THE DESIRE FOR RESPECT CAN DRIVE POLARIZATION

Kjell Arne Brekke and Karine Nyborg

Abstract:

Evangelical Christians are often viewed less favorably by Democrats than by Republicans, while the reverse hold, e.g., for Muslims. We show that if individual characteristics of positive, possibly small shares of the population are less respected by one side of the ideological spectrum, almost everyone's ideological views become polarized in equilibrium. Equilibrium views are more extreme if encounters across peer groups are infrequent, and if individuals are strongly reluctant to adopt views threatening their self-respect. Polarization limits affected individuals' exposure to disrespectful views, thus improving their welfare. Unlike previous contributions, homophily is not assumed but appears endogenously.

Keywords: Image preferences; affective and ideological polarization; reluctant social learning.

JEL Codes: D72; D91.

Addresses: Department of Economics, University of Oslo, P.O. Box 1095 Blindern, NO-0317 Oslo, Norway (both authors). Email: karine.nyborg@econ.uio.no (corresponding author), k.a.brekke@econ.uio.no.

Acknowledgements: We are grateful for comments to earlier versions by a substantial number of colleagues and conference and seminar participants, in particular Ingela Alger, Felix Dwinger, Sahar Fard, Marijn Keijzer, Bård Harstad, Frikk Nesje, Steinar Holden, Karine van der Straeten, and Morten Støstad. This work did not receive external funding. The authors have no relevant financial or other conflicts of interest.

1. Introduction

In a recent survey, 81% of US Democrats but only 40% of Republicans reported to view same-sex relationships as morally acceptable (Gallup 2024). Perhaps not surprisingly, then, LGBTQ voters have a strong tendency to vote Democratic (Michelson and Schmitt 2020). Similarly, atheists and Muslims are viewed unfavorably more often by Republicans than by Democrats, and lean heavily Democratic (Pew Research 2023); evangelical Christians, on the other hand, are viewed more unfavorably by Democrats (Pew 2023), and tend to support the Republican party (Pew 2024). For Jews, however, the pattern differs: while they are unlikely to be viewed negatively by either Republicans or Democrats, they lean strongly Democratic (Pew 2023).

Here, we present a formal model exploring the relationships between ideological views, social interaction, and a preference to be respected. We find that if a strictly positive but possibly small share of the population feels less respected by one side of the ideological spectrum than the other, almost everyone's views become polarized in the steady state. Those with characteristics more respected by one side end up supporting that side. The views of those equally respected by both sides also become polarized, but in less predictable ways: which side they turn to will depend on the composition of their peer group.

Feeling esteemed and respected by others as well as oneself is key to human well-being (Crocker and Wolfe 2001; Pyszczynski et al. 2004; Lieberman 2013). While the desire for social esteem and self-respect can be important motivators for prosocial behavior (Brekke et al. 2003; Benabou and Tirole 2006a; Nyborg et al. 2006; Brekke and Nyborg 2008, 2010; Nyborg 2011; Benabou et al. 2018, Falk 2021), the present paper points out a very different implication: under plausible conditions, the preference to be respected by oneself and others can drive society towards ideological and affective polarization. Ideological polarization occurs when ideological views cluster around the extremes rather than the middle ground (see Brady and Han 2006; Lee 2015).[1] Affective polarization means that proponents of each side have low esteem for their opponents and avoid social contact with them (Iyengar et al. 2019). While ideological and affective polarization do not necessarily go hand in hand, they often do.

The idea that people care about their image is by now widely explored in the economics literature (e.g., Akerlof and Kranton, 2000; Brekke et al. 2003; Santos-Pinto and Sobel 2005; Benabou and

[1] Examples include the current divide between US Republicans and Democrats (Lee, 2015; Alesina et al. 2020); the debate on slavery in the 19th century (Brady and Han 2006; Hetherington 2009); the political situation in Germany between the first and second World Wars (Caprettini et al., 2024); Brexit (Hobolt et al. 2021); Norway's 1972 referendum on whether to join the EU predecessor EEC (Holst 1975); and conflicts on LGBTQ rights (Hadler and Symons 2018; Castle 2019).

Tirole 2006, 2016; Corneo and Jeanne 2009; Shayo 2009; Ellingsen and Johannesson 2011; Bursztyn and Jensen 2017; Benabou et al. 2018; Bonomi et al. 2021). Nevertheless, the idea that different ideological views may be associated with respect for different individual characteristics has not, to our knowledge, previously been explored in this literature.

Assume, for example, that intellectuals are viewed less favorably by conservatives than by liberals, while the reverse is true for the wealthy. If so, intellectuals may be more comfortable among liberal peers, while the wealthy may prefer socializing with conservatives. Furthermore, to protect their self-respect, intellectuals may be reluctant to learn from their conservative peers; such biased learning will influence the intellecutal's own future views, and through social learning also her peers, even those who do not themselves feel less respected by one side.

Below, we formalize how this can drive long-term social dynamics helping individuals feel respected, causing polarization as an externality. Our assumptions are fairly general and, we believe, plausible.

We place no restrictions on individuals' initial views or peer group affiliations. In the short run, peer groups and ideological views are considered fixed; one cannot simply choose an ideological conviction that would be pleasant for oneself. Short-run utility depends on self-respect and social respect; in an extension we add costly effort to increase one's respect, but since this is unimportant for the dynamics, we start with the simplest specification.

The respect given to a person is assumed to depend on that person's exogenous characteristics and the ideological position of the person *making the judgement*. That is, j's respect for i depends on i's characteristics and j's ideological view – not, for example, the difference between i's and j's views. This differ from much of the previous literature on polarization and social cleavages, where assumptions of homophily – i.e., that people prefer to interact with others who are similar to themselves – often play a crucial role. For example, in Desmet and Wacziarg (2020), Axelrod et al. (2023), and Desmet et al. (2025), individuals are assumed to be drawn towards those who have similar views as themselves; Törnberg et al. (2021) assume a preference to interact with others with a similar social identification. Different but relatedly, Törnberg (2022) assume that individuals are more influenced by socially similar others, while Bonomi et al. (2021) assume that voters, after choosing their social identity, slant their beliefs towards the group they identify with. In the model below, similarity as such determines neither one's preferred peers nor by whom one is influenced; people are simply assumed to prefer peers who respect them more, while being reluctant to adopt views threatening their own self-image. Segregated, like-minded peer groups then arise endogenously. Adding assumptions of homophily like in, e.g., Axelrod et al. (2023), would reinforce the polarizing force in our model but are not required to derive it.

In our model, two dynamic social mechanisms interact in the long run: migration between peer groups and social learning of ideological views. We assume that every now and then, people reconsider their peer group affiliation, preferring peers respecting them more. Furthermore, ideological views are gradually learnt from others, predominantly one's peers (Algan et al., 2023); however, in line with the literature on biased learning, motivated beliefs, and psychological reactance (Brehm 1966; Babcock and Loewenstein 1997; Hart et al. 2009; Deffains et al. 2016; Rosenberg and Siegel 2018), such learning is taken to be reluctant (Brekke et al. 2010): one is less likely to be swayed towards others' views if doing so would decrease one's utility.

In this social learning process, individuals make judgement errors, since others' ideological positions cannot be observed with precision. Due to reluctance, however, errors implicitly benefiting the individual's self-respect will be given disproportionally large weight. This combination of uncertainty and reluctance allows ideological views to move beyond their initial range, potentially all the way to the extremes, even if all views initially happened to be quite moderate.

Unlike Desmet and Wacziarg (2020), Bonomi et al. (2021), and Desmet et al. (2025), we restrict ourselves to explore the case of one-dimensional ideological disagreement. We assume that ideological views can be sorted along a continuous scale ranging from 0 (left) to 1 (right), which could represent, for example, the range from extremely liberal to extremely conservative, support for democracy versus support for totalitarianism, or support for equal rights versus support for oppression of some group.[2] A crucial assumption is the following. If peers' ideological views approach one extreme, say 1, then individuals with some sets of exogenous characteristics, henceforth called the *L* type ("respected by the Left"), will experience *lower* social respect – whereas social respect *increases* for individuals with other sets of characteristics, called the *R* type ("respected by the Right"). Individuals whose social respect is unaffected by changes in peers' ideological views are termed the *O* type. Similarly, if one's *own* ideological view moves towards 1, self-respect is reduced for *L*'s, improved for *R*'s, and left unaffected for O's.[3]

We find that over time, *L* and *R* types gradually self-select into different peer groups, while *O* types have no reason for social migration. For *O* types, social learning of ideological views is unbiased; *L* types are biased towards the left, *R* types towards the right.

[2] Of course, this scale needs not necessarily correspond to the traditional political left – right spectrum, although we use the terms 'left' and 'right' for simplicity.
[3] That is, we disregard individuals whose received respect is non-monotonic in the observer's ideological view. The assumption that both social respect and self-respect vary in the same direction with the observer's view corresponds to the consistency criterion of Corneo and Jeanne (2009).

In the steady state, no peer group contains both *L* and *R* types. Everyone in a peer group with *L* types, including any *O*'s in that group, agrees on a leftist view, while everyone in a peer group with *R* types agrees on a view to the right. Only peer groups consisting exclusively of *O* types, if such groups exist, may hold intermediate equilibrium views. Except such groups, thus, the entire population is ideologically polarized.

Although it may not be too surprising that the *L* and *R* types are gradually drawn towards the ideological positions implicitly benefiting them, it is important to note that even the *O* types – who may well constitute the majority – are just as ideologically polarized in equilibrium as everyone else (except those in exclusively *O*-type groups, if such groups exist). Since everyone, including *O*'s, are influenced by their peers, *O*'s in groups including *L* peers become gradually more leftist, while *O*'s in groups including *R* peers move gradually to the right. A larger share of *O*'s will slow down this process but leave the equilibrium unaffected.

This ideological polarization is more extreme in equilibrium the more social learning occurs within rather than across peer groups. In the extreme case where learning takes place exclusively within peer groups, everyone in a group with *L*'s hold the most possible extreme leftist view in equilibrium, while everyone in a group with *R*'s hold the most possible extreme rightward view.

There is also affective polarization in the steady state. Recall that *L*'s are defined as those highly respected by the left, while *R*'s are highly respected by the right. In equilibrium, everyone in a group with *L*'s have become leftists, thus having low respect for *R*'s, all of whom are their opponents; conversely, all *R*'s now lean to the right, thus having low respect for their *L* opponents. Furthermore, there is no social interaction between opponents. Nevertheless, the presence of the *O*'s limits affective polarization in our model in the following sense: Disregarding groups with only *O*'s, if they exist, all *O*'s have low respect for their opponents, exactly like the *L*'s and *R*'s – but their opponents do not have particularly low respect for them.

In our model, we abstract from possible macro level effects of increased polarization such as political unrest and instability, less general trust, less efficient intergroup collaboration and cooperation on public goods, more bullying and violence, and so on. Having that in mind, we find that polarization enhances welfare – by limiting *L*'s and *R*'s exposure to views undermining their self-respect and social respect. When costly effort to gain respect is allowed, we show that optimal effort is minimized for everyone when polarization is maximally extreme.

The main novelty of our approach is to show that under plausible assumptions, potentially extreme ideological polarization can arise simply because people strive to be respected by themselves and others; homophily arise as a result but need not be assumed. In addition, our mechanism allowing

views to become more extreme than their initial range due to a combination of uncertainty and reluctance is, to our knowledge, new.

Brown et al. (2022) show that polarization and segregation may result when individuals compromise between their own and peers' attitudes; their analysis, however, is based on the idea that attitudes are represented by statistical distributions rather than single ideal points. Shayo (2009) and Sambanis and Shayo (2013) discuss endogenous group choice based on social identity, but like Akerlof and Kranton (2000), they do not explore the dynamic implications of social learning within groups. Schelling's (1978) segregation model does not involve changing attitudes, thus predicting segregation but not polarization.

Bonomi et al. (2021) analyze the case of two-dimensional disagreement, finding that social identification causes increased but not extreme political disagreement by distorting individuals' factual beliefs. The latter represents an element of limited rationality not required in our analysis: social learning in our model is not concerned with facts but with normative opinions, for which there are no objectively correct answers. Note, however, that if ideological positions are based on factual beliefs (e.g., views on climate policies being based on climate denial/the belief that climate change is man-made), then replacing the social learning of ideological beliefs in our model by reluctant social learning of factual beliefs could explain the observation that people's factual beliefs are correlated with their political views (Alesina et al., 2020).

