Moral responsibility as a driver of polarization

Karine Nyborg and Kjell Arne Brekke

Abstract:

In societies relying on voluntary public good provision, individuals may experience their responsibility to contribute as a moral burden. Considering a morally motivated population with exogenous income inequality, we explore social processes driven by this moral burden. Ethical views are assumed to be socially learnt, but reluctantly so: one is less prone to adopt views implying higher fair contributions for oneself. Egalitarian peers, demanding high contributions from the rich, are avoided by the rich, but sought by the poor. The resulting long-run equilibrium is extremely segregated and polarized, with minimal voluntary contributions. Public provision, however, limits the polarization.

Keywords: Private public good provision; endogenous ethical views; self-image; social image; segregation; inequality.

JEL Codes: D11; D31; D63; D64; 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 to Bård Harstad, Frédéric Deroïan, Maria Bigoni, Joël van der Weele, and seminar and conference participants at the University of Amsterdam, the University of Oslo, the Berlin Environmental Economics Seminar, LAGV 2023, European ESA 2023, and the Annual Meeting of the Norwegian Economic Association for comments and helpful suggestions.

1. Introduction

While narrowly self-interested individuals tend to free-ride on others' public good provision (Bergstrom et al., 1986), previous research has demonstrated that image concerns and moral considerations can give rise to more cooperative behavior (Brekke et al. 2003; Benabou and Tirole 2006a; Nyborg et al., 2006; Brekke and Nyborg 2008, 2010; Benabou et al., 2018). Perceived moral responsibility can be a burden, however (Nyborg 2011): Doing what you find morally best may involve costs, e.g., in terms of voluntary donations to some good cause, while deviating from your moral ideal may involve cognitive dissonance (Festinger 1957). Similarly, others' perception of your moral responsibility can also represent a burden: if your peers think you should contribute a lot, you must either do so or risk facing their disapproval.

Judging what is morally appropriate, however, is a subjective matter on which people disagree. As thoroughly demonstrated by psychological research (see, e.g., Aronson et al. 2005, p.166-188), cognitive dissonance can be reduced not only by changing one's behavior, but also by adjusting the

ideals one strives towards. In the same vein, peer pressure can be reduced by replacing one's peers by less demanding ones.

Here, we explore the long-run effects of social interaction on perceived moral responsibility and contribution behavior. In our framework, incomes differ exogenously. Ethical views may initially differ between individuals but develop endogenously over time. We assume that individuals have identical image preferences, comparing their actual behavior to what they themselves (self-image) and their peers (social image) consider morally ideal. Since there is no such thing as an 'objectively correct' normative view, however, we assume that over time, ethical views are learnt from one's peers (Algan et al., 2023), although reluctantly so (Brekke et al. 2010): in line with the literature on biased learning and motivated beliefs (Babcock and Loewenstein 1997; Hart et al. 2009; Deffains et al. 2016), an ethical view is less likely to be adopted if it implies a higher morally ideal contribution for oneself. Finally, we assume that there are several social peer groups, and that individuals occasionally revise which group they prefer to be part of.

With these assumptions, we find that in the steady state, society is characterized by strong ethical polarization, segregation according to income, and minimal voluntary public good contributions. This occurs although everyone is assumed to be morally motivated, in the sense that their self-image (social image) can only be improved by getting closer to what they (their peers) truly consider morally ideal – and although ethical views cannot simply be chosen, being fixed convictions in the short run. Over time, however, the combined forces of social learning and social migration gradually reduce individuals' exposure to peers whose ethical views would have imposed heavy moral responsibilities on them, allowing fairness views to become increasingly (implicitly) self-serving over time.

Policies increasing contact between diverse social groups, however, can limit polarization. Interestingly, if public good provision is left to the public sector, strong polarization need not arise: in contrast to voluntary contributions, the size of tax payments is not determined by the individual herself. This eases the moral burden associated with non-self-serving ethical views, removing the drive towards reluctant learning even at very low levels of polarization.

Our model is designed to highlight ethical disagreement concerning inequality and redistribution, and is thus not well suited to explore normative disagreement and polarization along other dimensions such as universalism, cultural or national identity, liberal versus conservative values, or democratic versus totalitarian views (see, e.g., Shayo 2009; Enke 2019; Alesina et al. 2020; Bonomi et al. 2021). Note, however, that our results follow from three major elements, which may be adapted to explore other segregation and polarization dimensions: first, preferences for self-image (behaving in line with one's own beliefs) as well as social image (behaving in line with peers' beliefs); second, exogenous heterogeneity in the burden imposed on individuals by holders of alternative beliefs; and third, the idea that beliefs are socially learnt, but reluctantly so. One could, for example, envision a model along similar lines addressing climate change denial: those who have invested heavily in continued use of fossil fuels (in terms of, e.g., education, work experience, or financial investments) may be more reluctant to accept the belief that continued fossil fuel use causes substantial negative externalities, and may also prefer to be surrounded by peers sharing their skepticism. This would yield a push over time towards social segregation between those who are more, respectively less,

heavily pre-invested in fossil fuel activities, with a self-reinforcing drive towards gradually more extreme beliefs within each group.

