IMAGE PREFERENCES AS A DRIVER OF POLARIZATION

Abstract:

We show how the preference to be highly regarded by others as well as oneself can drive polarization. Changed normative views may affect individuals' image, depending on their characteristics: for example, while single mothers or LGBTQ people may be viewed more favorably by liberals than conservatives, the reverse may hold for the wealthy or preachers. While normative views are fixed convictions in the short run, we assume that they are adopted from peers over time; adoption is more reluctant, however, for views decreasing one's utility. Over time, migration between peer groups is feasible. In the steady state, everyone in a given peer group shares an extreme normative view, which one depending on group members' characteristics. The only exception is groups where no-one's image is affected by changed normative views: if such groups exist, their members may share intermediate positions. We show that if views are partly learnt across and not just within peer groups, equilibrium polarization is less extreme. If image can be improved through effort, the steady state involves minimal optimal efforts.

Keywords: Polarization; segregation; reluctant social learning; image preferences.

JEL Codes: D11; D31; D63; D64; D91.

1. Introduction

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 a positive self-image 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 discusses a more troublesome aspect: under plausible conditions, we find that the preference for social image and self-image can drive society towards extreme segregation and polarization.

In political and/or ethical debate, normative views sometimes cluster around the extremes rather than the middle ground (political polarization); moreover, proponents of each side may have low esteem for their opponents, avoiding social contact with them (affective polarization, Iyengar et al. 2019). 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).

This paper shows how the desire to be highly regarded by oneself as well as others can drive such social divisions. The idea that people care about their image has by now become widely explored in the economics literature (e.g., Akerlof and Kranton, 2000; Brekke et al. 2003; Santos-Pinto and Sobel 2005; Benabou and Tirole 2006, 2016; Shayo 2009; Ellingsen and Johannesson 2011; Bursztyn and Jensen 2017; Benabou et al. 2018; Bonomi et al. 2021). Below, we use the observation that changing political or ethical views can affect individuals' image, depending on their characteristics: for example, single mothers or LGBTQ people may be viewed more favorably by liberals than by conservatives, while the reverse may hold for preachers or wealthy people. As a result, social

dynamics may arise helping individuals protect their image, while causing segregation and polarization as an externality.

Our assumptions are fairly general and, we believe, plausible. Short-run utility depends on self-image and social image, which may vary with one's characteristics and the normative views of the person making the judgement. Later, we also allow individuals to improve their image by means of costly effort. Over time, two social mechanisms interact: adoption of normative views from peers, and migration between peer groups.

Normative views are considered fixed convictions in the short run but are adopted from peers over time (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 adoption is taken to be *reluctant* (Brekke et al. 2010): one is somewhat less likely to adopt views decreasing one's utility (by reducing one's self-image). Furthermore, since others' views cannot be observed with precision, individuals make errors when judging others' views. Reluctance implies that errors benefiting the individual are given disproportionally large weight, allowing normative views to move beyond their initial range. Finally, people are assumed to occasionally reconsider which social peer group to be part of, preferring peers regarding them more favorably (increasing their social image).

The crucial idea causing the above assumptions to drive polarization is the following. We assume that normative views can be sorted along a continuous one-dimensional scale ranging from 0 to 1 (e.g., left to right; liberal to conservative; support of democracy to support of totalitarianism). If peers' normative views approach one extreme, say 1, then we assume that all else given, social image is *reduced* for individuals with some sets of characteristics, henceforth called the L type, whereas social image *increases* for individuals with other sets of characteristics, called the R type. Individuals whose social image is unaffected by changes in peers' normative views are termed the O type. Similarly, if one's *own* normative view moves towards 1, self-image is reduced for L types, is improved for R types, and is unaffected for O types.

We place no restrictions on individuals' initial normative views and peer group affiliations. Nevertheless, given the above assumptions, L and R types gradually self-select into different peer groups, while normative views converge within peer groups but diverge between peer groups. The steady state is extremely polarized and segregated: in equilibrium, no peer group contains both L and R types; everyone in a peer group with L types agrees on the view corresponding to 0 on the normative scale, while everyone in a peer group with R types agree on the view corresponding to 1. Only peer groups consisting exclusively of O types, if such groups exist, may hold intermediate equilibrium views.