So what? In our model, equilibrium polarization is decreasing in the probability of meeting someone from a different peer group but increasing in learning reluctance. Thus, exogenous shocks and policies affecting either of these are predicted to affect polarization. In the model, meeting someone simply means being exposed to that person's views, which can in practice take place in person or through media exposure. Thus, one example of a relevant exogenous shock is the introduction of the internet, apparently causing reduced consumption of cross-cutting media like local newspapers and national TV programmes but increased consumption of more fragmented and like-minded media (Mutz 2024). Possible policies counteracting the polarizing effect of this might include, for example, subsidizing bipartisan mass media, regulating the use of algorithms in social media, prohibiting or limiting options for personalized ideological advertising, or incentivizing people to familiarize themselves with opponent media (Akbiyik et al., 2024). Also, cross-cutting meetings may be stimulated through labor market regulations limiting segregation in the workplace, such as job security regulations prohibiting employers from hiring and/or firing employees based on the employee's opinions or job irrelevant personal characteristics; through education policies influencing which groups tend to attend the same schools and colleges; through cultural policies supporting high-quality literature, drama and screenplay which would likely be enjoyed by mixed audiences (assuming that artists' views are implicitly expressed through their art); as well as any policies stimulating diverse groups' social

encounters and joint participation in public debate (Benabou et al. 2018). While the degree of reluctance may be less straightforward to influence, schools may train students' curiosity and tolerance; academic training may help students judge others' views less self-defensively; while, conversely, politicians acting aggressively towards their opponents may possibly raise reluctance by making their supporters less open to the views of those opponents.

Finally, the fundamental driver of the polarization process in our model is the premise that lack of respect for some groups is associated with ideology. If both sides respected everyone equally, the polarizing force would vanish. Our model results are of a 'bang-bang' nature, but we conjecture that in a more complex but also more nuanced model, stronger associations between ideology and respect for specific groups would be predicted to increase polarization.

Below, we present our formal model, starting with the short run. We then turn to the dynamic mechanisms of social learning and migration, respectively, before merging all parts into an integrated dynamic model. For simplicity, we initially assume that all social learning occurs within peer groups, yielding a prediction of extreme ideological polarization, but subsequently modify this, showing what then determines the degree of polarization. Finally, we extend the model to allow the option of exerting costly effort to gain respect, before concluding.

2. Image preferences

We begin with the static part of our model. In the short term, individuals consider ideological views and social group affiliations fixed.

Let each individual i's ideological conviction be indexed by $q_i \in [0,1]$. Each individual belongs to one of a set of non-overlapping social peer groups, where $i \in G$ denotes that individual i belongs to peer group G. Let q_G denote the average ideological view q_i for members of peer group G.

Each individual i has an exogenously fixed vector of characteristics θ_i. This vector could for example include social class, education and other human capital, financial wealth, ethnicity, sexual orientation, personality traits, religion, and so on. Each individual's specific combination of characteristics will not matter below; the crucial assumption is that every feasible set of characteristics θ_i can be classified as belonging to one (and only one) of three sets or types, L (respected by the left), R (respected by the right), and O (equally respected by all): for every i, either $\theta_i \in L$, $\theta_i \in R$, or $\theta_i \in O$.

Individuals care about their self-respect (or self-image), I_i, as well as their social respect from peers (social image), S_i. Both i's self-respect and social respect depend on i's exogenous characteristics θ_i (where either $\theta_i \in L$, $\theta_i \in R$, or $\theta_i \in O$); self-respect also depends on i's *own* ideological view q_i, while social respect depends instead on i's peers' average ideological view q_G (where G is i's peer group). Individual utility U_i is assumed to be linearly separable for simplicity:

(1) $U_i = I_i(\theta_i, q_i) + S_i(\theta_i, q_G)$.

Note that it is the evaluator's ideological view that enters the image functions, not the views of the person being evaluated; the object of the evaluation is the evaluated individual's characteristics, not her ideological position.

As long as we abstract from the option of exerting effort, individuals have no choices to make in the short run; they must simply accept their own and others' judgements. If, for example, a person is gay but holds a very conservative ideological view, he may regard himself less favorably than he would if his views had been more liberal.

The following formalizes our assumption that a change in ideological views affect image differently for different types.

Assumption 1:

(i) I_i is decreasing in q_i for $\theta_i \in L$, increasing in q_i for $\theta_i \in R$, and is independent of q_i for $\theta_i \in O$.

(ii) S_i is decreasing in q_G for $\theta_i \in L$, increasing in q_G for $\theta_i \in R$, and is independent of q_G for $\theta_i \in O$, where G is i's peer group.

It follows that utility U_i is decreasing in q_i and q_G for $\theta_i \in L$, increasing in q_i and q_G for $\theta_i \in R$, and is independent of q_i and q_G for $\theta_i \in O$.

3. Social learning of ideological views

Let us now turn to the dynamics. We first consider how ideological views are learnt from others over time, keeping peer groups fixed for the moment. Social migration is added to the picture in the next section.

Ideological views represented by q_i are convictions that cannot simply be chosen. However, since such value judgements are inherently subjective and cannot be deduced from facts and logic alone, it seems reasonable to assume that they are at least to some extent instilled by parents, school, friends and role models – for example through observation of others' behaviors and statements, shared deliberation and reflection, or explicit ethical debate.

Let us begin with discrete time, using a superscript t to denote the time period; when moving to continuous time below, we omit this. For simplicity, we also assume for the moment that all social learning takes place within peer groups; this assumption will be relaxed later on.

When a person i meets another person j, i cannot know the other's view q_j^t perfectly. Consider first the case of unbiased social learning. Assume that each period, i meets with a random individual j in her social group (a new random draw for each period) and adjusts her ideological view q_i^t a fraction $\delta > 0$ in the direction of what i perceives to be j's view, $\tilde{q}_{ji}^t$. Let this perception be established with some noise, although unbiased: $E\tilde{q}_{ji}^t = q_j^t$. To avoid truncating the distribution of $\tilde{q}_{ji}^t$, we assume that the distribution is symmetric and has support in [0,1].[4] With unbiased learning, the change in i´s view would thus be

$$(2) \qquad \Delta q_i^t = q_i^{t+\Delta t} - q_i^t = \delta\left(\tilde{q}_{ji}^t - q_i^t\right)\Delta t$$

where $i, j \in G_i^t$ (where G_i^t is i's peer group in period t). If we now move to continuous time by letting the time step approach zero, i´s view is pulled towards an average of all $\tilde{q}_{ji}^t$ observed during any fixed time interval, converging to the group average q_G^t. Thus, we get the continuous process

$$(3) \qquad \dot{q}_i^t = \delta\left(q_G^t - q_i^t\right).$$

That is, each member of the group gradually adjusts their view toward the group average, eventually leading to $q_i^t \approx q_G^t$ for all group members. Averaging over all i, this means that with unbiased social learning within a fixed group, the average ideological view of the group stays unchanged:

$$(4) \qquad \dot{q}_G^t = \delta(q_G^t - q_G^t) = 0.$$

Note that since all group members' views converge towards the initial group average q_G^0, in-group variation is gradually reduced. Hence, if the initial average view in group G, q_G^0, is different for two peer groups, ideological disagreement is gradually reduced within each group but not between groups (given unbiased social learning and no migration).

However, when uncertain observation of others' views is combined with *reluctant social learning*, groups' average ideological view can change beyond their initial values over time, moving potentially all the way to the extremes. The idea here is that although individuals gradually learn their ideological view from others, they have a slight reluctance to pay attention to views that would be to their disadvantage. Due to this reluctance, errors in the perception of others' ideological views do not cancel out over time. Let us again begin with discrete time.

[4] This implies that the distribution's variance must depend on $\tilde{q}_{ji}^t$, approaching 0 as $\tilde{q}_{ji}^t$ approaches 0 or 1. We return to this in the discussion of migration below.

Definition (unbiased and reluctant social learners): Let $1 > \delta > 0$ and $1 > r > 0$. Assume that each period t, i meets with a random individual j in her social group (a new random draw for each period). An *unbiased social learner* i then adjusts her ideological view q_i^t a fraction δ in the direction of what i perceives to be j's view, $\tilde{q}_{ji}^t$. A *reluctant social learner* i adjusts her ideological view q_i^t a fraction δ in the direction of $\tilde{q}_{ji}^t$ when doing so weakly increases U_i^t, but adjusts her ideological view q_i^t only a fraction $\delta(1-r)$ in the direction of $\tilde{q}_{ji}^t$ otherwise.

Below, we assume that social learning is reluctant.

Consider, now, the situation where individual i meets j. If $\theta_i \in L$, i is reluctant to adopt an *increase* in q_i^t, but if instead $\theta_i \in R$, she is reluctant to adopt a *decrease* in q_i^t. Hence, if $\theta_i \in R$, we have

$$\Delta q_i^t(\theta_i \in R) = \begin{cases} \delta(\tilde{q}_{ji}^t - q_i^t)\Delta t & if\ \tilde{q}_{ji}^t \geq q_i^t \\ \delta(1-r)(\tilde{q}_{ji}^t - q_i^t)\Delta t & if\ \tilde{q}_{ji}^t < q_i^t \end{cases}$$

$$= \delta(\tilde{q}_{ji}^t - q_i^t)\Delta t + \begin{cases} 0 & if\ \tilde{q}_{ji}^t \geq q_i^t \\ -\delta r(\tilde{q}_{ji}^t - q_i^t)\Delta t & if\ \tilde{q}_{ji}^t < q_i^t. \end{cases}$$

Similarly, for $\theta_i \in L$:

$$\Delta q_i^t(\theta_i \in L) = \delta(\tilde{q}_{ji}^t - q_i^t)\Delta t + \begin{cases} 0 & if\ \tilde{q}_{ji}^t < q_i^t \\ -\delta r(\tilde{q}_{ji}^t - q_i^t)\Delta t & if\ \tilde{q}_{ji}^t \geq q_i^t. \end{cases}$$

To simplify notation, let $(\tilde{q}_j^t - q_i^t)^-$ denote the negative part of $(\tilde{q}_j^t - q_i^t)$, and let $(\tilde{q}_{ji}^t - q_i^t)^+$ denote the positive part.[5] A more concise way to write the change over time in q_i^t, for either type is then

$$(5) \qquad \Delta q_i^t = \delta(\tilde{q}_{ji}^t - q_i^t)\Delta t \begin{cases} +r\delta(\tilde{q}_{ji}^t - q_i^t)^-\Delta t & if\ i \in R \\ -r\delta(\tilde{q}_{ji}^t - q_i^t)^+\Delta t & if\ i \in L. \end{cases}$$

We can now introduce a measure of expected learning biases. Let $B_{ij}^{+t} = E(\tilde{q}_{ji}^t - q_i^t)^+ > 0$ be the expected positive part. Similarly, let $B_{ij}^{-t} > 0$ be the expected negative part. Furthermore, to indicate that, for example, $\theta_i \in R$ and $\theta_j \in L$ and hence the expectation must be taken over $\theta_i \in R$ and $\theta_j \in L$, let us write $B_{RL}^{+t} = E[(\tilde{q}_{ji}^t - q_i^t)^+ | \theta_i \in R \text{ and } \theta_j \in L]$. These variables are proportional to the expected size of the learning biases in the various cases. Since an individual $\theta_i \in R$ is only reluctant to adopt the

[5] That is, $(\tilde{q}_{ji}^t - q_i^t)^- = 0$ if $\tilde{q}_{ji}^t > q_i^t$ and $-(\tilde{q}_{ji}^t - q_i^t)$ if $\tilde{q}_{ji}^t < q_i^t$. Similarly, $(\tilde{q}_{ji}^t - q_i^t)^+ = (\tilde{q}_{ji}^t - q_i^t)$ if $\tilde{q}_{ji}^t > q_i^t$ and 0 if $\tilde{q}_{ji}^t < q_i^t$. Note that both the negative and the positive parts are positively signed.

perceived view of j if $\tilde{q}_{ji}^t < q_i^t$, only B_{RR}^{-t} and B_{RL}^{-t} matter for $\theta_i \in R$; similarly, only B_{LL}^{+t} and B_{LR}^{+t} matter for $i \in L$.