Our framework provides one possible explanation why individuals' fairness views tend to be strongly correlated with one's own economic situation in implicitly self-serving ways (Konow 2000; Di Tella et al. 2007; Fehr and Vollman 2022; Hvidberg et al., 2023; Cohn et al. 2023; Lobeck and Støstad 2023; Amasino et al. 2024), and why people tend to interact mainly with others of a socio-economic status similar to their own (e.g., Eika et al. 2019; Chetty et al. 2020, 2022). Our conclusions on minimal equilibrium voluntary contributions are in line with the experimental findings of Nikiforakis et al. (2012), Gangadharan et al. (2017), and Koch et al. (2021), who find that in heterogeneous groups, normative conflict limits the scope for efficient voluntary cooperation.

The mechanisms we describe are largely driven by endogenous *formation* of ethical principles, a feature that, to our knowledge, distinguishes our approach from most previous analyses. While our migration mechanism is related to Schelling's (1978) segregation model, for example, Schelling's model involves neither changing attitudes nor voluntary public good contributions. 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. Alesina and Angeletos (2005) and Benabou and Tirole (2006b) explore the joint development of fairness beliefs and tax and transfer policies. In Alesina and Angeletos (2005), policy influences descriptive beliefs on the luck versus merit origin of individual incomes, thus affecting views of whether the income distribution is fair; the normative fairness principle itself, however, is kept exogenously fixed. Similarly, the endogenous ideological beliefs in Benabou and Tirole (2006b) are associated with optimistic versus pessimistic descriptive beliefs about the degree to which individual effort pays off, while ethical views per se do not change.[1]

Below, we begin by presenting our set-up for individuals' short-term utility maximization, keeping ethical views as well as social neighborhoods fixed. Individuals' short-term choice consist of two steps, which we will discuss in reverse order: first, each individual consults her current ethical principles to determine what she considers the morally ideal contribution for each member of society, including herself; then, given her view of this ideal, she chooses how much to actually contribute to the public good. We then turn separately to the long-term mechanisms of social learning and migration, before merging all mechanisms into our integrated dynamic model. After this we demonstrate our minimal contribution result; show that policies encouraging contact between social groups can modify equilibrium polarization; and finally demonstrate that polarization may be limited by leaving public good provision to the government.

[1] A number of previous contributions explore endogenous change of normative values based on different approaches and aims than ours, and without discussing polarization and segregation. For example, Benabou et al. (2018) explore the spread of narratives influencing moral reasoning; Besley and Persson (2023) allow consumers' environmental values to change endogenously based on rational foresight; and Alger and Weibull (2016) show that what they call Kantian morality is evolutionary stable (see also Alger et al. 2020).

2. Basic model set-up

In the short run, individuals take their ethical views and social group affiliation as given. Based on her own ethical views, each individual derives corresponding beliefs about how much each member of society should ideally contribute to the public good. We return below to the endogenous determination of these beliefs; however, since ethical views are fixed in the short run, perceptions of morally ideal contributions can also be regarded as fixed in the short run. [2]

Let there be $N > 1$ individuals in the economy. Each individual i has an exogenously given disposable income $Y_i > 0$ in each period, constant over time, that within each period t can be spent on private consumption $c_i^t \geq 0$ or contributions $e_i^t \geq 0$ towards a global public good E^t (we disregard saving and borrowing):

$$(1) \qquad Y_i = c_i^t + e_i^t$$

In each period the total supply of the public good is given by the sum of individuals' voluntary contributions:

$$(2) \qquad E^t = \sum_{i=1}^{N} e_i^t$$

Each individual regards contributions from others as exogenously given, unaffected by her own behavior.

Individuals have identical utility functions. An individual i's utility in period t depends on her private consumption c_i^t, the total supply of the public good E^t, and her image as a decent person, the latter consisting of her self-image I_i^t and her social image S_i^t. For simplicity, let preferences be linearly separable and identical for all i, as follows:

$$(3) \qquad U_i^t = u\left(c_i^t\right) + \gamma E^t + I_i^t + S_i^t,$$

where u is a strictly increasing and strictly concave function (further down, we will specify $u\left(c_i^t\right) = \ln c_i^t$), while $\gamma > 0$. To focus on the case in which individuals would be free-riders in the absence of image preferences, we assume that for all i, $u'(Y_i) > \gamma$ (primes denote derivatives).