In our model, we abstract from possible macro level effects of increased polarization such as political unrest and instability; more bullying and violence; less general trust; less efficient intergroup collaboration; or lower willingness to contribute to public goods. Furthermore, people are implicitly assumed to care only about their peers' regard, not the regard of their opponents. Given these essential caveats, the steady state maximizes each individual's utility: each holds the normative conviction that maximizes her self-image, and is surrounded by peers agreeing to the view maximizing her social image. Similarly, we show that if costly effort can be used to improve one's image, optimal efforts are minimized in the steady state. Thus, in spite of its substantial negative effects not modelled here, polarization may serve a social purpose, providing one reason why it emerges - namely to let people feel highly regarded by themselves as well as their peers.

How can extreme polarization be prevented or reversed? Within our formal framework, it is surprisingly difficult to come up with good answers. Our key result on this, however, is that equilibrium polarization is reduced if normative views are partly adopted across peer groups, not only within peer groups. Hence, starting from an initial situation with substantial inter-group learning due to widespread consumption of cross-cutting news and opinion media, such as local newspapers and national TV programmes, polarization should be expected to increase if technological developments make people's media exposure more fragmented, replacing their consumption of cross-cutting media by like-minded media (Mutz 2024). Conversely, starting from a polarized state, policies incentivizing people to learn from opponent media may reduce polarization (Akbiyik et al., 2024).

The main novelty of our approach is to show how dynamic social mechanisms helping each individual feel highly regarded by herself as well as others can cause strong polarization and segregation as a by-product. We are not aware of other explorations of such mechanisms in previous literature. 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, while underlying attitudes as such are kept fixed. Schelling's (1978) segregation model does not involve changing attitudes, thus predicting segregation but not polarization. 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. In Axelrod et al.'s (2023) agent-based analysis of ideological polarization, changes in normative views over time occur because interaction between agents with similar views mechanically reduces the difference between them, while interaction between dissimilar actors mechanically increases their difference.

Unlike us, Bonomi et al. (2021) focus on the analysis of two-dimensional disagreement. They find that social identification causes increased but not extreme political disagreement compared to the case without social identification. In their model, social identification distorts individuals' factual beliefs about the world, thus invoking an element of limited rationality. The adoption of peers' normative views in our model is not based on limited rationality, even with reluctance, since normative views cannot be established on purely factual or logical grounds but must necessarily involve subjective judgement. Note, however, that if normative disagreement is based on different factual beliefs, and reluctant social learning occurs not only for normative but even for factual beliefs, the mechanism we describe can also help explain the observation that people's factual beliefs are correlated with their political views (Alesina et al., 2020).[1]

[1] For example, climate deniers may approve of fossil fuel intensive activities while finding green activities foolish, while those concerned about man-made climate change may approve of green activities while finding fossil fuel intensive activities immoral. Those heavily pre-invested in fossil fuels will then tend to prefer climate denying peers, who will view them more favorably, while the opposite holds for those with green pre-investments. If *factual* beliefs are adopted reluctantly from peers, those with green pre-investments will be more eager to adopt climate concerned beliefs while the reverse is true for those invested in fossil fuels. Hence the

Below, we present our formal model in more detail. We begin by presenting individuals' short-run utility function, considering normative views and social group affiliation fixed. We then turn to the dynamic mechanisms of social learning and migration, respectively, before merging all parts into an integrated dynamic model. After this, we demonstrate that policies stimulating learning across social groups can modify equilibrium polarization. Finally, we introduce costly effort to improve one's image, before concluding.

2. Image preferences

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

Let each individual i believe in some normative principle indexed by $q_i \in [0,1]$, serving as the basis of i's judgements of others as well as themselves (Figure 1). This may represent, for example, the scale from left-wing to right-wing views, from egalitarianism to libertarianism, from trust to mistrust in scientific reasoning, from traditional to liberal values, or from believing that all humans have equal intrinsic value to finding individuals of certain characteristics worthless.