For simplicity, we first consider the case where the set of type *O* is empty. Now, let us again move to continuous time, and as we will always stay at one point in time we drop the superscript t. Let s_G be the share of R types in group G. The probability that an R type meets another R type is then s_G. We noted above that without reluctance $\dot{q}_G = 0$, hence we can now restrict ourselves to consider the effects of reluctance, which is different for *L* and *R* types. The dynamic effect of reluctance on *R* types' views is then

$$\dot{q}_R^r = s_G r\delta B_{RR}^- + (1 - s_G) r\delta B_{RL}^-, \tag{6}$$

where $\dot{q}_R^r$ is the change in the average ideological view due to reluctance for $\theta_i \in R$. Similarly, for $\theta_i \in L$ we have

$$\dot{q}_L^r = -s_G r\delta B_{LR}^+ - (1 - s_G) r\delta B_{LL}^+. \tag{7}$$

Since the dynamic for the group's average view $\dot{q}_G$ is only due to reluctance, it is the weighted average of the two, which gives

$$\dot{q}_G = r\delta[s_G^2 B_{RR}^- + (1 - s_G)s_G(B_{RL}^- - B_{LR}^+) - (1 - s_G)^2 B_{LL}^+] \equiv r\delta\Pi_G. \tag{8}$$

Hence, $r\delta\Pi_G$, as defined by eq. (8), is a measure of the total effect of reluctance in social group G.

Assumption 2 below specifies our assumptions concerning the probability distribution of $\tilde{q}_{ji}$.

Assumption 2. Let the probability distribution for $\tilde{q}_{ji}$ be binary, symmetric, unbiased with $E(\tilde{q}_{ji}) = q_j$, and with support on [0,1]. Specifically, let $\tilde{q}_{ji} = q_j \pm \phi d_j$ with equal probability, where $d_j = min(q_j, 1 - q_j)$ is the distance between q_j and the border of [0,1], and $0 < \phi \leq 1$.

While a binary distribution simplifies the problem, we also discuss the case of a uniform distribution in Appendix 1.

We must now allow the set of *O* types to be non-empty. Note that while *O* players are themselves effectively unbiased in their learning (since their self-respect is unaffected by their own ideological position), types *L* and *R* behave reluctantly when meeting an *O* player if they perceive the *O* player's view as pulling in the "wrong" direction. The resulting bias measure will be denoted B_{RO}^- and B_{LO}^+, defined in a similar fashion as above.

Let $(1-\alpha_G)$ denote the share of *O*'s in group *G*, and let us now interpret s_G as the share of *R's* among the non-*O*'s in the group (i.e., among the remaining share α_G: for example, if $s_G = 0.5$ there are equally many *R*'s and *L*'s in group *G*).

We can now state our Proposition 1, which will be an important building block towards our main results: If there is no migration, and the strength of individuals' learning reluctance is limited, then if there are more *R*'s than *L*'s in the peer group, everyone in the group holds the extreme view $q_i = 1$ in the only asymptotically stable steady state; conversely, if there are more *L*'s than *R*'s in the peer group, everyone in the group holds the opposite extreme view $q_i = 0$ in the only asymptotically stable steady state.

Proposition 1. Assume that the composition and size of social groups are fixed, and that learning is reluctant. Then, in a steady state, all $i \in R$ in a given peer group hold the same view $q_i = q_R$, all $i \in L$ in the group hold the same view $q_i = q_L$, and all $i \in O$ in the group hold the same view $q_i = q_O$. Moreover, for $r < \frac{2}{5}$, and given Assumption 1 - 2,

I. For all values of s_G, state *a*: $q_L = q_R = 0$ and state *b*: $q_L = q_R = 1$ are stable states for which $\dot{q}_L = \dot{q}_R = \dot{q}_G = 0$.
II. If $s_G < \frac{1}{2}$, only state *a* is asymptotically stable, while for $s_G > \frac{1}{2}$, only state *b* is asymptotically stable.
III. For $s_G = \frac{1}{2}$, all states with $q_R - q_L < \phi \min(d_L, d_R)$ are stable states, but none of them are asymptotically stable.
IV. There are no further stable states.

Proof: See Appendix 1.

To see the main intuition of the proof (which is itself rather tedious), note that when people learn from each other within a fixed group, this reduces the heterogeneity in q_i within each type, and in the case of limited reluctance also between types. Still, however, reluctance pulls *R* types towards higher q_i and *L* types towards lower q_i. The relative strength of these forces depends on whether there are more

L than *R* types. This drives the group average q_G towards zero if the majority of non-O's consists of L types but towards one if the majority of non-O's consists of R types.[6]

4. Choosing one's peers

Let us now allow migration between peer groups. In contrast to ideological view updating, which is basically an inference, changing one's peer group is a choice. Nevertheless, inspired by evolutionary game theory (Weibull 1995), we assume that individuals revise their peer group affiliation only every now and then, and that they do so myopically, not taking into account that migration might affect their own future ideological view. [7]

Assumption 3: When reconsidering at time t which peer group G to be part of, i prefers the group that would give the highest utility U_i^t.

O types have no incentives to migrate, as their utility is independent of q_G. As migration thus only involves *L* and *R* types, we first consider the case with only *L* and *R* types, then demonstrating that the main results are still valid when *O* types are present.

Assume now that there are only two equally large peer groups, A and B, and that the population consists of equally many L and R types. We relax these assumptions in Appendix 3, but for now they simplify the calculations below considerably, as we will only need one state variable: knowing the number of L types in social group *A*, the number of R and L in each peer group follows. Also, let individuals disregard the potential effect of their own migration on q_G in either peer group.

In the current framework, the only reason why individuals care about social group affiliation is their preference for social respect from their peers. When revising which peer group they want to be part of, L types thus prefer the social group with lower q_G, while R types prefer the neighborhood with higher

[6] Appendix 1 also analyzes the case with a uniform distribution (with the same support) but with no O-types. The main difference is that in this case, there is an asymptotically stable state with an average view $q \approx \frac{1}{2}$ in an interval around $s_G = \frac{1}{2}$ (Theorem A1-2.) When such a stable state exists, we can provide numerical bounds for the width of this interval around $s_G = \frac{1}{2}$, demonstrating that the interval is small: for example, if $r < 0.1$, the relevant interval is contained in $s_G \in (\frac{1}{2} - 10^{-5}, \frac{1}{2} + 10^{-5})$. Outside of this interval, result II in Proposition 1 holds.

[7] We do not believe that rational foresight would change our main results, except that a coordination problem may arise, since individuals would not know in advance which groups would end up having high and low q_G. When social group membership is revised myopically and only occasionally, group composition and thus q_G are stable in the short run.

q_G. For example, a single mother may prefer to be surrounded by liberal peers, while a wealthy person may prefer conservative peers.

The share of each type within a peer group can now vary over time. Let us again move to continuous time by shortening period length towards zero. Let s_G denote the share of R types in group $G \in \{A, B\}$ at a given moment in time (still assuming no *O* types). Note that if $q_A > q_B$, R types prefer A, so only the share of R types who are in B, $1 - s_A$, have incentives to move. Denoting by $\rho > 0$ the share of individuals who revise their neighborhood affiliation in each period, this can now be expressed as

$$(9) \qquad \dot{s_A} = -\dot{s_B} = \begin{cases} (1 - s_A)\rho(q_A - q_B) & \text{when } q_A \geq q_B \\ s_A\rho(q_A - q_B) & \text{when } q_A < q_B \end{cases}.$$

Eq. (9) shows that for migration to come to a rest, i.e., $\dot{s}_A = 0$, we must have either $q_A = q_B$, or complete segregation between Ls and Rs: $s_A = 1$ and $q_A \geq q_B$ or $s_A = 0$ and $q_A < q_B$.

When looking for possible stable equilibria, we must also take into account the dynamics of the ideological view updating, which is what we now turn to.

5. Total dynamics

Let us now bring the elements above together in a complete dynamic model. Eq. (9) above describes the dynamic development in the share of each exogenous type in each peer group. Eq. (8) describes the dynamics of ideological views caused by reluctant social learning in fixed groups, but without taking into account the direct effect of migration on the average ideological view in each peer group.

Writing eq. (8) separately for groups A and B (still for the moment ignoring the short-run changes in q_A and q_B as a direct result of migration, as well as the *O*'s), using the measure $r\delta\Pi_G$ of the total effect of reluctance in social group G defined in that equation, we have:

$$(10) \qquad \dot{q_A} = r\delta\Pi_A$$

$$(11) \qquad \dot{q_B} = r\delta\Pi_B.$$

The set of equations (9) - (11) has one interior solution, $q_A = q_B$ and $s_G = \frac{1}{2}$, which is unstable: a slight deviation causing the ideological views in the two peer groups to differ, say $q_A > q_B$, would attract Rs to A and Ls to B. Thus, if ignoring the direct effects of migration on q_G, reluctance would pull views gradually towards a higher q_A (for example, the wealthy attracted to A would be reluctant to adopt more liberal views), while the opposite happens in B. This process would only stop at the border where $q_A = 1$ and $q_B = 0$ and where $s_A = 1$: L and R types would be in different peer groups; L types would hold the view $q_i = 0$ (e.g., extremely liberal), while R types would hold the view $q_i = 1$ (e.g., extremely conservative).

Migration increases q_A directly to the extent that Rs moving from B to A hold a higher q_i than the Ls migrating in the other direction. Thus, to consider the full effects of migration, an extra term $(q_{RB} - q_{LA})\dot{s}_A$, where $q_{\theta G}$ denotes the average q_i among type $\theta = L, R$ in peer group $G = A, B$, must be added to expression (10), similarly for eq. (11).[8]

Inserting for $\dot{s}_A$ from eq. (9) in the case where the R dominated group is A, and hence $q_A \geq q_B$, gives

(12) $\quad \dot{q}_A = r\delta\Pi_A + \rho(1 - s_A)(q_A - q_B)(q_{RB} - q_{LA})$.

Similarly, for the L dominated group, B,

(13) $\quad \dot{q_B} = r\delta\Pi_B + \rho s_B(q_A - q_B)(q_{LA} - q_{RB})$.

These additional terms do not affect the equilibrium, however, because $\dot{s}_A = (q_{LB} - q_{RA}) = 0$ when the dynamic process has come to a rest (eq. (9)), and similarly for $\dot{s}_B$.

Outside of the steady state, the term $(q_{LA} - q_{RB})$ can in general be either positive or negative, depending on whether the Ls in A on average hold higher or lower q_i than the Rs in B. Since we have not imposed any restrictions on the relationship between individuals' initial ideological view q_i and their exogenous type, it is conceivable that migration temporarily contributes to reductions in q_A and increases in q_B. Nevertheless, over time, reluctance pushes Ls towards gradually lower q_i and Rs towards gradually higher q_i (see Appendix 1), so such reverse movements cannot persist over time.

Intuitively, the average q_i in a given peer group is influenced by two factors, reluctance and migration. In the steady state, migration is by definition zero. Hence, the only possible asymptotically steady state is when reluctance has pushed the average q_i to one of its boundaries, 0 or 1, and thus cannot push it any further.

As mentioned above, *O* types do not migrate. Nor do they contribute to the direction of the movement of q_G within each group, since their social learning is unbiased (Proposition 1). Hence the above discussion is equally valid with *O* types present, now interpreting s_G as the share of R types among the non-*O*'s in group G. Thus, one group will have no L types and agree on the view $q_G = 1$, while the other group will have no R types and agree on the view $q_G = 0$.

We summarize the above discussion in a Proposition, establishing that there is extreme segregation and polarization in the long-run equilibrium:

---

[8] To see this, note that during a time interval Δt, $\dot{s}_A\Delta t$ R-types will move into A, and at the same time equally many L-types will move from A to B. The resulting change in q_A will be $\Delta q_A = (q_{RB} - q_{LA})\dot{s}_A\Delta t$. Letting $\Delta t \to 0$, we get the given equation.

Proposition 2. In the only asymptotically steady states, no group has both *L* and *R* members. Any group with L members has $q_G = 0$; any group with R members has $q_G = 1$. If there are groups with only *O* types, these groups can have $0 \leq q_G \leq 1$.

Since a given social group can either be the one with R types or the one with L types, there are two asymptotically stable states.

Proof: See Appendix 2.

So far in our analysis of migration, we have assumed two equally sized peer groups and equal shares of *L*'s and *R*'s. In Appendix 3, we show that even with unequal shares of L and R types, unequal and possibly endogenous social group sizes, and/or more than two social groups, there is no asymptotically stable state without extreme ideological polarization, such that the first paragraph of Proposition 2 still holds.