As long as i's actual contribution e_i^t falls short of her view of the morally *ideal* contribution for herself, e_{ii}^{t*}, the individual's self-image is better the smaller the distance between the two:

$$(4) \qquad I_i^t = \begin{cases} -\frac{\alpha}{2}\left(e_i^t - e_{ii}^{t*}\right)^2 & for\ e_i^t < e_{ii}^{t*} \\ 0 & otherwise, \end{cases}$$

[2] For simplicity and ease of intuitive understanding, we keep the discrete time and population set-up from Brekke et al. (2003) in our description of individuals' short-term utility maximization. When turning to the dynamics later on, however, we move to continuous time and population, letting the time period length go to zero.

where $\alpha > 0$.[3] The utility loss of contributing less than one's ideal may be interpreted as representing cognitive dissonance.

Individuals also care about their social image; that is, they would like others to consider them as morally responsible or decent people. Since individuals may have different ethical views, however, others' views of the morally ideal behavior for a person may differ from that individual's own view. Let e_{ij}^{t*} be the morally ideal contribution for i as judged by individual j (in period t). Furthermore, let e_{iG}^{t*} be the average e_{ij}^{t*} in i's social group or neighborhood (or, in the case of only one group, the entire population): $e_{iG}^{t*} = \frac{\sum_{j \in G} e_{ij}^{t*}}{N_G}$.[4]

Social image can now be specified in a manner similar to self-image, except that i's actual contribution e_i^t is now compared to *others'* view of i's morally ideal contribution:

$$(5) \qquad S_i^t = \begin{cases} -\frac{\beta}{2}\left(e_i^t - e_{iG}^{t*}\right)^2 & for\ e_i^t < e_{iG}^{t*} \\ 0 & otherwise, \end{cases}$$

where $\beta > 0$. Others' ethical views, and thus e_{iG}^{t*}, are considered exogenous by individuals.[5]

To prevent confusion between social groups and income groups, we will sometimes use the term social neighborhoods rather than groups. This implies no presumption about the division being geographical; the crucial aspect of a social neighborhood or group is that people are considerably more likely to meet within groups than across groups (since social learning occurs within groups): any individual i in neighborhood G_i is considerably more likely to meet another individual $j \neq i$ if $G_i = G_j$ than if $G_i \neq G_j$.

3. Short-term utility maximization

Let us now consider individuals' choice of how much to contribute to the public good in a given period, which is determined by short-run utility maximization taking neighborhood affiliation and ethical views as given. We demonstrate two core results: first, actual contributions are always strictly lower than the highest ideal (e_{ii}^{t*} or e_{iG}^{t*}); second, utility is (weakly) decreasing in each of these ideals.

Consider first the contribution choice. If individual i and her peers agreed on the morally ideal contribution for i, i.e., $e_{ii}^{t*} = e_{iG}^{t*}$, i would always contribute less than this shared ideal: if we did have $e_i^t = e_{ii}^{t*} = e_{iG}^{t*} > 0$, the marginal image gain would be zero, while the marginal cost in terms of lost consumption benefits would be strictly positive (Brekke et al. 2003). :

[3] Unlike Brekke et al. (2003), we restrict the quadratic loss function to apply for contributions below the ideal. In Brekke et al. (2003), such restrictions would not matter since equilibrium contributions were always strictly lower than the ideal. In the present analysis, this is not necessarily the case. While strictly negative image effects of 'too large' contributions would complicate the model, we do not think substantial new insight would be added.

[4] The logical but cumbersome notation would be $e_{iG_i}^{t*}$; we hope the simplified notation is sufficiently clear.

[5] The set-up so far implicitly assumes that incomes, actual contributions, as well as e_{iG}^{t*} are observable. In the dynamic analysis further below, we introduce uncertainty in beliefs about others' ethical views. Taking such uncertainty explicitly into account at the present point, however, would only complicate the notation without changing anything, since e_{iG}^{t*} is considered exogenously fixed by i.

The mechanism is quite similar even if $e_{ii}^{t*} \neq e_{iG}^{t*}$, as long as these ideals are not too different. There are three possible cases. If $e_i^t < \min(e_{ii}^{t*}, e_{iG}^{t*})$, the first order condition for an interior utility maximum with respect to e_i^t is (primes denote derivatives)[6]

(6) $\quad e_i^t = \frac{\alpha e_{ii}^{t*} + \beta e_{iG}^{t*}}{(\alpha+\beta)} - \frac{u'-\gamma}{\alpha+\beta}$,

implying that actual contribution falls short of the weighted average ideal by $\frac{u'-\gamma}{\alpha+\beta} > 0$.