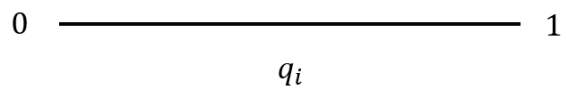

Figure 1: Possible values for the index q_i representing normative views.

equilibrium will be polarized (McCright and Dunlap 2011; Falkenberg et al., 2022): green pre-investors will socialize with each other, entertaining climate concerned beliefs, while fossil pre-investors similarly socialize with each other, holding climate denier beliefs.

Each individual is assumed to belong 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 normative view q_i for members of peer group G.[2]

Each individual i has an exogenously determined vector of characteristics θ_i. This vector could for example include social class, education and other human capital, financial wealth, ethnicity, personality traits, and religion. 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, which we will call L, R, and O. Hence, for every i, either $\theta_i \in L$, $\theta_i \in R$, or $\theta_i \in O$.[3] As specified more precisely in Assumption 1 below, type L can intuitively be thought of as *those regarded more highly by the Left*, type R as *those regarded more highly by the Right,* while type O are *those whose regard is independent of the evaluator's normative view.* Note, however, that the term "Left" here refers to low values of q_i while "Right" refers to high values of q_i, corresponding to the visual presentation in Figure 1; this does not, of course, need to correspond to the traditional left-wing versus right-wing political scale, although that is of course one possible application.

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

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

Note that it is the *evaluator's* normative view that enters the image functions, not the views of the person *being evaluated*. My social image is judged by my peers, so it is their views that matter for how

[2] The logical but cumbersome notation would be q_{G_i}, where G_i is i's peer group; we use the simpler notation hoping there will be no confusion.

[3] In some examples the set of type O may be empty. For example, the set L may consist of everyone without a college/university degree while R are those with a college/university degree; or L may be one ethnic group while R is another.

highly they regard me. The object of their evaluation is my characteristics θ_i, not my normative position q_i (the latter would be more similar to the approach of Axelrod et al. (2023).

In the short run, individuals have no choices to make; they must simply accept their own and others' judgements. If, for example, a person is gay but holds a very conservative normative view, he may regard himself less favorably than he would if his views had been more liberal. Further below we modify this, allowing individuals to exert effort to improve their image; however, since the introduction of effort does not matter substantially for the dynamics, we keep things simple for now.

Our results will depend crucially on the following assumption, namely that a change in normative views affects 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 $\theta_i \in O$.

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 ethical views

Let us now turn to the dynamics, considering first how ethical views are affected by one's peers over time, which we may think of as social learning. For the moment, we keep groups fixed; migration will be added to the picture in the next section.

We regard normative views represented by q_i as convictions that cannot simply be chosen. However, since normative value judgements are inherently subjective and cannot be deduced from facts and logic alone, it seems reasonable to assume that such convictions 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.

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 normative 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), since j is part of i's own social group. 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

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

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 ethical view of the group stays unchanged:

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

[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.

Note, however, 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, normative 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 normative view can change beyond their initial values over time, moving all the way to the extremes. The idea here is that although individuals gradually learn their normative view from others, they have a slight reluctance to adopt views that would be to their disadvantage. Due to this reluctance, errors in the perception of others' normative views do not cancel out over time. Let us again begin with discrete time.

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 normative 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 normative view q_i^t a fraction δ in the direction of $\tilde{q}_{ji}^t$ when doing so increases U_i^t, but adjusts her normative view q_i^t only a fraction $\delta(1-r)$ in the direction of $\tilde{q}_{ji}^t$ otherwise.

Below, we will assume that social learning is reluctant.