In equilibrium, there is not only ideological but also affective polarization: First, people are not interacting socially with their opponents. Moreover, opponents have low regard for each other, even if ideological views are not assumed to be an object of respectability judgements: in equilibrium, R types hold the view that would give an L type the lowest possible social image if socializing with *R*'s, and *L*s similarly hold the view which would give an R type the lowest possible social image if socializing with *L*'s.

The presence of O types limits both kinds of polarization somewhat. First, if there are groups with only O's, the average view q_G in such groups stays constant over time, limiting ideological polarization. Secondly, O types limit affective polarization – not because they judge others differently than their fellow non-O group members (they do not), but because their characteristics are judged less harshly by their opponents.

The highly polarized and segregated steady states described in Proposition 2 display a striking feature: No other combinations of ideological views and sorting into peer groups can improve utility U_i for any individual i. This can easily be seen by recalling eq. (1): $U_i = I_i(\theta_i, q_i) + S_i(\theta_i, q_G)$. Assumption 1 implies that for any $\theta_i \in L$, self-respect I_i is maximized if i's ideological conviction corresponds to $q_i = 0$, while her social respect S_i is maximized if her peers' average view corresponds to $q_G = 0$, both of which hold in equilibrium. Similarly, for $\theta_i \in R$, self-respect is maximized if i's ideological conviction corresponds to $q_i = 1$, while his social respect is maximized if his peers' average view corresponds to $q_G = 0$, both of which hold in equilibrium. For $\theta_i \in O$, utility depends on neither q_i nor q_G, so no other combination of ideological views and sorting into groups can improve their utility.

This is not, of course, a claim that extreme polarization and segregation are generally welfare maximizing. First, we have abstracted from any preferences for respect from non-peers, which would

work in the opposite direction. More importantly, our model abstracts from macro level consequences of polarization such as mistrust, instability, political unrest and so on, which can presumably have substantial negative welfare consequences. Nevertheless, it is worth noting that the social processes we describe here do help people feel respected by themselves as well as their peers, through limiting exposure to views critical to one's personal characteristics.

6. Meeting the others

Above, we simplified the analysis by assuming that learning of ideological views occurs only within peer groups. However, individuals from different peer groups may attend the same schools, colleges and universities; they may interact at work, in sports, and as neighbors; they may consume the same news and entertainment media, read the same novels and watch the same movies, be influenced by pop stars or successful athletes, and even participate in direct ideological debates with their opponents. Thus, it seems implausible that *all* learning of ideological views would occur within peer groups.

For simplicity, let us go back to the assumption of only two peer groups. Recall that in eq. (5), we modelled the change over time in an individual's ideological view q_i^t as the sum of two parts: the unbiased learning from meeting a random peer group member j, plus a term reflecting reluctance. Now, assume instead that with probability κ, where $0 < \kappa < \frac{1}{2}$, the random other person j meets is from the other peer group ($G_i^t \neq G_j^t$,), while with probability $(1-\kappa)$ the other is from one's own peer group ($G_i^t = G_j^t$).

This will not affect migration directly, compared to the situation where $\kappa = 0$; however, it will change the dynamics of social learning. As established in Proposition 3 below, we find that provided that meetings across peer groups do occur, i.e., $\kappa > 0$, ideological polarization is less than extreme in equilibrium but increases in reluctance r. Moreover, from a starting point with moderate to large polarization, polarization is decreasing in κ.

While we relegate the proof of Proposition 3 to Appendix 4, the intuition can be understood as follows. Remember that when i meets j, i is reluctant to learn from j if adopting j's *believed* view $\tilde{q}_{ji}$ would reduce i's self-respect. But such judgements are uncertain, so i will sometimes make errors. Reluctance means that over time, implicitly self-serving errors are given larger weight in i's learning process, pushing polarization towards the extremes.

The variance of $\tilde{q}_{ji}$, however, decreases as views move towards the extremes (Assumption 2). Judgements then become less erroneous, leaving less room for the polarizing effect described above.

When $\kappa > 0$, there is in addition an anti-polarizing force: ideological views are partly learnt from one's opponents, although reluctantly so. This force becomes stronger towards the extremes, since the

difference between q_A and q_B is then large. Consequently, at some point before we get to $q_A = 1$ and $q_B = 0$, the learning effect of meeting opponents from the other group outweighs the polarizing combined effect of uncertainty and reluctance.

Proposition 3 identifies two equilibria when $\kappa > 0$, and these may in fact exist under the same conditions. However, such multiple equilibria exist only within a small range of κ values; they are very similar in terms of ideological views, and both are asymptotically stable.

Proposition 3. Assume that $\kappa > 0$. Then,

i) ideological polarization is incomplete in equilibrium.

ii) Assuming $q_B < q_A$, the equilibrium solution is $q_A = 1 - q_B$, with:

- $q_B = \frac{2\kappa(1-r)}{r\phi(1-\kappa)+4\kappa(1-r)}$ if $\kappa < \frac{r}{2-r}$, and
- $q_B = \frac{\kappa(2-r)}{r\phi(1-\kappa)+2\kappa(2-r)}$ if $\kappa > \frac{r}{2}$.

If $\frac{r}{2} < \kappa < \frac{r}{2-r}$, both equilibria exist.

iii) Both equilibria are asymptotically stable.

Proof: See Appendix 4.

q_B is concave and increasing in κ for both equilibria. A small increase in κ can thus have a large effect when κ is small: starting from a situation with low κ and thus close to full polarization, minor increases in contact across peer groups may reduce polarization substantially. Figure 1 shows q_B as a function of κ, assuming $r = 0.1$ and $\phi = 0.4$. With these parameters, q_B exceeds 0.4 already when $\kappa = \frac{r}{2} = 0.05$ (the left dashed threshold line).

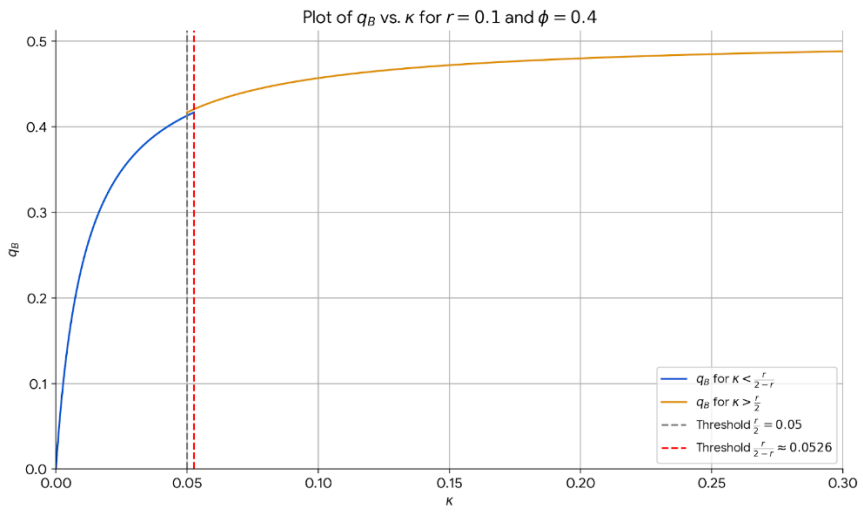

Figure 1: Equilibrium values of q_B as functions of κ.

Policies stimulating meetings across groups would thus help reduce polarization, and so would policies reducing reluctance. Note, however, that the latter presupposes that $\kappa > 0$: if one never met people from other peer groups, it would not matter for ideological view formation whether one would, hypothetically, have listened to them.

7. Exerting effort to gain respect

In the framework presented above, individuals are stuck with their image in the short run and can do nothing to improve it, which may seem unreasonable. Thus, let us now take a brief look at what would happen if individuals may exert costly effort to improve their self-respect and/or social respect.

Let $e_i = e_i^I + e_i^S$ denote the effort i exerts to gain respect, where e_i^I is effort to improve self-respect and e_i^S is effort to improve one's social respect, and let us rewrite eq. (1) as follows:

$$\left(1'\right) \quad U_i = I_i^e\left(e_i^I, \theta_i, q_i\right) + S_i^e\left(e_i^S, \theta_i, q_G\right) - c_i(e_i; \theta_i),$$

where the functions I_i^e and S_i^e are concave and strictly increasing in e_i, while c is increasing and strictly convex in e_i. [9]

From Assumption 1 we can now establish Corollary 1, which demonstrates that the basic foundation for the dynamics above remains the same:

Corollary 1: Utility U_i is decreasing in q_i and q_G for $\theta_i \in L$, increasing in q_i and q_G for $\theta_i \in R$, and independent of q_i and q_G for $\theta_i \in O$.

Proof: Let

$f_i(q_i, q_G) = \max_{e_i}[I_i^e\left(e_i^I, \theta_i, q_i\right) + S_i^e\left(e_i^S, \theta_i, q_G\right) - c_i(e_i; \theta_i)]$, where $e_i = e_i^I + e_i^S$.

Then by the envelope theorem $\frac{\partial f_i}{\partial q_i} = \frac{\partial I_i}{\partial q_i}$ and $\frac{\partial f_i}{\partial q_G} = \frac{\partial S_i}{\partial q_G}$. By Assumption 1 the Corollary follows. ■

Thus, the first general conclusion is that the option to improve image through short-term effort involve no changes to the above dynamics, as individual utility varies with ideological views in the same way

[9] As demonstrated in Nyborg and Brekke (2024), main conclusions hold even if I_i^e and S_i^e are concave but only weakly increasing in e_i. Since would make the proof a bit more cumbersome, however, we keep to the simplest approach here.

as before. By adding more structure to how effort affects image, however, we can derive further conclusions.

Since O types' image is unaffected by ideological views, it is natural to assume that their effort is independent of ideological views too. If so, the sorting of O types has no impact on effort. For simplicity, and without loss of generality, let us thus assume that there are no O types. Moreover, assume that when a good image becomes harder to get, it is also optimal to work harder to get it (see Nyborg and Brekke 2024 for an application where this follows endogenously). Formally, let e_i^* be the utility-maximizing effort level for person i, that is, $e_i^* = \arg\max_{e_i} [I_i^e(e_i^I, \theta_i, q_i) + S_i^e(e_i^S, \theta_i, q_G) - c_i(e_i; \theta_i)]$, where $e_i = e_i^I + e_i^S$, and add Assumption 4:

Assumption 4:

(i) e_i^* is non-decreasing in q_i for $\theta_i \in L$ but non-increasing in q_i for $\theta_i \in R$.

(ii) e_i^* is non-decreasing in q_G for $\theta_i \in L$ but non-increasing in q_G for $\theta_i \in R$.

It now follows that the long-run equilibrium of Proposition 2 (that is, the equilibrium arising in the extreme case with no learning across peer groups) not only represents the strongest possible segregation and ideological polarization – it also represents an absolute effort minimum in the sense defined below.

Definition (absolute effort minimum). A combination of sorting and ethical views is an *absolute effort minimum* if there is no other combination of ethical views and sorting into social groups that would yield strictly lower optimal effort e_i^* for any individual i.

Proposition 4. The long-term equilibrium described in Proposition 2 is an absolute effort minimum.

Proof: Proposition 2 shows that $q_i = 1$ and $q_G = 1$ for all $i \in R$ while $q_i = 0$ and $q_G = 0$ for all $i \in L$ in equilibrium. For an R type, effort is non-increasing in both q_i and q_G by Assumption 4, hence effort achieves a minimum when $q_i = 1$ and $q_G = 1$. Similarly, effort is minimal for L types, as effort for them is non-decreasing in both q_i and q_G. ■

8. Discussion

Before concluding, let us briefly and informally discuss a few possible modifications. First, could benefits of trade between diverse peer groups modify or prevent polarization? If there was a profit to be gained by interacting with one's opponents, some people may choose to join or stay in a peer group giving them lower social respect, in order to seek such profit. Proposition 1, however, establishes that ideological polarization arises even in the absence of segregation (assuming $r < \frac{2}{5}$). Hence, such individuals would gradually adopt the ideological view held by the majority of their group, regardless

of their own type; segregation of L and R types would be incomplete, but extreme polarization (in the absence of meetings across groups) would remain.

However, if benefits of trade instead cause more meetings between individuals from different peer groups, which seems plausible, such benefits would help limit polarization. Moreover, if individuals are better positioned to reap such gains when they have a reasonable understanding of their opponents' way of thinking, benefits of trade could conceivably also reduce learning reluctance.