If α and/or β are large and the ideals are very different, however, e_i^t as specified by eq. (6) may violate the assumption $e_i^t < \min(e_{ii}^{t*}, e_{iG}^{t*})$: individuals may choose to contribute more than they find morally ideal to gain social approval, or to contribute more than considered ideal by their peers to improve their self-image.

Consider first the case where $e_{ii}^{t*} < e_{iG}^{t*}$ and $e_{ii}^{t*} < \frac{\alpha e_{ii}^{t*} + \beta e_{iG}^{t*}}{(\alpha+\beta)} - \frac{u'-\gamma}{\alpha+\beta}$. Actual contributions will now be in the area where $I_i^t = 0$. The first order condition for utility maximization now becomes

(6′) $\quad e_i^t = e_{iG}^{t*} - \frac{u'-\gamma}{\beta}$.

By symmetry, if $e_{ii}^{t*} > e_{iG}^{t*}$,

(6″) $\quad e_i^t = e_{ii}^{t*} - \frac{u'-\gamma}{\alpha}$.

In all three cases, the actual contribution is strictly lower than the highest ideal: one strives towards the ideals, but unless $e_{ii}^{t*} = e_{iG}^{t*} = 0$, one never quite reaches both of them.

How would the contribution choice respond to changing ideals? If the actual contribution already exceeded the ideal, a change in the ideal would be inconsequential; otherwise, the actual contribution is strictly increasing in the ideal. Using total differentiation of eq. (6), we have

(7) $\quad \frac{de_i^t}{de_{ii}^{t*}} = \frac{\alpha}{\alpha+\beta-u''} \geq 0$, with strict inequality whenever $e_i^t < e_{ii}^{t*}$, and

(8) $\quad \frac{de_i^t}{de_{iG}^{t*}} = \frac{\beta}{\alpha+\beta-u''} \geq 0$, with strict inequality whenever $e_i^t < e_{iG}^{t*}$.

Utility, however, is non-increasing in the ideals: moral responsibility is a burden. By the envelope theorem, we get

(9) $\quad \frac{dU_i^t}{de_{ii}^{t*}} = \begin{cases} \alpha\left(e_i^t - e_{ii}^{t*}\right) < 0 & for\ e_i^t < e_{ii}^{t*} \\ 0 & otherwise, \end{cases}$

(10) $\quad \frac{dU_i^t}{de_{iG}^{t*}} = \begin{cases} \beta(e_i^t - e_{iG}^{t*}) & for\ e_i^t < e_{iG}^{t*} \\ 0 & otherwise. \end{cases}$

[6] Derived by inserting eqs. (1), (2), (4) and (5) in eq. (3), then differentiating with respect to e_i^t while regarding e_{ii}^{t*} and e_{iG}^{t*} as exogenous.

Total public good supply is given by the sum of everyone's contributions. In the case where $e_i^t < \min(e_{ii}^{t*}, e_{iG}^{t*})$ for all i, for example, inserting the expressions for contribution choice in eq. (2) gives

$$(11) \quad E^t = \sum_{i=1}^{N} \frac{\alpha e_{ii}^{t*} + \beta e_{iG}^{t*} - u' + \gamma}{(\alpha+\beta)},$$

which is increasing in the moral ideals e_{ij}^{t*} for every i, j (recall that e_{iG}^{t*} is only the average of all e_{ij}^{t*} in group G). Hence, moral responsibility is a burden; this burden, however, helps secure a higher voluntary supply of the public good.

We summarize the above discussion in a proposition.

Proposition 1.

a) The contribution by individual i ($i = 1, \dots, N$) in period t, e_i^t, is given by (6), (6'), or (6'').
b) Utility is weakly decreasing in e_{ii}^{t*}. The decrease is strict if $e_i^t < e_{ii}^{t*}$.
c) Utility is weakly decreasing in e_{iG}^{t*}. The decrease is strict if $e_i^t < e_{iG}^{t*}$.
d) Total voluntary public good provision in period t is weakly increasing in e_{ii}^{t*} for any $i = 1, \dots, N$, and strictly increasing for those i for whom $e_i^t < e_{ii}^{t*}$.
e) Total voluntary public good provision in period t is weakly increasing in e_{iG}^{t*} for any $i \in G$. The increase is strict for those i, G such that $e_i^t < e_{iG}^{t*}$.