Recall that due to Assumption 1, utility is *decreasing* in q_i^t for individuals of type L but *increasing* in q_i^t for type R. For example, someone who is gay may be slightly reluctant to accept a more conservative view than the one he already holds, while a preacher may be slightly reluctant to accept a more liberal view. The assumption that people have self-image preferences is crucial for this: one's self-image is determined by exogenous characteristics and one's own normative view, meaning that if adopting views in the "wrong" direction, utility decreases. Individuals of type O will not be reluctant to move in any direction, as their self-image is unaffected.

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 rather reluctant to adopt a *decrease* in q_i^t. Hence, if $\theta_i \in R$, we have

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

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

And 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

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

Let us 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 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 $i \in L$.

[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.

For simplicity we first consider the case where the set of type O is empty. Now, let us again move to continuous time. 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. The dynamic of R types' views is thus

(6) $$\dot{q}_R = (1 - s_G)\delta E[(q_j - q_i)|\theta_i \in R \text{ and } \theta_j \in L] + s_G r\delta B_{RR}^{-} + (1 - s_G) r\delta B_{RL}^{-},$$

where $\dot{q}_R$ is the change in the average normative view for $\theta_i \in R$. Similarly, for $\theta_i \in L$ we have

(7) $$\dot{q}_L = s_G\delta E[(q_j - q_i)|\theta_i \in L \text{ and } \theta_j \in R] - s_G r\delta B_{PR}^{+} - (1 - s_G) r\delta B_{PP}^{+}.$$

The dynamic for the group's average view $\dot{q}_G$ is the weighted average of the two, which gives

(8) $$\dot{q}_G = r\delta[s_G^2 B_{RR}^{-} + (1 - s_G)s_G(B_{RP}^{-} - B_{PR}^{+}) - (1 - s_G)^2 B_{PP}^{+}] \equiv \Pi_G.$$

Recall that with unbiased learning, we would have $\dot{q}_G = 0$. Hence, Π_G, as defined by eq. (8), can be interpreted as a measure of the total effect of reluctance in social group G. Note that while players of type O are not reluctant, types L and R are reluctant when meeting an O-player and perceiving their view as pulling in the "wrong" direction. The resulting reluctance will be denoted B_{RO}^{-} and B_{PO}^{+}, and defined in the similar fashion as above.

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

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 briefly below and more thoroughly in Appendix 1.

We must now allow the set of O types to be non-empty. Let $(1 - \alpha_G)$ denote the share of the O type in group G, and let us now interpret s_G as the share of the R type among the remaining share α_G. For example, if $s_G = 0.5$ there are equally many of type R as of type L in group G.

We can now state our Proposition 1, which will be an important building block towards our main polarization and segregation result when later allowing for social migration: If there is no migration, and the strength of individuals' learning reluctance is limited, then if the majority of non-O's in the social group consists of R types, everyone in the group holds the extreme view $q_i = 1$ in the only asymptotically stable steady state. Conversely, if the majority of non-O's in the social group consists of L types, 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 depend on whether there are more

L than R types. This drives the group average q_G towards zero if the majority on 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 social groups. In contrast to ethical view updating, changing one's social peer group is a choice, not an inference. Nevertheless, inspired by evolutionary game theory (Weibull 1995), we assume that individuals revise their social group affiliation only every now and then, and that they do so myopically, not taking into account that migration might affect their own future normative view. [7]

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

Note that 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, and then later demonstrate that the main results are still valid when there are O-types too.

Assume now that there are only two equally large social neighborhoods, 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:

[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 may 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.

knowing the number of L types in social group A, the number of R and L in each social group follows. Also, let individuals disregard the potential effect of their own migration on q_G in either social group.

In the current framework, the only reason why social group affiliation matters to individuals is their preference for social image: they want to be regarded highly by their peers. When revising which peer group they want to be part of, L types prefer the social group with lower q_G, while R types prefer the neighborhood with higher 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 social group can now vary over time. 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, and moving again to continuous time by shortening period length towards zero, 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 ethical views 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 social neighborhood. Eq. (8) describes the dynamics of ethical views caused by reluctant social learning in fixed groups, but without taking into account the direct effect of migration on the average normative view in each neighborhood.