A cost of migration will not necessarily prevent polarization (Nyborg and Brekke 2024). Clearly, it cannot *reverse* polarization: In a polarized equilibrium, no-one wants to move anyway, so introducing a migration cost at that point would change nothing. What if a fixed, strictly positive migration cost is introduced *before* segregation is complete? So far, we have only specified the direction in which q_G affects utility, depending on the individual's type (L, R, or O). Assume now that there is heterogeneity in the *size* of such utility changes within each type, such that individual benefits of moving between groups differ between individuals of similar type. A migration cost would then prevent some marginal potential movers – those who are close enough to indifference between peer groups – from moving. This would not hinder polarization, however, again due to Proposition 1: polarization arises even when segregation is limited.

As already hinted at in the discussion of trade benefits above, it may be plausible to let κ, the probability of meeting someone from a different peer group, depend endogenously on incentives for such meetings. In addition to trade benefits, the expected respect or disrespect from the other may be one such incentive. Similarly, while r, the degree of reluctance in ideological learning, is exogenous in our model, this might possibly vary with the ideological distance between oneself and the perceived position of the other, on the respect one experiences from the other, or the possible economic gains of being able to secure a profitable trade with the other. Nevertheless, we leave such extensions for future work.

9. Conclusions

Above, we have demonstrated how the wish to be respected by oneself and others can push society towards ideological and affective polarization. In equilibrium, almost everyone's views are polarized; opponents are segregated in separate social environments, and there is widespread mutual disrespect. While the negative externalities associated with such social division are presumably substantial, our framework shows that from a narrow individual perspective, it also has the welfare-improving effect of limiting people's exposure to ideological views that would have threatened their self-respect and social image.

We find that two key parameters determine the level of equilibrium polarization: the probability that a random meeting is with someone from outside one's own peer group, and the degree of reluctance in ideological learning. Thus, exogenous shocks and policies affecting these parameters can affect polarization.

In our model, meeting with someone essentially means being exposed to their views. Hence, the fragmentation of media information flows after the establishment of the internet can be considered one such negative shift in the first parameter, increasing equilibrium polarization. Policies possibly counteracting this effect could include, for example, subsidizing cross-cutting media, stricter regulation of social media platforms and personalized advertising, or, as in an interesting experiment by Akbiyik et al. (2024), providing incentives to familiarize oneself with opponent media. Other policies stimulating meetings across peer groups may include, e.g., worker protection regulations limiting employers' ability to hire or fire workers based on their opinions or job irrelevant personal characteristics; education policies encouraging a mix of diverse student groups in schools; subsidized public childcare involving contact between diverse families; or cultural policies stimulating cross-cutting literature, art, and sports experiences.

Even when occasionally meeting one's opponents, people may be reluctant to learn from them. In our framework, learning is biased in the sense that one is less easily convinced by views threatening one's self-respect – and in equilibrium, this will often be the case for opponent views. Although reluctance reflects a psychological phenomenon which can hardly be regulated directly, the degree of reluctance may still conceivably be affected by, for example, training students' open-mindedness and curiosity to others' views, or, in the other direction, role models acting aggressively and scornfully towards their opponents in public.

Finally, note that the basic assumption driving the polarizing dynamics in our framework is the idea that those on opposite sides of the ideological spectrum do not equally respect all individual characteristics. If everyone were equally respected by all – i.e., in the language of the model, if everyone were an O type – the social dynamics considered above would drive neither segregation nor polarization.

Appendix 1: On asymptotically stable states in the case without migration

In the case without migration, reluctance pulls in the direction of increasing q_i for *R* types while decreasing it for *L* types; this drives the group average q_G^t towards zero if the majority is in *L* but towards 1 if the majority is in *R*. However, since all *R* (*L*) within a given group are subject to the same dynamic, they become increasingly homogenous over time. The purpose of the present Appendix is to explore whether there may exist steady states in which both types within the same group converge to different views (given no migration).

Let us simplify notation by writing $s_G = s$ for the share of *R* individuals in the group, and suppress the explicit notation of time. Without loss of generality, let $\delta = 1$, which only affects the speed of convergence but not the direction or state to which it converges.

We first establish that in a stable state all *R* will have homogenous views, and so will all *L*:

Lemma *A1-1: Ideological views q_i^t converge to one common view q_R for all $i \in R$ and one common view q_L for all $i \in L$, finally q_i^t converge to $q_O = s_R q_R + (1 - s_R) q_L$ for $i \in O$.*

Proof: We prove this by first considering the effect of two different *R* individuals conditional on meeting the same *j* and forming the same belief of $j^{'}s$ view. Then we take the unconditional expectation.

From (5), if $i \in R$ meets an individual *j*, then

$$(A1-1) \qquad \Delta q_i^t = \left(\tilde{q}_{ji}^t - q_i^t\right)\Delta t + r\left(\tilde{q}_{ji}^t - q_i^t\right)^{-}\Delta t$$

If we consider two different *R* individuals $i, i^{'}$, with $q_i^t > q_{i^{'}}^t$, meeting the same *j*, and perceiving the same $\tilde{q}_{ji}^t$ then

$$(A1-2) \qquad \Delta\left(q_i^t - q_{i^{'}}^t\right) = -\left(\left(q_i^t - q_{i^{'}}^t\right) + r\left(\left(\tilde{q}_{ji}^t - q_i^t\right)^{-} - \left(\tilde{q}_{ji}^t - q_{i^{'}}^t\right)^{-}\right)\right)\Delta t$$

$$< -(1-r)\left(q_i^t - q_{i^{'}}^t\right)\Delta t < 0$$

Thus, contingent on meeting the same *j* and perceiving the same $\tilde{q}_{ji}^t$, the views of the two *R* individuals move closer together. As we move to continuous time and infinite population, the randomness concerning which *j* one meets cancels out, and we are left with the unconditional expectation

$$(A1-3) \qquad \left|\dot{q}_i^t - \dot{q}_{i'}^t\right| < -(1-r)\left|\left(q_i^t - q_{i'}^t\right)\right|.$$

The same argument applies to any two L individuals.
A similar argument applies to O-types, but as noted above ideological views tend to the group average in the absence of reluctance. Hence with a share $(1-a)$ of O-types and s_R being the share of R types among the remaining, we get $q_O = (1-a)q_O + a(s_R q_R + (1-s_R)q_L)$, which implies $q_O = s_R q_R + (1-s_R)q_L$. ■

Lemma A1-1 implies that eventually all R will hold approximately the same view, and similarly with all L. Hence, to consider asymptotic stability we can limit attention to the case where all L within a given group hold exactly the same view q_L, while all R hold the same view q_R.

Under this assumption, the dynamic of the view of the two types are (from eqs. (6) and (7) in the main text):

$$(A1-4) \qquad \dot{q}_R = (1-s)(q_L - q_R) + srB_{RR}^- + (1-s)rB_{RL}^-$$

and

$$(A1-5) \qquad \dot{q}_P = s(q_R - q_L) - srB_{LR}^+ - (1-s)rB_{LL}^+.$$

We will be particularly interested in the dynamics of how the different groups differ and how the average evolves. Note that these can be simplified as

$$(A1-6) \qquad \dot{q}_R - \dot{q}_L = -(q_R - q_L) + r(s(B_{RR}^- + B_{LR}^+) + (1-s)(B_{RL}^- + B_{LL}^+))$$

And, if we let q denote the average ethical view in the entire group (recall that group composition is still assumed to be fixed),

$$(A1-7) \qquad \dot{q} = s\dot{q}_R + (1-s)\dot{q}_L = r\left((s^2 B_{RR}^- - (1-s)^2 B_{LL}^+) + s(1-s)(B_{RL}^- - B_{LR}^+)\right).$$

We note that if $(q_R - q_L)$ is large and r is small, the first term in (A1-6) will dominate and $\dot{q}_R - \dot{q}_L \approx -(q_R - q_L)$. Hence the difference in view between types will decline over time. Eventually they will become rather similar, and then $B_{RR}^- \approx B_{LL}^+$ and $B_{RL}^- \approx B_{LR}^+$, and by $(A1-7)$ the average q_i will decline with a L majority and increase with a R majority. The following proof makes this argument precise. Note in particular that the biases are only approximately equal, thus there is a possibility that reluctance will pull harder on one group than the other.

Recall that in the main text, we assumed that the distribution of $\tilde{q}_{ji}$ is symmetrical, unbiased and has support in [0,1]. Two alternative distributions that satisfy this are:

Alternative assumptions on the probability distribution of $\tilde{q}_{ji}$

(a) *Binary distribution:* $\tilde{q}_{ji} = q_j \pm \phi d_j$ *with equal probability*

(b) *Uniform distribution:* $\tilde{q}_{ji} \sim U(q_j - \phi d_j, q_j + \phi d_j)$

where $d_j = min(q_j, 1 - q_j)$ *is the distance between* q_j *and the border of* [0,1], *and* $0 < \phi \leq 1$.

We first consider the case of a binary distribution.

Binary distribution

We first consider the case of a binary distribution. Here we have the following theorem:

Theorem A1-1: *If* $r < \frac{2}{5}$, *the only asymptotically stable states are* $q = 0$ *if* $s < \frac{1}{2}$ *and* $q = 1$ *if* $s > \frac{1}{2}$.

First, we establish that if the two types hold sufficiently different views, their views will approach each other.

Lemma A1-2: *With a binary distribution, if* $q_L \leq \frac{1}{2}$, $q_R > (1 + \phi)q_L$ *and* $r < \frac{2}{5}$ *then* $\dot{q}_R - \dot{q}_L < 0$.

Proof: Let $\Delta = q_R - q_L$. Note first that by the assumption of the lemma, $\Delta > \phi q_L$. Moreover, $\phi q_R = \phi q_L + \phi\Delta < (1 + \phi)\Delta$. Given the binary distribution,

$$B_{RR}^{-} = \frac{\phi}{2} q_R < \frac{(1+\phi)}{2}\Delta$$

$$B_{LL}^{+} = \frac{\phi}{2} q_L < \frac{1}{2}\Delta$$

We have also assumed that views are so different that *R* are always reluctant when meeting a *P* type:

$$B_{RL}^{-} = \Delta$$

For the last bias, note that the perceived level of q_R is $q_R \pm \phi d_R$, and $d_R = \min(q_R, 1 - q_R)$. For the last bias there are two cases:

$$B_{LR}^{+} = \begin{cases} \Delta & for\ q_R - \phi d_R > q_L \\ \frac{1}{2}(q_R + \phi d_R - q_L) & for\ q_R - \phi d_R > q_L \end{cases}$$

Note that, $\phi d_R \leq \phi q_R < (1 + \phi)\Delta$. It follows that

$$B_{LR}^{+} < \frac{(2+\phi)}{2}\Delta.$$

Let $q = sq_R + (1-s)q_L$.

$$\dot{q}_R - \dot{q}_L = -\Delta + sr(B^-_{RR} + B^+_{LR}) + (1-s)r(B^-_{RL} + B^+_{LL})$$

$$< -\Delta + sr\left(\frac{(1+\phi)}{2} + \frac{(2+\phi)}{2}\right)\Delta + (1-s)r(1+\frac{1}{2})\Delta$$

$$= -\Delta + r\left(s\left(\frac{3}{2}+\phi\right) + (1-s)\frac{3}{2}\right)\Delta = \left(1 - r\left(\frac{3}{2}+\phi\right)\right)\Delta$$

$$< 0 \quad \text{if} \quad r < \frac{2}{5}$$

■

We next want to extend this to the case with an O-type. Disregarding reluctance for the moment, then $\dot{q}_R = a(1-s)(q_L - q_R) + (1-a)(q_O - q_R)$ and $\dot{q}_L = as(q_R - q_L) + (1-a)(q_O - q_L)$ and it follows that $\dot{q}_R - \dot{q}_L = (q_L - q_R) = -\Delta$, in the absence of reluctance. Hence

$$\dot{q}_R - \dot{q}_L = -\Delta + r(as(B^-_{RR} + B^+_{LR}) + (1-s)a(B^-_{RL} + B^+_{LL}) + (1-a)(B^-_{RO} + B^+_{LO})$$

Now we claim reluctance when meeting an O type is less than the average reluctance when meeting an $B^-_{RO} \leq sB^-_{RR} + (1-s)B^-_{RL}$, and similar for B^+_{LO}.

Next, we need to show that when the views of the two types are sufficiently close, then everyone will move toward the view most favorable to the majority.