4. The judgement of morally ideal contributions

A central aspect of the current analysis is that views about morally ideal contributions are derived through ethical deliberation. Thus, before turning to social learning of ethical principles in the next section, let uselaborate a bit on individuals' assumed ethical reasoning.

Along the lines of Harsanyi (1955), Brekke et al. (1996), and Nyborg (2000), we formalize ethical views in terms of subjective social welfare functions, representing individuals' normative views of the good society. While we assume that these subjective social welfare functions are welfaristic and additive, the relative weights μ_{ij}^t placed by i on j's well-being may vary:

$$(12) \quad W_i^t = \sum_{j=1}^{N} \mu_{ij}^t \left(u\left(c_j^t\right) + \gamma E^t\right)$$

where W_i^t is i's judgement of social welfare in period t, and $\sum_{j=1}^{N} \mu_{ij}^t = 1$.[7] Note that the subjective social welfare functions include only material well-being – that is, utility derived from private consumption and public goods: when judging social welfare, individuals do not take image concerns into account. While it can be argued that image benefits matter to individuals and thus should count in welfare judgements, one may also argue that it is unreasonable to incorporate the desire to do

[7] Identical preferences allow well-being comparisons once others' incomes are known, since access to the public good is identical for all. Implicitly, thus, we assume that incomes are observable and that it is common knowledge that material well-being preferences are identical. Although the linear form may seem restrictive, non-linear social welfare functions may be represented by their linear approximations if the contribution decision is a marginal one.

what is right in the very judgement of what is right. To avoid complicating the analysis, we adopt the latter view here.[8]

Based on the idea that moral action should be based on some universal law applicable to everyone, such as Confucius' *Do not do to others what you would not want others to do to you*, the Biblical Jesus' Golden Rule, or Kant's Categorical Imperative, Brekke et al. (2003) defined the morally ideal contribution as *the contribution that would maximize social welfare in the hypothetical case that everyone acted exactly like oneself*. To account for individual heterogeneity in income as well as ethical views, we modify this principle by assuming that the morally ideal contribution for individual j as judged by i is understood as *the contribution by j that would maximize i's subjective social welfare function in the hypothetical case that every other individual $k \neq j$ also contributed their morally ideal contribution (as judged by i).*

This approach can be considered Kantian in the following sense: the morally ideal behavior is to act according to a rule that one would wish everyone to follow. Note, however, that "would wish" here refers to individuals' views in their role as ethical observers, referring to their subjective social welfare functions, thus involving preferences concerning distribution. This is in contrast to several other economic approaches to Kantian morality, such as Laffont (1975), Bilodeau and Gravel (2004), Roemer (2015), and Alger and Weibull (2016), where "would wish" simply refers to individuals' own utility functions.

Note also that although the morally ideal contributions are derived according to a version of Kant's categorical imperative, individuals do not *behave* according to the categorical imperative; when deciding how much to actually contribute, they make trade-offs between image concerns and consumption.

Formally, individual i's judgement of the morally ideal contribution for j in period t, e_{ji}^{t*}, is found by maximizing W_i^t with respect to e_j^t, subject to eqs. (1), (2), (12), and the (hypothetical) assumption that $e_k^t = e_{ki}^{t*}$ for all $k = 1, \dots . N$.

To simplify expressions below, we henceforth specify the utility of consumption as $u(c_i^t) = \ln c_i^t$. Given this, the first order condition for an interior social welfare maximum with respect to e_j^t, as judged by i, is

$$(13) \quad e_{ji}^{t*} = Y_j - \frac{\mu_{ij}^t}{N\gamma}.$$

Due to differences in income and welfare weights, it may now be the case that $e_{ji}^{t*} \neq e_{ki}^{t*}$: i thinks the morally ideal contribution for i differs from the morally ideal contribution for k. Also, if $\mu_{ik}^t \neq \mu_{jk}^t$ for some k, i and j may disagree on the morally ideal contribution for individual k ($e_{ki}^{t*} \neq e_{kj}^{t*}$).

Notice that due to equations (2) and (13), i's view of the socially optimal public good level, E_i^{t*}, is given by $E_i^{t*} = \sum_{j=1}^N e_{ji}^{t*} = (\sum_{j=1}^N Y_j) - \frac{1}{N\gamma} = E^*$. Hence, all $i = 1, \dots, N$ agree on the socially optimal

[8] Including only self-image in the social welfare function would be straightforward: when considering the morally ideal action, i assumes that $e_j^t = e_{ji}^{t*}$ for all j, implying $I_j^t = 0$ for everyone (Brekke et al. 2003). However, if $e_{ii}^{t*} \neq e_{iG}^{t*}$, including both self-image and social image concerns would complicate the analysis.

public good supply E^*; what differs is the normative view of the fair burden-sharing, i.e., the socially optimal distribution of contributions.