Writing eq. (8) separately for neighborhoods A and B (still for the moment ignoring the short-run changes in q_A and q_B as a direct result of migration), using the measure Π_G of the total effect of reluctance in social group G defined in that equation, we have:

(10) $\dot{q_A} = \Pi_A$

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

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 normative views in the two neighborhoods 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 (since, for example, the wealthy attracted to A are reluctant to adopt more liberal/leftist 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 $s_A = 1$: Groups would be perfectly segregated according to their type (e.g., income); 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 $(q_{LB} - q_{RA})\dot{s}_A$, where $q_{\theta G}$ denotes the average q_i among type $\theta = L, R$ in neighborhood $G = A, B$, must be added to expression (10), similarly for eq. (11). A more detailed explanation of why is provided in Appendix 2.

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) $\dot{q}_A = \Pi_A - \rho(1 - s_A)(q_A - q_B)(q_{LB} - q_{RA})$.

Similarly, for the L dominated group, B,

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

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 $\dot{s}_B$. Note further that since close to the steady state, $1 - s_A \approx 0$ and $s_B \approx 0$, and in the long run, as $t \to \infty$, $s_A \to 1$ and $s_B \to 0$ by eq. (9), the migration terms will eventually be negligible and hence not affect the asymptotic stability of the equilibrium.

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 normative 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 group is influenced by two factors, reluctance and migration. In the steady state, migration is by definition zero. Hence, the only possible 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 they have no reason for reluctance (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:

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.

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 polarization between groups with L members and groups with R members.

The highly polarized and segregated steady states described in Proposition 2 display a striking feature: No other combinations of normative views and sorting into social 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-image is maximized if i's normative conviction corresponds to $q_i = 0$, while her social image 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-image is maximized if i's normative conviction corresponds to $q_i = 1$, while his social image 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 normative views and sorting into groups can improve their utility.

Our model abstracts from possibly crucial macro level consequences such as more mistrust, instability, political unrest and possibly violence, less efficient cooperation, for example in team production, between members from different groups. Further, in contrast to our assumptions, people might care about social approval even from non-peers. We would thus by no means conclude that extreme polarization and segregation are generally welfare maximizing phenomena. Nevertheless, this finding illustrates one possible reason why polarization and segregation do occur: in societies where normative disagreement makes people hold different judgements of the worthiness of various individual characteristics, segregation and polarization help us escape negative judgements, allowing us to feel highly valued both by ourselves and others.

In equilibrium, there is not only political (or more generally, normative) polarization but also affective polarization. First, people are not interacting socially with their opponents. Moreover, opponents have low regard for each other, even if i's social image does not depend her own normative views in our framework: in equilibrium, R types hold the view that would give an L type the lowest possible social image if moving to a social group with R's, and Ls similarly hold the view which would give an R type the lowest possible social image if moving to a social group with L's.

Note that the presence of O type limits polarization somewhat. First, if there are groups with only O's, the average view q_G in such groups will stay constant over time, limiting political (normative) 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.

6. Limiting equilibrium polarization

A key driver of the polarization result above is the assumption that individuals learn their normative views from peers in their own social group. Here, we show that if some learning of normative views takes place between peer groups, steady state polarization is less extreme. Hence, policies stimulating learning across groups – for example, making kids from diverse neighborhoods attend the same schools, encouraging attendance in shared cultural experiences, or stimulating diverse groups' joint participation and encounters in public debate – could help reduce or prevent polarization (Benabou et al. 2018).

As a point of departure, consider the equilibrium described in Proposition 2 with complete polarization and segregation. Assume now that some exogenous change – a policy stimulating contact between diverse social groups, or technological or institutional changes facilitating cross-cutting social learning – is introduced in this situation, causing social learning of normative views to partly take place between social neighborhoods.

While this will not affect migration directly, it will change the dynamics of social learning. Recall that in eq. (5), we modelled the change over time in an individual's ethical view q_i^t as the sum of two parts: the unbiased learning from meeting a random group member j, plus a term reflecting reluctance. Now, assume instead that with probability κ, the other individual j is from the other social group ($G_i^t \neq G_j^t$, keeping the assumption of two groups for simplicity), while with probability $(1-\kappa)$ the other is from one's own neighborhood ($G_i^t = G_j^t$).