Lemma A1-3: *With a binary distribution, if* $q_L \leq \frac{1}{2}$, $q_R \leq (1+\phi)q_L$, *then* $\dot{q} = s\dot{q}_R + (1-s)\dot{q}_L < 0$, if $s < \frac{1}{2}$, and $\dot{q} > 0$ if $s > \frac{1}{2}$.

Proof: Consider first the case with no O-types. As before $B^-_{RR} = \frac{\phi}{2}d_R$ and $B^+_{LL} = \frac{\phi}{2}d_L$. Moreover, $B^+_{LR} = \frac{1}{2}(q_R - q_L + \phi d_R)$, while $B^-_{RL} = \frac{1}{2}(q_R - q_L + \phi d_L)$. Remember from eq. (A1-7) that

$$\dot{q} = s\dot{q}_R + (1-s)\dot{q}_L = sr(sB^-_{RR} + (1-s)B^-_{RL}) - (1-s)r(sB^+_{LR} + (1-s)B^+_{LL})$$

$$= sr\left(s\frac{\phi}{2}d_R + (1-s)\frac{1}{2}(q_R - q_L + \phi d_L)\right) - (1-s)r\left(s\frac{1}{2}(\phi d_R + q_R - q_L) + (1-s)\frac{\phi}{2}d_L\right)$$

$$= \frac{\phi r}{2}(s^2 d_R - (1-s)^2 d_L) + \frac{s(1-s)r}{2}(q_R - q_L) - \frac{s(1-s)r}{2}(q_R - q_L) - \phi\frac{s(1-s)r}{2}(d_R - d_L)$$

$$= \frac{\phi r}{2}\left(s^2 d_R - (1-s)^2 d_L - s(1-s)(d_R - d_L)\right)$$

$$= \frac{\phi r}{2}\left(s\big(s-(1-s)\big)d_R - (1-s)(1-s-s)d_L\right)$$

$$= \frac{\phi r}{2}(s(2s-1)d_R - (1-s)(1-2s)d_L)$$

$$= \frac{\phi r}{2}(2s-1)(sd_R + (1-s)d_L)$$

We see that $\dot{q} > 0$ for $s > \frac{1}{2}$ and $\dot{q} < 0$ for $s < \frac{1}{2}$.

Now, if a share $(1-a)$ are of type O, and the remaining are L or R, where s is the share of these being of type R, then the total dynamics becomes:

$$\dot{q} = (1-a)\dot{q}_O + a(s\dot{q}_R + (1-s)\dot{q}_L)$$

As above, there would be no movement in the average in absence of reluctance, so only consider the effect of reluctance, which is limited to the term $s\dot{q}_R + (1-s)\dot{q}_L$. Note here that R and L types are so close that reluctance only applies in one of the two points in the binary distribution. With O types in between, reluctance when meeting an O-type also only applies in one of the two points in the support of the distribution. Hence the reluctance part of $s\dot{q}_R + (1-s)\dot{q}_L$ becomes

$$sr(asB^-_{RR} + a(1-s)B^-_{RL} + (1-a)B^-_{RO}) - (1-s)r(asB^+_{LR} + a(1-s)B^+_{LL} + (1-a)B^-_{LO})$$

$$= asr(sB^-_{RR} + (1-s)B^-_{RL}) - (1-s)r(sB^+_{LR} + (1-s)B^+_{LL}) + a(1-a)r(sB^-_{RO} - (1-s)B^-_{LO})$$

The first part of this is calculated above, so we focus on the last term

$$sB^-_{RO} - (1-s)B^-_{LO} = \frac{s}{2}(q_R - q_O - \phi d_O) + \frac{1-s}{2}(q_O - q_L + \phi d_O)$$

$$= \frac{s}{2}(q_R - q_O) - \frac{1-s}{2}(q_O - q_L) - \frac{s}{2}\phi d_O + \frac{1-s}{2}\phi d_O$$

$$\frac{1}{2}(q_O - ((1-s)q_L + sq_R))) + \frac{1-2s}{2}\phi d_O$$

$$= \frac{1}{2}\big((1-2s)\phi d_O\big)$$

We have here used the fact that in the long run $q_O = s_R q_R + (1-s_R)q_L$. Combining this with the calculation above, for the case with no O-types, we find

$$\dot{q} = \frac{\phi r}{2}(2s-1)a(asd_R + a(1-s)d_L + (1-a)d_O)$$

As above the direction of the movement is determined by the sign of $2s-1$, that is, q increases if there are more R than L types and vice versa.

■

Recall the claim of the theorem we want to prove: *If* $r < \frac{2}{5}$, *the only asymptotically stable states are* $q = 0$ *if* $s < \frac{1}{2}$ *and* $q = 1$ *if* $s > \frac{1}{2}$.

Proof of Theorem A1-1: Note first that we have chosen s to be the share of R and $q = 1$ to represent the view implicitly most self-serving to R types. Alternatively we could use $\check{s} = 1 - s$ as the share of L and $\check{q} = 1 - q$ denoting the view most favorable to L types. Here L will have a higher $\check{q}$ than R. As the model is symmetric, Lemma A1-2 would still be valid, with $\check{s}$ and $\check{q}$ replacing s and q and with R and L changing roles. Now, by this extension of Lemma A1-2, we know that for $\check{q}_R \leq \frac{1}{2}$ and $\check{q}_L > (1 + \phi)\check{q}_R$, the difference in view between the two groups will be declining when $\check{s} < \frac{1}{2}$. But $\check{q}_R \leq \frac{1}{2}$ is equivalent to $q_R \geq \frac{1}{2}$. From Lemma A1-2 we already know that the conclusion holds for $q_R \leq \frac{1+\phi}{2}$. Thus by this symmetry we conclude that the conclusion holds everywhere. The same symmetry argument extends the conclusions of Lemma A1-3 to hold everywhere.

Thus, using the lemmas above, we see that by Lemma A1-1 the views of all R will tend toward a common view, and the same for the L. If the two types hold sufficiently different views, they will tend toward each other by Lemma A1-2. Thus, we know that they will hold sufficiently similar views for Lemma A13-3 to apply. Note that the inequalities $\dot{q}_R - \dot{q}_L < 0$ and $\dot{q} < 0$ for $s < \frac{1}{2}$, are strict, and hence the movement will go back toward the stable state after a small deviation. The states are thus asymptotically stable. ■

Note that the theorem does not state what happens when $s = \frac{1}{2}$. However, from the proof of Lemma A1-3 we see that $\dot{q} = 0$ if $q_L < \frac{1}{2}$ and $q_R \leq (1 + \phi)q_L$, and by symmetry this also applies with $q_L < \frac{1}{2}$ and $d_L \leq (1 + \phi)d_R$. Thus any state satisfying these conditions are stable. With a slight deviation from this state we still have $\dot{q} = 0$, thus there is no movement back to the original stable state so none of these stable states are asymptotically stable.

Uniform distribution

With a uniform distribution we will need to use a numerical approximation to get a better picture of the asymptotically stable states. We start by proving a key result:

Lemma A1-4: *With a uniform distribution and no O types*, if $(q_R - q_L) = \Delta \leq \min(d_L, d_R)$ then for $q_L < q_R \leq \frac{1}{2}$,

$$\dot{q} = \frac{\phi r}{4}(2s - 1)(sd_R + (1 - s)d_L) + \frac{s(1 - s)r}{4\phi d_R d_L}\Delta^3$$

Proof: Remember from A1-7 that

$$\dot{q} = s\dot{q}_R + (1-s)\dot{q}_L = sr(sB_{RR}^- + (1-s)B_{RL}^-) - (1-s)r(sB_{LR}^+ + (1-s)B_{LL}^+)$$

With a uniform distribution, $B_{RR}^- = \frac{\phi}{4}d_R$ and $B_{LL}^+ = \frac{\phi}{4}d_L$. By the assumption of the lemma, q_R is inside the support of the distribution around q_L and similarly the other way around. It follows from the properties of a uniform distribution that $B_{LR}^+ = \frac{1}{2}\frac{(\Delta+\phi d_R)^2}{2\phi q_R}$ and $B_{RL}^- = \frac{1}{2}\frac{(\Delta+\phi d_L)^2}{2\phi d_L}$. We collect terms and simplify:

$$sr(sB_{RR}^-) - (1-s)r(1-s)B_{LL}^+ = \frac{\phi r}{4}(s^2 d_R - (1-s)^2 d_L)$$

and

$$sr\big((1-s)B_{RL}^-\big) - (1-s)r(sB_{LR}^+) = \frac{s(1-s)r}{4\phi d_L d_R}(d_R(\Delta+\phi d_L)^2 - d_L(\Delta+\phi d_R)^2),$$

Now,

$$d_R(\Delta+\phi d_L)^2 - d_L(\Delta+\phi d_R)^2 = +\Delta^3 - \phi^2 d_R d_L \Delta$$

For $q_L < q_R \leq \frac{1}{2}$, $q_L = d_L$ *and* $q_R = d_L$ so

$$\begin{aligned}
\dot{q} &= \frac{\phi r}{4}(s^2 q_R - (1-s)^2 q_L) + \frac{s(1-s)r}{4\phi}\left(-\phi^2\Delta + \frac{\Delta^3}{q_R q_L}\right) \\
&= \frac{\phi r}{4}\big(s^2 q_R - (1-s)^2 q_L - s(1-s)(q_R - q_L)\big) + \frac{s(1-s)r}{4\phi q_R q_L}\Delta^3 \\
&= \frac{\phi r}{4}(2s-1)(sq_R + (1-s)q_L) + \frac{s(1-s)r}{4\phi q_R q_L}\Delta^3 \qquad \blacksquare
\end{aligned}$$

Note that the lemma implies that $\dot{q} > 0$ when $s \approx \frac{1}{2}$, indicating that there is an area $s \in [\frac{1}{2} - \epsilon, 1]$ where the average q is increasing for $q_L < q_R \leq \frac{1}{2}$. Using the transformation with $\check{s}$ and $\check{q}$ as in the proof of the theorem, we conclude, by symmetry, that there is an area $s \in [0, \frac{1}{2} + \epsilon]$ where the average q is decreasing for $\frac{1}{2} \leq q_L < q_R$. Hence there will be a stable equilibrium with $q_L < \frac{1}{2} < q_R$ in the interval $s \in [\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon]$.

Lemma A1-5: *If for some* $K \geq 1$, $\Delta > \frac{\phi}{K} min(d_R, d_L)$ *and* $r < \frac{4}{5+2K}$, *then* $\dot{q}_R - \dot{q}_L < 0$.

Note that the lemma allows us to conclude that in the long run: $\Delta \leq \frac{\phi}{K} min(d_R, d_L)$. Thus the larger we may choose K the smaller we can assume Δ to be. On the other hand, a higher K implies that we must

assume smaller reluctance as $r < \frac{4}{5+2K}$. Let $q_G^h = q_G + \phi d_G$ and $q_G^l = q_G - \phi d_G$ denote the high and low border of the support of the uniform distribution, for each group G.

Proof: Note that the condition implies that $\phi q_L < K\Delta$. And $\phi d_R \leq \phi q_R = \phi(q_L + \Delta) < (K + \phi)\Delta$. Moreover,

$$B_{RL}^{-} = \begin{cases} (q_R - q_L)\ for\ q_L^h < q_R \\ \frac{1}{2}\frac{(q_R - q_L^l)^2}{q_L^h - q_L^l} < \frac{K+1}{2}\Delta \quad for\ q_L^h \geq q_R \end{cases}$$

The inequality follows as for $q_L^h \geq q_R$ then $\frac{q_R - q_L^l}{q_L^h - q_L^l} \leq 1$. Next, in a similar fashion.

$$B_{LR}^{+} = \begin{cases} (q_R - q_L) & for\ q_R^L > q_L \\ \frac{1}{2}\frac{(q_R^h - q_L)^2}{q_R^h - q_R^l} < \frac{(K+1+\phi)}{2}\Delta & for\ q_R^L \leq q_L \end{cases}$$

Let $q = sq_R + (1-s)q_L$

$$(A1-8) \qquad \begin{aligned} \dot{q}_R - \dot{q}_L &= -(q_R - q_L) + sr(B_{RR}^{-} + B_{LR}^{+}) + (1-s)r(B_{RL}^{-} + B_{LL}^{+}) \\ &\leq -\Delta + sr\left(\frac{\phi}{4}d_R + \frac{3+K}{4}\Delta\right) + (1-s)r\left(\frac{3+K}{4}\Delta + \frac{\phi}{4}d_L\right) \\ &\leq -\Delta\left(1 - \frac{2+2K+2\phi}{4}r\right) + \frac{r\phi}{4}\bar{d} \leq -\Delta\left(1 - \frac{2+3K+3\phi}{4}r\right) \\ &< 0\ for\ r < \frac{4}{5+2K} \qquad \blacksquare \end{aligned}$$

Theorem A1-2: With a uniform distribution, and with $r < \frac{1}{2}$, and $K = \frac{4-5r}{2r}$ there are $\underline{s}$ and $\underline{s}$ such that $\frac{1}{2} - \frac{\phi}{8K^3} \leq \underline{s} < \frac{1}{2} < \overline{s} \leq \frac{1}{2} + \frac{\phi}{8K^3}$, and such that $q = 0$ is asymptotically stable for $s \in [0, \underline{s})$. And $q = 1$ is asymptotically stable for $s \in (\overline{s}, 1]$. For $s \in I \subset (\underline{s}, \overline{s})$ there is a stable state with $q_L < \frac{1}{2} < q_R$.