5. Utilitarians and status quo supporters

Different subjective social welfare functions – i.e., different sets of social well-being weights μ_{ij}^t – imply different redistribution preferences, and thus different views on who should contribute how much to the public good.

Note first that some ethical observers may find the exogenous initial income distribution socially optimal, preferring no redistribution. For example, if the current income distribution is believed to reflect individual effort, this may be viewed as fair according to a meritocratic principle. Such a view would imply social well-being weights that are inversely proportional to the individual's marginal utility of consumption (evaluated in the initial state, i.e., without contributions), often called Negishi weights. [9] Let us call an individual subscribing to Negishi weights in their subjective social welfare function a *status quo supporter.*

Given that the social well-being weights μ_{ij}^t are relative, and $u\left(c_i^t\right) = \ln c_i^t$, Negishi weights in i's social welfare function correspond to $\mu_{ij}^t = \frac{Y_j}{\sum_k Y_k} = \frac{Y_j}{N\bar{Y}}$, with $\bar{Y} = \frac{\sum_k Y_k}{N}$ denoting average income ($k = 1, \dots, N$). Assuming an interior welfare optimum, eq. (13) then implies $e_{ji}^{t*} = Y_j - \frac{Y_j}{\bar{Y}N^2\gamma}$; i.e., ideal contributions are proportional to each individual's income.

Other social well-being weights than Negishi weights involve a preference for redistribution, affecting the individual's view of who should ideally contribute how much. In particular, with an unweighted utilitarian social welfare function, corresponding to $\mu_{ij}^t = \frac{1}{N}$ for all j, the morally ideal or fair contribution for person j as judged by i is given by $e_{ji}^{t*} = Y_j - \frac{1}{N^2\gamma}$. That is, the morally ideal contribution is so much higher for those with higher incomes as to completely cancel out the exogenous income differences between individuals: the utilitarian would ideally prefer everyone to have the same consumption level. In the present context, thus, unweighted utilitarianism implies egalitarianism.

On a scale of redistribution preferences ranging from complete equalization of private consumption to no preferred redistribution at all, Negishi weights and (equally-weighted) utilitarianism thus represent polar cases in the current context.[10]

[9] «[A] competitive equilibrium is a maximum point of a social welfare function which is a linear combination of utility functions of consumers, with the weights in the combination in inverse proportion to the marginal utilities of income» (Negishi 1960, p.92).

[10] This may seem puzzling, since utilitarianism does not generally involve inequality aversion. However, in our context, social welfare functions displaying explicit inequality aversion would not prescribe more redistribution than the utilitarian one. While prioritarianism (see, e.g., Adler and Norheim 2022) generally gives more priority to those with lower well-being levels, any well-being differences would be cancelled out even in the utilitarian social optimum in the present setting.

Importantly, for any j with an income below average, $Y_j < \bar{Y}$, Negishi weights imply a higher burden of moral responsibility – that is, a higher e_{ji}^{t*} – than equal weights. For high-income individuals, i.e., any j for whom $Y_j \geq \bar{Y}$, the opposite is true. For ease of reference, let R be the set of all j for whom $Y_j \geq \bar{Y}$, called 'the rich' below; similarly, let P be the set of all j such that $Y_j < \bar{Y}$, called 'the poor' below.

Reasonable moral reasoning should be based on universal principles. As a minimum, let the weights μ_{ij}^t satisfy anonymity in the sense that if $Y_j^t = Y_k^t$, then $\mu_{ij}^t = \mu_{ik}^t$ (recall that material well-being preferences are identical, and since E^t is a pure public good, it is equally accessible to all). Further, let the weights μ_{ij}^t satisfy continuity in the sense that if $Y_j^t > Y_k^t > Y_l^t$, we require μ_{ik}^t to lie between μ_{ij}^t and μ_{ik}^t.[11] A simple, one-dimensional measure of the strength of i's redistribution preference, as in eq. (14) below, satisfies these requirements..

Assume that every individual $i = 1, \dots, N$ subscribes to a subjective social welfare function that can be placed on a scale ranging from (unweighted) utilitarianism to status quo support (thus excluding preferences for further increased inequality). Let q_i^t reflect i's degree of status quo support, i.e., the degree to which i finds the status quo income distribution socially optimal at time t, such that $q_i^t = 1$ corresponds to using Negishi weights while $q_i^t = 0$ corresponds to unweighted utilitarianism. q_i^t can now be defined implicitly by i's weights μ_{ij}^t as follows: for each $i, j = 1, \dots, N$,

$$(14) \quad \mu_{ij}^t = \frac{(1-q_i^t)\bar{Y}+q_i^t Y_j}{N\bar{Y}}.$$

Whenever $q_i^t \neq 1$, the associated morally ideal contributions will partly serve a purpose of (hypothetically) redistributing income.