Let A be the social group with R types and possibly O types. Then, with continuous time, incorporating reluctance, and adding the direct effect of migration on average ethical views in the group (see eqs. (12) - (13) and the discussion thereof), the equilibrium condition now becomes

$$\dot{q}_A = \delta\kappa(q_B - q_A) + \Pi_A = 0. \quad (14)$$

In equilibrium, with A as the R type group, $s_A = 1$. Thus, in equilibrium

$$\kappa\delta(q_A - q_B) = r\Pi_A\,. \quad (15)$$

This rules out $q_A = 1$ and $q_B = 0$, since if $q_A = 1$ we must have $\Pi_A = 0$. Consequently, we no longer get complete polarization. The overall effect of reluctance approaches zero as $q_A \to 1$; hence, at some point before we get to $q_A = 1$, the effect of meeting people in the other group will cancel out the effect of reluctance. As a result, equilibrium polarization is limited.[8] We summarize this as a Proposition.

Proposition 3. If social learning takes place partly between groups, polarization is incomplete in equilibrium: $q_A - q_B < 1$.

This result can of course also be interpreted in terms of a reverse movement: Assume that the economy is initially in the *in*complete polarization equilibrium described by Proposition 3. Then, if an exogenous shock occurs decreasing the share of learning taking place across groups, polarization will increase. The introduction of the internet and the reduced consumption of cross-cutting media that

[8] On the other hand, it is not sufficient to rule out *any* polarization, since if $q_A = q_B = \frac{1}{2}$ then $r\Pi_A > 0$ for $s_A > \frac{1}{2}$.

followed (Mutz 2024), as well as social media algorithms favoring the display of like-minded views, may possibly represent such exogenous shocks.

7. Minimal efforts

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 allow individuals to exert costly effort to improve their self-image and/or social image. Let $e_i = e_i^I + e_i^S$ denote the effort i exerts to improve her image, where e_i^I is effort to improve self-image and e_i^S is effort to improve her social image, and let us rewrite eq. (1) as follows:

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

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. ■

[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 slightly more cumbersome, however, we keep to the simplest approach here.

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 normative views in the same way as before. By adding more structure to how effort affects image, however, we can derive further conclusions.

Since O types' image is unaffected by normative views, it is natural to assume that their effort is independent of normative 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 not only represents the strongest possible segregation and 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: counteracting mechanisms

Above, we demonstrated that with some learning across social groups, equilibrium polarization is limited. Within our formal framework, however, it is surprisingly hard to come up with other reasonably simple mechanisms counteracting polarization. To see why, Proposition 1 is key: This result shows that given limited reluctance, polarization arises even without migration – thus even in the absence of segregation.

Introducing an individual migration cost, for example, would make segregation between L and R types incomplete, but would not necessarily limit equilibrium polarization – although the effects would depend on the image functions and the distribution of θ_i. To see this, let the polarized equilibrium described in Proposition 2 be the starting point, assuming only two social groups, with group A being the one with R types. Assume now that the social image benefit of belonging to group A rather than B varies among Rs, being small for some but increasing throughout the population of Rs. Assume that the social image benefit of belonging to group B rather than A varies similarly among Ls. Then, introducing a cost of moving would just make some marginal group members – those who are close to indifferent between peer groups – abstain from moving, while everything else would go through as before (Nyborg and Brekke 2024).

One may also expect that trading benefits would reduce equilibrium segregation and thus reduce polarization: if there is 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 less than maximal social image in order to seek such profit. However, Proposition 1 indicates that over time, such individuals would adopt the normative view held by the majority of their group; thus, although segregation according to type would be less than complete, extreme polarization would remain.