Proof: Lemma A1-5 shows that with $r < \frac{1}{2} < \frac{4}{5+2}$ we can ensure that Δ satisfies the conditions of Lemma A1-4. Lemma A1-4 implies that under the conditions of the lemma, $\dot{q} > 0$ when $s \approx \frac{1}{2}$. That is, there is an area $s \in [\frac{1}{2} - \epsilon, 1]$ where the average q is increasing if $q_L < q_R \leq \frac{1}{2}$, and $q_R - q_L$ is sufficiently small. Using the transformation with $\check{s}$ and $\check{q}$ as in the proof of the theorem, we conclude, by symmetry, that there is an area $s \in [0, \frac{1}{2} + \epsilon]$ where average q is decreasing for $\frac{1}{2} \leq q_L < q_R$.

Hence there will be a stable equilibrium with $q_L < \frac{1}{2} < q_R$ in the interval $s \in [\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon]$. Outside this interval the term $\frac{\phi r}{4}(2s-1)(sd_R + (1-s)d_L)$ will dominate.

It remains to show the bounds on the interval. We can utilize $\Delta \leq \frac{\phi}{K} min(q_R, q_L)$ when $q_L < q_R \leq \frac{1}{2}$ to give an estimate of the interval around $\frac{1}{2}$ where there exists an intermediate asymptotically stable state. Since $\Delta < \frac{\phi}{K} q_L$,

$$\frac{s(1-s)r}{4\phi q_R q_L}\Delta^3 < \frac{\phi r}{4}\left(\frac{s(1-s)}{K^2}\Delta\right) < \frac{\phi r}{4}\left(\frac{s(1-s)\phi}{K^3}\right)q$$

Hence

$$\dot{q} < \frac{\phi r}{4}\left[(2s-1) + \frac{\phi}{K^3}s(1-s)\right]q$$

We consider the sign of the expression inside the brackets. This can be simplified as $(2s-1) + \Phi s(1-s)$ with $\Phi = \frac{\phi}{K^3}$. Note that $(2s-1) + \Phi s(1-s)$ is negative for $s < \frac{1}{2} - \frac{\sqrt{\Phi^2+4}-2}{2\Phi} \leq \frac{1}{2} - \frac{\Phi}{8} = \frac{1}{2} - \frac{\phi}{8K^3}$. To see this, Let $f(\Phi) = \frac{\sqrt{\Phi^2+4}-2}{2\Phi}$ then $f(\Phi) = f(0) + f\left(\Phi^{'}\right)\Phi$ with $0 \leq \Phi^{'} \leq \Phi$. Using L'Hopital we find $\lim_{\Phi \to 0} f(\Phi) = 0$. Moreover $f^{'}(\Phi) = \frac{\sqrt{\Phi^2+4}-2}{\Phi^2\sqrt{\Phi^2+4}}$ which is declining and $\lim_{\Phi \to 0} f^{'}(\Phi) = \frac{1}{8}$, using L'Hopital. Thus[10] $f(\Phi) \leq \frac{\Phi}{8}$. This gives the approximation that q is declining for $s < \frac{1}{2} - \frac{\phi}{8K^3}$. ■

If we choose $r = 0.1$, then $K = 17.5$. Now, with $\phi = 0.5$ we find $\frac{\phi}{8N^3} = 1.17 \cdot 10^{-5}$. In this case q is declining for $s < \frac{1}{2} - \frac{\phi}{8K^3} \approx \frac{1}{2} - 1.17 \cdot 10^{-5}$.

Note that since we have to revert to approximations, we cannot prove that there is an intermediate stable state for all $s \in (\underline{s}, \overline{s})$, only that this is true in a subset I which must include $s = \frac{1}{2}$.

Appendix 2: Proof of Proposition 2

[10] This estimate is rather precise as $f^{''}(\Phi) \approx 0$ around $\Phi = 0$, ($|f^{''}(\Phi)| < 0.0001$ for $\Phi < 0.1$, evaluated numerically).

Proof: In a stable state, $\dot{s}_{RA}^t = -\dot{s}_{RB}^t = -\dot{s}_{LA}^t = \dot{s}_{LB}^t = 0$ and $\dot{q}_A^t = \dot{q}_B^t = 0$. Migration adds a term $(q_{LB} - q_{RA})\dot{s}_{RA}$ to (10) and similarly to (11). But as $\dot{s}_{RA}^t = 0$ in a stable state, this addition vanishes in the stable state; thus any stable state would also be a stable state without migration. Proposition 1 shows that, without migration, assuming A is a group with a L majority, there is only one stable state: $q_A = 0$. In this case B would be a group with a R majority, with $q_B = 1$ as the only stable state. Since $q_A = 0$ and $q_B = 1$, then when we allow migration the Rs in A will migrate to B, while the Ls in B migrate in the other direction. As long as one group has slightly higher average q_G, R-types will migrate towards that group. With a small perturbation from the stable state with $q_i = q_R = 1$, R-types will have incentives to migrate toward the group with a majority of R-types. Hence this stable state is asymptotically stable. Thus, the only stable state is when all Rs are in one group and all Ls in the other. Exactly the same argument applies with groups A and B interchanged.
If A has an equal share of each type, there is an additional stable state as discussed in Proposition 1, Part III, in which R and L within the same group converge to different but less extreme views. However, this state is asymptotically unstable: any slight deviation making the average view in one group lower than the other initiates a migration toward one of the asymptotically steady states discussed above. ■

Appendix 3: Generalization

Proposition 3-1. With unequal shares of $R\acute{}\ s$ and $L\acute{}\ s$, and with several groups of exogenous or endogenous and potentially different size, there is no asymptotically stable state without complete segregation and polarization.

Proof: Note first that by Proposition 2, groups would either hold view $q_G = 0$, $q_G = \frac{1}{2}$ or $q_G = 1$. One intermediate group with $q_G = \frac{1}{2}$ and equally many R and L cannot coexist with groups with either $q_G = 0$ or $q_G = 1$, due to migration, as R types would move to groups with $q_G = 1$ and L types would move to groups with $q_G = 0$. The intermediate alternative requires equally many R and L types in all groups, which is impossible if the shares of R and L types in the population are different, and the state where $q_G = \frac{1}{2}$ is not asymptotically stable either. We are left with the two extremes, $q_G = 0$ and $q_G = 1$. To avoid social migration in a situation with both R and L types in at least one of the groups, both groups must hold the same ideological view, e.g. $q_G = 0$ in both groups. But if the average view in one group changes slightly, migration would start; and once R types constitute the majority in one group, $q_G = 0$ is no longer a stable situation in that group. Hence the only stable state is when the two types are segregated in different groups, and R types hold the view $q_i = 1$ while L types hold the view $q_i = 0$. Group size, which could potentially be endogenous to migration, does not matter for the argument above. Note also that the presence of O types does not affect the argument. ■

Appendix 4: Proof of Proposition 3

Proposition 3. Assume that $\kappa > 0$. Then,

iv) ideological polarization is incomplete in equilibrium.

v) Assuming $q_B < q_A$, the equilibrium solution is $q_A = 1 - q_B$, with:

- $q_B = \frac{2\kappa(1-r)}{r\phi(1-\kappa)+4\kappa(1-r)}$ if $\kappa < \frac{r}{2-r}$, and
- $q_B = \frac{\kappa(2-r)}{r\phi(1-\kappa)+2\kappa(2-r)}$ if $\kappa > \frac{r}{2}$.

If $\frac{r}{2} < \kappa < \frac{r}{2-r}$, both equilibria exist.

vi) Both equilibria are asymptotically stable.

Proof: Let A be the social group with R types (and possibly O types), such that in equilibrium $q_B \leq \frac{1}{2} \leq q_A$ and $s_A = 1$ while $s_B = 0$.

Assume for the moment that views in peer groups A and B are sufficiently different to ensure that members of A are always reluctant when meeting someone from B and vice versa. Then, with continuous time, incorporating reluctance and adding the direct effect of migration on q_G (see eqs. (12) - (13) and the discussion thereof), the equilibrium condition now becomes

$$(A3-1) \qquad \dot{q}_A = \delta\kappa(1-r)(q_B - q_A) + r\delta(1-\kappa)\Pi_A = 0.$$

The first term reflects the effect of meeting someone from the other group with probability κ, while the second term is as above but adjusted for the lower probability of meeting someone from one's own group. Thus, in equilibrium

$$(A3-2) \qquad \kappa(1-r)\delta(q_A - q_B) = r\delta(1-\kappa)\Pi_A.$$

Similarly, for group B,

$$(A3-3) \qquad \kappa(1-r)\delta(q_B - q_A) = r\delta(1-\kappa)\Pi_B.$$

This rules out $q_A = 1$ and $q_B = 0$, since if $q_A = 1$ we must have $\Pi_A = 0$. Similarly for B. Consequently, we no longer get complete polarization.

Now, recall that $\Pi_G = [s_G^2 B_{RR}^- + (1-s_G)s_G(B_{RL}^- - B_{LR}^+) - (1-s_G)^2 B_{LL}^+]$. Inserting s_G from above, i.e., $s_A = 1$ and $s_B = 0$, this implies that $\Pi_A = B_{RR}^-$ while $\Pi_B = -B_{LL}^+$.

Next, recall from above (e.g. Proof of Lemma A1-2) that $B_{RR}^{-} = \frac{\phi}{2}\min(q_R, 1-q_R)$, while $B_{LL}^{+} = \frac{\phi}{2}\min(q_L, 1-q_L)$. Thus $\Pi_A = B_{RR}^{-} = \frac{\phi}{2}(1-q_A)$ while $\Pi_B = -B_{LL}^{+} = -\frac{\phi}{2}q_B$. The equilibrium conditions $(A3-2)$ and $(A3-3)$ can now be written

$$(A3-4) \qquad \kappa(1-r)\delta(q_A - q_B) = r\delta(1-\kappa)\frac{\phi}{2}(1-q_A)$$

and

$$(A3-5) \qquad \kappa(1-r)\delta(q_A - q_B) = r\delta(1-\kappa)\frac{\phi}{2}q_B.$$

As the left-hand side is identical in both these equations, this implies that $q_B = 1-q_A$, so $q_A - q_B = 1-2q_B$, and the second condition can be rewritten

$$\frac{\kappa(1-r)}{(1-\kappa)r}(1-2q_B) = \frac{\phi}{2}q_B.$$

Solving yields

$$(A3-6) \qquad q_B = \frac{\frac{\kappa(1-r)}{r(1-\kappa)}}{\frac{\phi}{2}+2\frac{\kappa(1-r)}{r(1-\kappa)}} = \frac{2\kappa(1-r)}{\phi r(1-\kappa)+4\kappa(1-r)}.$$

Next, as $q_A = 1-q_B$ we can see, after some computation, that $q_A - q_B > \phi q_B$ iff $\kappa < \frac{r}{2-r}$.

