949*.

Gallego, Aina and Thomas Kurer (2022), "Automation, digitalization, and artificial intelligence in the workplace: implications for political behavior." *Annual Review of Political Science*, 25, 463–484.

Glaeser, Edward L (2005), "The political economy of hatred." *The Quarterly Journal of Economics*, 120, 45–86.

Glaeser, Edward L, Giacomo AM Ponzetto, and Jesse M Shapiro (2005), "Strategic extremism: Why republicans and democrats divide on religious values." *The Quarterly journal of economics*, 120, 1283–1330.

Golab-Andrzejak, Edyta (2023), "The impact of generative AI and chatgpt on creating digital advertising campaigns." *Cybernetics and Systems*, 1–15.

Golub, Benjamin and Matthew O Jackson (2010), "Naive learning in social networks and the wisdom of crowds." *American Economic Journal: Microeconomics*, 2, 112–149.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer (2019), "Fake news on twitter during the 2016 US presidential election." *Science*, 363, 374–378.

Guess, Andrew M, Brendan Nyhan, and Jason Reifler (2020), "Exposure to untrustworthy websites in the 2016 US election." *Nature human behaviour*, 4, 472–480.

Gupta, Aastha (2023), "Deceptive advertising, regulation and naive consumers." *International Journal of Industrial Organization*, 91, 103026.

Guriev, Sergei and Elias Papaioannou (2022), "The political economy of populism." *Journal of Economic Literature*, 60, 753–832.

Hobbs, Renee, Igor Kanižaj, and Luis Pereira (2019), "Digital literacy and propaganda." *Media Studies*, 10, 1–7.

Hotelling, Habold (1929), "Stability in competition." *The economic journal*, 39, 41–57.

Hristakieva, Kristina, Stefano Cresci, Giovanni Da San Martino, Mauro Conti, and Preslav Nakov (2022), "The spread of propaganda by coordinated communities on social media." In *Proceedings of the 14th ACM Web Science Conference 2022*, 191–201.

Hruska, Jan and Petra Maresova (2020), "Use of social media platforms among adults in the united states—behavior on social media." *Societies*, 10, 27.

Hudders, Liselot, Eva A Van Reijmersdal, and Karolien Poels (2019), "Digital advertising and consumer empowerment." *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 13.

Huerta, EA, Ben Blaiszik, L Catherine Brinson, Kristofer E Bouchard, Daniel Diaz, Caterina Doglioni, Javier M Duarte, Murali Emani, Ian Foster, Geoffrey Fox, et al. (2023), "Fair for ai: An interdisciplinary and international community building perspective." *Scientific data*, 10, 487.

Kemp, Simon (2020), "Digital 2020: Global digital overview."

Langguth, Johannes, Konstantin Pogorelov, Stefan Brenner, Petra Filkuková, and Daniel Thilo Schroeder (2021), "Don't trust your eyes: image manipulation in the age of deepfakes." *Frontiers in Communication*, 6, 632317.

Lazer, David MJ, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. (2018), "The science of fake news." *Science*, 359, 1094–1096.

Levy, Ro'ee (2021), "Social media, news consumption, and polarization: Evidence from a field experiment." *American economic review*, 111, 831–870.

Lewandowsky, Stephan, Ullrich KH Ecker, and John Cook (2017), "Beyond misinformation: Understanding and coping with the "post-truth" era." *Journal of applied research in memory and cognition*, 6, 353–369.

Lindbeck, Assar and Jörgen W Weibull (1993), "A model of political equilibrium in a representative democracy." *Journal of public Economics*, 51, 195–209.

Lu, Zhuoran, Patrick Li, Weilong Wang, and Ming Yin (2022), "The effects of ai-based credibility indicators on the detection and spread of misinformation under social influence." *Proceedings of the ACM on Human-Computer Interaction*, 6, 1–27.

Mas-Colell, Andreu, Michael Dennis Whinston, Jerry R Green, et al. (1995), *Microeconomic theory*, volume 1. Oxford university press New York.

McCarty, Nolan (2007), "The policy effects of political polarization." *The transformation of American politics: Activist government and the rise of conservatism*, 223–55.

Meurer, Michael and Dale O. Stahl (1994), "Informative advertising and product match." *International Journal of Industrial Organization*, 12, 1–19.

Milgrom, Paul and Chris Shannon (1994), "Monotone comparative statics." *Econometrica: Journal of the Econometric Society*, 157–180.

Molina, Carlos Molina (2024), *Essays on Political Economy*. Ph.D. thesis, Massachusetts Institute of Technology.