Nevertheless, it may still be the case that economic or other benefits of cross-cutting interaction could play an important role to limit polarization. Proposition 3 says that *learning* across social groups, not contact per se, limits equilibrium polarization. Starting from a strongly polarized state, it may be hard to convince people to listen to and thus potentially learn from opponents, even if exposed to their views. However, economic or other benefits associated with *understanding opponents' positions* would provide an incentive to listen.

9. Conclusions

Above, we have demonstrated how image preferences can push society towards social segregation and political/normative as well as affective polarization. In equilibrium, normative views are extreme rather than moderate; opponents avoid each other's company, and have low regard for each other.

We have assumed that one's image, whether self-image or social image, depends on one's exogenous individual characteristics. The evaluation of these characteristics, however, may depend on the evaluator's normative views. For example, an LGBTQ person or a single mother may be more highly regarded by their peers if their peers are more liberal, while a preacher or a wealthy person may be more highly regarded by conservative peers. Over time, thus, individuals migrate to social peer groups appreciating their characteristics. Within peer groups, people gradually adopt each other's normative views; moreover, since we assume a social learning process giving disproportionally low weight to perception errors threatening one's self-image, normative differences between peer groups become extreme over time. As a result, the steady state is characterized by deep social division: individuals have self-selected into separate social groups according to their characteristics, where groups with different types of individuals hold extremely different normative views.

We also show, however, that if some learning of normative views takes place across social groups rather than only within groups, equilibrium polarization is less extreme. Starting from a such equilibrium with limited polarization, exogenous changes reducing cross-cutting learning should be expected to increase polarization. The increased fragmentation of information flows after the

establishment of the internet could represent such a change, for example through reduced consumption of cross-cutting news and entertainment media (Mutz 2024); social media and search algorithms favoring display of like-minded views; and local stores, where neighbors would meet and chat, being replaced by online trade. A key to reduce equilibrium polarization, then, would be to find ways to incentivize not only contact but learning across polarized groups (Akbiyik et al. 2024).

Author affiliations:

Department of Economics, University of Oslo (both authors).

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: Normative 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 normative 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),

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

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 P 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

$$sB_{RO}^- - (1-s)B_{LO}^- = \frac{s}{2}(q_R - q_O - \phi d_O) + \frac{1-s}{2}(q_O - q_P + \phi d_O)$$

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

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

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

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

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

$$= \frac{\phi r}{4}\big(s^2 q_R - (1-s)^2 q_L - s(1-s)(q_R - q_L)\big) + \frac{s(1-s)r}{4\phi q_R q_L}\Delta^3$$

$$= \frac{\phi r}{4}(2s-1)(sq_R + (1-s)q_L) + \frac{s(1-s)r}{4\phi q_R q_L}\Delta^3 \qquad \blacksquare$$

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

$$(A3-8) \qquad \dot{q}_R - \dot{q}_L = -(q_R - q_L) + sr(B_{RR}^{-} + B_{LR}^{+}) + (1-s)r(B_{RL}^{-} + B_{LL}^{+})$$

$$\leq -\Delta + sr\left(\frac{\phi}{4} d_R + \frac{3+K}{4}\Delta\right) + (1-s)r\left(\frac{3+K}{4}\Delta + \frac{\phi}{4} d_L\right)$$

$$\leq -\Delta\left(1 - \frac{2+2K+2\phi}{4} r\right) + \frac{r\phi}{4}\bar{d} \leq -\Delta\left(1 - \frac{2+3K+3\phi}{4} r\right)$$

$$< 0 \; for \; r < \frac{4}{5+2K} \qquad \blacksquare$$

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 $f(\Phi) = f(0) + f(\Phi')\Phi$ 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: On asymptotically stable states in the case with migration

Equations (10) and (11) represent a continuous-time version of the discrete-time eq. (8) (separately for groups A and B), reflecting the change in ethical views due to social and possibly reluctant learning. In the total dynamics, we must also take into account the change in average ethical views in each group (q_A and q_B) caused by migration between the two groups. This is most easily seen starting, again, from a discrete time formulation.