So far, we have considered the case where individuals in A are always reluctant when meeting someone from B and vice versa. However, if $q_A - q_B < \phi q_B$, then a person from A will be reluctant when meeting someone from B only if the perceived position of the other happens to be $q_B - \phi q_B$ but not if it happens to be $q_B + \phi q_B$. Thus the probability of reluctance is $\frac{1}{2}$, and the equilibrium conditions becomes

$$(A3-7) \qquad \kappa\left(1-\frac{r}{2}\right)\delta(q_A - q_B) = r\delta(1-\kappa)\frac{\phi}{2}(1-q_A)$$

$$(A3-8) \qquad \kappa\left(1-\frac{r}{2}\right)\delta(q_A - q_B) = r\delta(1-\kappa)\frac{\phi}{2}q_B,$$

where the only difference from $(A3-2)$ and $(A3-3)$ above is the $\left(1-\frac{r}{2}\right)$ rather than $(1-r)$ on the left hand side. By the same reasoning as above, we cannot have full polarization. Hence part i) of the Proposition.

Solving as above, we get

$$(A3-9) \qquad q_B = \frac{2\kappa\left(1-\frac{r}{2}\right)}{\phi r(1-\kappa)+4\kappa\left(1-\frac{r}{2}\right)} = \frac{\kappa(2-r)}{r\phi(1-\kappa)+2\kappa(2-r)}.$$

This solution satisfies the requirement $q_A - q_B < \phi q_B$ iff $\kappa > \frac{r}{2}$.

As the thresholds for κ are overlapping, both the value of q_B specified in $(A3-6)$ and the value specified in $(A3-9)$ define an equilibrium if $\frac{r}{2} < \kappa < \frac{r}{2-r}$. Thus the second part of the Proposition.

Finally, to see that the solution is asymptotically stable, note first that in general, the solution to a linear system of the form $\dot{\boldsymbol{x}} = A\boldsymbol{x} + \boldsymbol{z}$, where A is a matrix and $\boldsymbol{z}$ a vector, will be of the form $\boldsymbol{x}(t) = \boldsymbol{x}^* + (\boldsymbol{x}(0) - \boldsymbol{x}^*)Pe^{\Lambda t}P^{-1}$ where $\boldsymbol{x}^*$ is the stable state, P is a matrix of eigenvectors and Λ is the diagonal matrix of eigenvalues. It follows that the stable state is asymptotically stable if all eigenvalues are negative. The system above can be written

$$\begin{bmatrix}\dot{q}_A \\ \dot{q}_B\end{bmatrix} = \begin{bmatrix}-(\kappa(1-\hat{r})\delta + r\delta(1-\kappa)\frac{\phi}{2}) & \kappa(1-\hat{r})\delta \\ \kappa(1-\hat{r})\delta & (\kappa(1-\hat{r})\delta + r\delta(1-\kappa)\frac{\phi}{2})\end{bmatrix}\begin{bmatrix}q_A \\ q_B\end{bmatrix} + \begin{bmatrix}r\delta(1-\kappa)\frac{\phi}{2} \\ 0\end{bmatrix},$$

where $\hat{r} = r$ or $\hat{r} = \frac{r}{2}$ depending on whether $q_A - q_B > \phi q_B$ or not. Here it can easily be verified that the matrix has eigenvalues $\lambda_1 = -r\delta(1-\kappa)\frac{\phi}{2} < 0$ and $\lambda_2 = -\left(2\kappa(1-\hat{r})\delta + r\delta(1-\kappa)\frac{\phi}{2}\right) < 0$. As both eigenvalues are negative, the system is asymptotically stable. ■

References

Akbiyik, A., J. Bowles, H. Larreguy, and S. Liu (2024): Polarization and Exposure to Counter-Attitudinal Media in a Nondemocracy. Working paper, presented at Democracy Under Threat: the Norms and Behavioral Change Conference 2024, Center for Social Norms & Behavioral Dynamics, University of Pennsylvania.

Akerlof, G.A., and R.E. Kranton (2000): Economics and Identity, *Quarterly Journal of Economics* 115 (3), 715–53.

Alesina, A., A. Miano, S. Stantcheva (2020): The polarization of reality, *AEA Papers and Proceedings* 110, 324-328.Algan, Y., N. Dalvit, Q.-A. Do, A. Le Chapelain, Y. Zenou (2023): Friendship Networks and Political Opinions: A Natural Experiment among Future French Politicians, CeSifo Working Paper 10753.

Axelrod, R., J.J. Daymude, and S. Forrest (2023): Preventing extreme polarization of political attitudes, *PNAS* 118 (50), e2102139118.

Babcock, Linda, Loewenstein, George (1997): Explaining bargaining impasse: the role of self-serving biases. *Journal of Economic Perspectives* 11 (1), 109–126.

Benabou, R., A. Falk, J. Tirole (2018): Narratives, Imperatives, and Moral Reasoning. NBER Working Paper 24798.

Benabou, R., and J. Tirole (2006): Incentives and prosocial behavior, *American Economic Review* 96 (5), 1652-1678.

Bénabou, R., and J. Tirole (2016): Mindful Economics: The Production, Consumption, and Value of Beliefs, *Journal of Economic Perspectives* 30 (3), 141–64.

Bonomi, G., N. Gennaioli, G. Tabellini (2021): Identity, Beliefs, and Political Conflict, *Quarterly Journal of Economics* 136 (4), 2371–2411.

Brady, D.W., and H.C. Han (2006): "Polarization Then and Now: A Historical Perspective" in: Nivola, P.S, and D.W. Brady, Eds.: *Red and Blue Nation? Characteristics and Causes of America's Polarized Politics*, Brookings Institution Press, 119-174.

Brehm, J.W. (1966): *A Theory of Psychological Reactance*, Oxford, UK: Academic Press.

Brekke, K.A., G. Kipperberg, and K. Nyborg (2010): Social Interaction in Responsibility Ascription: The Case of Household Recycling, *Land Economics* 86(4), 766-784.

Brekke, K. A., S. Kverndokk, and K. Nyborg (2003): An Economic Model of Moral Motivation, *Journal of Public Economics* 87 (9-10), 1967-1983.

Brekke, K. A., and K. Nyborg (2008): Attracting Responsible Employees: Green Production as Labor Market Screening, *Resource and Energy Economics* 39, 509-526.

Brekke, K.A., and K. Nyborg (2010): Selfish Bakers, Caring Nurses? A Model of Work Motivation, *Journal of Economic Behavior and Organization* 75, 377-394.

Brown, G.D.A., S. Lewandowsky, Z. Huang (2022): Social Sampling and Expressed Attitudes: Authenticity Preference and Social Extremeness Aversion Lead to Social Norm Effects and Polarization, *Psychological Review* 129 (1), 18–48.

Bursztyn, L., and R. Jensen (2017): Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure, *Annual Review of Economics* 9, 131-153.

Caprettini, B., M. Caesmann, H.-J. Voth, and D. Yanagizawa-Drott (2024): Going Viral: Protests and Polarization in 1932 Hamburg. CEPR Discussion Paper 16356 (updated version downloaded Feb. 21, 2025 from https://mcaesmann.github.io/research/hamburg/GoingViral_Feb2024.pdf).

Castle, J. (2019). New Fronts in the Culture Wars? Religion, Partisanship, and Polarization on Religious Liberty and Transgender Rights in the United States, *American Politics Research* 47(3), 650-679.

Corneo, G., and O. Jeanne (2009): A theory of tolerance, *Journal of Public Economics* 93 (5–6), 691-702.

Crocker, J., and Wolfe, C.T. (2001): Contingencies of self-worth, *Psychological Review* 108(3), 593–623. https://doi.org/10.1037/0033-295X.108.3.593.

Deffains, Bruno, Romain Espinosa, Christian Thöni, (2016), Political self-serving bias and redistribution, *Journal of Public Economics* 134, 67–74.

Desmet, K, I. Ortuno-Ortin, and R. Wacziarg (2025): Latent Polarization. Working paper, available at https://www.anderson.ucla.edu/faculty_pages/romain.wacziarg/downloads/2025_partitions.pdf (retrieved August 21, 2025).

Ellingsen, T., and M. Johannesson (2011): Conspicuous Generosity, *Journal of Public Economics* 95 (9-10), 1131-1143.

Falk, A. (2021): Facing yourself – A note on self-image, *Journal of Economic Behavior & Organization* 186, 724-734.

Gallup. (2024, October 16). Same-sex relations, marriage still supported by most in U.S. Retrieved from https://news.gallup.com/poll/646202/sex-relations-marriage-supported.aspx.

Hadler, M., and J. Symons (2018): World Society Divided: Divergent Trends in State Responses to Sexual Minorities and Their Reflection in Public Attitudes, *Social Forces* 96 (4), 1721–1756, https://doi.org/10.1093/sf/soy019.

Hart, William, Dolores Albarracín, Alice H. Eagly, Inge Brechan, Matthew J. Lindberg, Lisa Merrill (2009): Feeling validated versus being correct: A meta-analysis of selective exposure to information, *Psychological Bulletin*, 135 (4), 555–588.

Hetherington, M.J. (2209): Putting Polarization in Perspective, *British Journal of Political Science* 39(2), 413-448, doi:10.1017/S0007123408000501.

Hobolt, S.B., T.J. Leeper, J. Tilley (2021): Divided by the Vote: Affective Polarization in the Wake of the Brexit Referendum. *British Journal of Political Science* 51(4), 1476-1493, doi:10.1017/S0007123420000125.

Holst, J.J. (1975): Norway's EEC Referendum: Lessons and Implications, *World Today* 31, 3, 114-120.

Iyengar, S., Y. Lelkes, M. Levendusky, N. Malhotra, S.J. Westwood (2019): The Origins and Consequences of Affective Polarization in the United States, *Annual Review of Political Science* 22, 129-146.

Lee, F. (2015): How Party Polarization Affects Governance, *Annual Review of Political Science* 18, https://doi.org/10.1146/annurev-polisci-072012-113747.

Michelson, M. R., and E. Schmitt (2020): Party Politics and LGBT Issues in the United States. *Oxford Research Encyclopedia of Politics*. https://doi.org/10.1093/acrefore/9780190228637.013.1208

Mutz, D. (2024): The Implosion of the Public Sphere, keynote presented at Democracy Under Threat: the Norms and Behavioral Change Conference 2024, Center for Social Norms & Behavioral Dynamics, University of Pennsylvania.

Nyborg, K. (2011): I Don't Want to Hear About it: Rational Ignorance among Duty-Oriented Consumers, *Journal of Economic Behavior and Organization* 79, 263-274.

Nyborg, K., R. B. Howarth, and K. A. Brekke (2006): Green Consumers and Public Policy: On Socially Contingent Moral Motivation, *Resource and Energy Economics* 28 (4), 351-366.

Pew Research Center. (2023, March 15): Americans feel more positive than negative about Jews, mainline Protestants, Catholics, https://www.pewresearch.org/religion/2023/03/15/americans-feel-more-positive-than-negative-about-jews-mainline-protestants-catholics/pf_2023-03-15_religion-favorability_00-010-png/.

Pew Research Center (2024, April 9): Party identification among religious groups and religiously unaffiliated voters. https://www.pewresearch.org/politics/2024/04/09/party-identification-among-religious-groups-and-religiously-unaffiliated-voters/.

Pyszczynski, T., Greenberg, J., Solomon, S., Arndt, J., and Schimel, J. (2004): Why Do People Need Self-Esteem? A Theoretical and Empirical Review, *Psychological Bulletin* 130 (3), 435–468. https://doi.org/10.1037/0033-2909.130.3.435.

Rosenberg, B.D., and J.T. Siegel (2018): A 50-Year Review of Psychological Reactance Theory: Do Not Read This Article, *Motivation Science* 4 (4), 281-300.

Sambanis, N., and M. Shayo (2013): Social Identification and Ethnic Conflict, *American Political Science Review* 107 (2), 294-325.

Santos-Pinto, L., and J. Sobel (2005): A model of positive self-image in subjective assessments, American Economic Review 95 (5), 1386-1402.

Schelling, T. C. (1978). *Micromotives and Macrobehavior*. Norton.Shayo, M. (2009): A Model of Social Identity with an Application to Political Economy: Nation, Class, and Redistribution, *American Political Science Review* 103(2), 147-174.

Törnberg, P. (2022): How digital media drive affective polarization through partisan sorting, *PNAS* 119 (42), https://doi.org/10.1073/pnas.2207159119.

Törnberg, P., C. Andersson, K. Lindgren, S. Banisch (2021): Modeling the emergence of affective polarization in the social media society. *PLoS ONE* 16 (10): e0258259. https://doi.org/10.1371/journal.pone.0258259.

Weibull, J.W. (1995): *Evolutionary Game Theory*. Cambridge, MA: MIT Press.