It follows that the more status quo supportive an ethical observer i is (the higher q_i^t), the lower is her view of the morally ideal contribution for those with high incomes, and the higher is her view of the morally ideal contribution for those with low incomes:

$$(15) \quad e_{ji}^{t*} = \left(Y_j - \frac{1}{N\gamma\bar{Y}}\right) - \frac{q_i^t}{N\gamma}\frac{(Y_j-\bar{Y})}{\bar{Y}}.$$

Note that since eq. (15) is linear in q_i^t, the average judgement in j's social group of the ideal contribution for j, e_{jG}^{t*}, can now be written as a function of the average degree of status quo support in j's social group or neighborhood G, $q_G^t = \frac{\sum_{k\in G} q_k^t}{N^G}$, where N^G is the number of members in j's social group:

$$(16) \quad e_{jG}^{t*} = \left(Y_j - \frac{1}{N\gamma\bar{Y}}\right) - \frac{q_G^t}{N\gamma}\frac{(Y_j-\bar{Y})}{\bar{Y}}.$$

[11] More precisely, if $Y_j^t > Y_k^t > Y_l^t$, then we require that either $\mu_{ij}^t \geq \mu_{ik}^t \geq \mu_{il}^t$ or $\mu_{ij}^t \leq \mu_{ik}^t \leq \mu_{il}^t$.

Proposition 2. Let $q_i^t \in [0,1]$ be the relative degree of status quo support in i's welfare evaluation weights at time t, μ_{ij}^t, in such a way that $\mu_{ij}^t = \frac{(1-q_i^t)\bar{Y}+q_i^t Y_j}{N\bar{Y}}$ holds for every $i, j = 1, \dots, N$. Then, the following holds:

a) For any $j \in P$ (the poor), the morally ideal contribution e_{ji}^{t*} ascribed to j by i is increasing in q_i^t. Conversely, for any $j \in R$ (the rich), the morally ideal contribution e_{ji}^{t*} ascribed to j by i is decreasing in q_i^t.
b) For any $i \in P$ (the poor), U_i^t is decreasing in q_i^t. Conversely, for any $i \in R$ (the rich), U_i^t is increasing in q_i^t. The ranking may be weak for some individuals.
c) For any $i \in P$ (the poor), U_i^t is decreasing in q_G^t. Conversely, for any $i \in R$ (the rich), U_i^t is increasing in q_G^t. The ranking may be weak for some individuals.

Proof:

Part a) follows from eqs. (15) and (16) above. For b) and c), note that, as a direct consequence of a) and Proposition 1b) and 1c), we know the following: For the poor, period utility U_i^t is nonincreasing in q_i^t, and strictly decreasing for the poor whose contribution falls short of the implied ideal; for the rich, period utility U_i^t is nondecreasing in q_i^t, and strictly increasing for the rich whose contribution falls short of the implied ideal. For the poor, period utility U_i^t is nonincreasing in q_G^t, and strictly increasing for the poor whose contribution falls short of the implied ideal; for the rich, period utility U_i^t is nondecreasing in q_G^t, and strictly increasing for the rich whose contribution falls short of the implied ideal. These two observations imply b) and c). ■

Note that we assume individuals to be myopic in the sense that when choosing their voluntary contributions to the public good, they do not take into account the possible consequences of this choice for their own future learning process and migration decisions. Although a formal analysis of this remains to be done, we do not believe that changing this assumption would have substantial consequences for our results. This would essentially correspond to all choices being made immediately, which would presumably just speed up the process to be described below; however, coordination failures might arise in the migration process, since one would not know in advance which social group would be the most status quo supportive.

6. Social learning of ethical views

Let us now turn to social learning of ethical views. For the moment, we disregard migration, keeping groups fixed; migration will be added to the picture in the next section.

Ethical views – here represented by social welfare functions – cannot be chosen freely; they represent, rather, convictions. In the short run, thus, it seems reasonable to consider subjective social welfare functions as given. Nevertheless, precisely because social welfare functions are inherently normative, there is no such thing as an objectively "correct" social welfare function: ethical principles reflect subjective value judgements, and cannot be deduced from facts and logic alone. So how do such convictions arise? Presumably, they are at least partly shaped by social interaction – through learning, observation, and debate (norms instilled by parents or school;

observing role models' behaviors and statements; shared ethical discussion, deliberation and reflection in a social context).[12]

An update of i's social welfare function could change q_i^t, i's degree of status quo support. As can be seen from eq. (14), this would involve a change in i's social well-being weights μ_{ij}^t, and thus also her view of the morally ideal contribution e_{ji}^{t*} for any member of society $j = 1, \dots, N$, including herself. It is crucial to note that in general, an increased q_i^t would imply a *lower* e_{ji}^{t*} if j is rich but a *higher* e_{ji}^{t*} if j is poor: For the poor, the burden of moral responsibility is increasing in the degree of status quo support; for the rich, moral responsibility increases in the degree of utilitarianism.