Mostagir, Mohamed, Asuman Ozdaglar, and James Siderius (2022), "When is society susceptible to manipulation?" *Management Science*, 68, 7153–7175.

Mostagir, Mohamed and James Siderius (2022a), "Learning in a post-truth world." *Management Science*, 68, 2860–2868.

Mostagir, Mohamed and James Siderius (2022b), "Naive and bayesian learning with misinformation policies." Technical report, Massachusetts Institute of Technology.

Mostagir, Mohamed and James Siderius (2023), "Social inequality and the spread of misinformation." *Management Science*, 69, 968–995.

Mostagir, Mohamed and James Siderius (2024), "When should platforms break echo chambers?" Technical report, Massachusetts Institute of Technology.

Nayak, Sameera S, Timothy Fraser, Costas Panagopoulos, Daniel P Aldrich, and Daniel Kim (2021), "Is divisive politics making americans sick? associations of perceived partisan polarization with physical and mental health outcomes among adults in the united states." *Social Science & Medicine*, 284, 113976.

Nyhan, Brendan (2020), "Facts and myths about misperceptions." *Journal of Economic Perspectives*, 34, 220–236.

Ott, William and James Yorke (2005), "Prevalence." *Bulletin of the American Mathematical Society*, 42, 263–290.

O'Callaghan, Derek, Derek Greene, Maura Conway, Joe Carthy, and Pádraig Cunningham (2015), "Down the (white) rabbit hole: The extreme right and online recommender systems." *Social Science Computer Review*, 33, 459–478.

Pennycook, Gordon and David G Rand (2019), "Fighting misinformation on social media using crowdsourced judgments of news source quality." *Proceedings of the National Academy of Sciences*, 116, 2521–2526.

Pew Research Center (2014), "Political polarization in the american public." URL https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/. Accessed: 2024-12-18.

Pew Research Center (2020), "Americans who mainly get their news on social media are less engaged, less knowledgeable." URL https://www.pewresearch.org/journalism/2020/07/30/americans-who-mainly-get-their-news-on-social-media-are-less-engaged-less-knowledgeable/. Accessed: 2025-01-18.

Pew Research Center (2024), "Social media and news fact sheet." URL https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/. Accessed: 2025-01-18.

Piccolo, Salvatore, Piero Tedeschi, and Giovanni Ursino (2018), "Deceptive advertising with rational buyers." *Management Science*, 64, 1291–1310.

Quinn, Kevin M, Andrew D Martin, and Andrew B Whitford (1999), "Voter choice in multi-party democracies: a test of competing theories and models." *American Journal of Political Science*, 1231–1247.

Rathje, Steve, Claire Robertson, William J Brady, and Jay J Van Bavel (2024), "People think that social media platforms do (but should not) amplify divisive content." *Perspectives on Psychological Science*, 19, 781–795.

Rochlin, Nick (2017), "Fake news: belief in post-truth." *Library hi tech*, 35, 386–392.

Ross Arguedas, Amy, Craig Robertson, Richard Fletcher, and Rasmus Nielsen (2022), "Echo chambers, filter bubbles, and polarisation: A literature review." *Reuters Institute for the Study of Journalism*.

Simchon, Almog, Matthew Edwards, and Stephan Lewandowsky (2024), "The persuasive effects of political microtargeting in the age of generative artificial intelligence." *PNAS nexus*, 3, pgae035.

Statista Research Department (2024a), "Digital media consumption in the u.s. in 2024." URL https://www.statista.com/topics/1536/media-use/. Accessed: 2025-01-18.

Statista Research Department (2024b), "Political advertising in the u.s. - statistics and facts." URL https://www.statista.com/topics/4942/political-advertising-in-the-us/. Accessed: 2025-01-18.

Suarez-Lledo, Victor and Javier Alvarez-Galvez (2021), "Prevalence of health misinformation on social media: systematic review." *Journal of medical Internet research*, 23, e17187.

Sunstein, Cass (2018), *# Republic: Divided democracy in the age of social media*. Princeton university press.

Topkis, Donald M (1998), *Supermodularity and complementarity*. Princeton university press.

Törnberg, Petter (2018), "Echo chambers and viral misinformation: Modeling fake news as complex contagion." *PLoS one*, 13, e0203958.

Tucker, Joshua A, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan (2018), "Social media, political polarization, and political disinformation: A review of the scientific literature." *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*.

Veerasamy, Namosha and Heloise Pieterse (2022), "Rising above misinformation and deepfakes." In *International Conference on Cyber Warfare and Security*, volume 17, 340–348.

Zeng, Jinghan (2020), "Artificial intelligence and china's authoritarian governance." *International Affairs*, 96, 1441–1459.