Assume now that the timing in each period t is as follows: first, at point in time t', individuals determine their contributions, taking ethical views and group affiliation as fixed; then, at t'', ethical views are updated; and finally, at t''', group affiliation is updated. We know already that the expected change in q_A^t between t' and t'' due to ethical updating equals $(2s_{RA}^t - 1)vr$ (eq. (12)). What we are missing is an expression for the change from t'' to t''', reflecting migration between A and B.

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

At t'', after the period's ethical updating has taken place but before migration, the average normative view in A can be written as

$$q_A^{t''} = s_{RA}^{t-1} q_{RA}^{t''} + (1 - s_{RA}^{t-1}) q_{LA}^{t''}$$

where $q_{\theta G}^t$ is the average normative view among income group θ in neighborhood G at t.

If $q_A^t = q_B^t$, there is no incentive to move, so no migration takes place. The interesting case is when the average normative view differs between A and B. Assume that $q_A^t > q_B^t$. The Ls in A who revise their neighborhood affiliation, i.e., $\rho(1 - s_{RA}^{t-1})$, will now move to B; Rs in B who revise, i.e., $\rho(1 - s_{RA}^{t-1})$, will move to A. (Recall that $s_{RA}^{t-1} = 1 - s_{RB}^{t-1} = 1 - s_{LA}^{t-1}$, thus $s_{RB}^{t-1} = s_{PA}^{t-1}$.) Since the normative view updating has already been done, the remaining individuals change neither their views nor their social group affiliation between t'' and t'''. Thus, the change in average normative views in A between t'' and t''', entirely due to the direct effect of migration, is

$$q_A^{t'''} - q_A^{t''} = \rho(1 - s_{RA}^{t-1})[q_{RB}^{t''} - q_{LA}^{t''}].$$

Similarly, by symmetry, the change in average normative views due to migration in B is

$$q_B^{t'''} - q_B^{t''} = \rho(1 - s_{RA}^{t-1})[q_{LA}^{t''} - q_{RB}^{t''}].$$

Stating this as differential equations and adding the relevant expressions to the direct effect of normative updating as specified in eqs. (10) and (11), disregarding now the within-period timing of updating decisions, we get the following adjusted equations for the change in moral views when both ethical updating and the short-run effect of migration are taken into account:

$\dot{q_A^t} = (2s_{RA} - 1) + vr - (q_{LA}^t - q_{RB}^t)\dot{s}_A$, and

$\dot{q_B^t} = (1 - 2s_{RA})vr + (q_{LA}^t - q_{RB}^t)\dot{s}_A$.

We are now equipped to prove Proposition 2, restating it here for convenience:

Proposition 2

Given Assumption 2 and reluctant learning, there are only two asymptotically stable states. In both

states, all $i \in R$ are located in one social group and hold an ethical view $q_i = q_R = 1$, while all $i \in L$ are located in the other social group, holding a view $q_i = q_L = 0$. Since a given neighborhood can either be the one with the R types or the one with the L types, there are two asymptotically stable states.

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. 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 neighborhoods 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 more egalitarian 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 Rs and Ls and with several groups of endogenous of 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 normative 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. ■

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.

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.

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

Enke, B. (2019): Kinship, Cooperation, and the Evolution of Moral Systems, *Quarterly Journal of Economics* 134(2), 953-1019.

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

Falkenberg, M., A. Galeazzi, M. Torricelli, N. Di Marco, F. Larosa, M. Sas, A. Mekacher, W. Pearce, F. Zollo, W. Quattrociocchi and A. Baronchelli (2022): Growing polarization around climate change on social media, *Nature Climate Change* 12, 1114–1121.

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): *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.

Hvidberg, K.B., C.T. Kreiner, S. Stantcheva (2023): Social Positions and Fairness Views on Inequality, *Review of Economic Studies* 90 (6), 3083–3118.

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.

McCright, A.M., R.E. Dunlap (2011): The politicization of climate change and polarization in the American's public view of global warming, 2001-2010, *Sociological Quarterly* 52, 155–194.

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.

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.

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