Assume now that ethical views are not perfectly observable: Although social interaction between individuals provide indications of their views, a person i cannot infer another's views q_j^t exactly. To fix ideas, let us continue with a discrete time setup for the moment, with a time-step Δt.

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 status quo support q_i^t a fraction $\delta > 0$ in the direction of what i perceives to be j's view, $\tilde{q}_{ji}^t$. As i cannot observe j´s view accurately, the perception is established with some noise, but we take it to be unbiased: $E\tilde{q}_{ji}^t = q_j^t$. To avoid truncating the distribution of $\tilde{q}_{ji}^t$, we assume the distribution is symmetric and has support in [0,1].[13]

With unbiased learning, the change in i´s view is thus

$$(17) \quad \Delta q_i^t = q_i^{t+\Delta t} - q_i^t = \delta\left(\tilde{q}_{ji}^t - q_i^t\right)\Delta t,$$

where $i, j \in G_i$, since j is part of i's own social group.

Let us now move to continuous time by letting the time step approach zero. i´s view will then be 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

$$(18) \quad \dot{q}_i^t = \delta\left(q_G^t - q_i^t\right).$$

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, we find that, with unbiased social learning within a fixed group, the average ethical view of the group stays unchanged:

$$(19) \quad \dot{q}_G^t = \delta(q_G^t - q_G^t) = 0.$$

Nevertheless, since individual views adjust over time, all group members' views converge towards the initial group average q_G^0: in-group variation is gradually reduced, eliminating over time the initial differences between q_i^t and q_j^t for any $i \neq j$ for whom $G_i = G_j$. Hence, if there are several fixed

[12] For example, the process of reconsidering one's principles as well as one's emotional reactions to these principles' specific implications, as implied by Rawls' idea of a reflective equilibrium (Rawls 1971), may well occur in a social context, involving discussion and argument between peers.

[13] Note that this requirement on the distribution of $\tilde{q}_{ji}^t$ implies that the 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.

groups or neighborhoods, and the initial average degree of status quo support q_G^0 differs between groups, there is a weak sense in which one may say that social learning of ethical views causes polarization between neighborhoods even with unbiased social learning and no migration: as time passes, ethical disagreement is reduced within groups, but not between groups.

Let us now turn to the case of reluctant social learning. Reluctant learners are assumed to be somewhat less prone to change their view if moving towards the other's view would increase their own perceived moral burden. This could either be because there is a small probability that the other's view is disregarded in such cases, or because i's view takes a shorter step in the direction of j's.

To clarify (going back, for a moment, to discrete time), note that eq. (17) implies $q_i^{t+\Delta t} = q_i^t + \delta(\tilde{q}_{ji}^t - q_i^t)\Delta t$. For an unbiased learner i, it follows that as i's status quo support changes from q_i^t to $q_i^{t+\Delta t}$, i's perceived morally ideal contribution also changes, from e_{ii}^{t*} to $e_{ii}^{t+\Delta t\,*}$ (due to eq. 15). We assume that a reluctant learner is less prone to adopt the views of a peer if $e_{ii}^{t+\Delta t\,*} > e_{ii}^{t*}$, since doing so would increase her burden of moral responsibility. In our main model, $e_{ii}^{t+\Delta t\,*} > e_{ii}^{t*}$ simplifies to $\tilde{e}_{ij}^{t*} > e_{ii}^{t*}$. Thus, for reluctant social learners, we assume that the strength of movement in q_i^t equals δ as above whenever $e_{ii}^{t*} \geq \tilde{e}_{ij}^{t*}$, i.e., when one thinks the other's view implies a lower moral responsibility, but limited to $\delta(1-r)$, where $0 < r < 1$, whenever $e_{ii}^{t*} < \tilde{e}_{ij}^{t*}$.

Consider the situation where individual i meets j. If i is rich, then a *reduction* of q_i^t would imply a higher morally ideal contribution e_{ii}^{t*} for i. Conversely, if i is poor, an *increase* of q_i^t would increase e_{ii}^{t*}. If i is rich, this means (returning for the moment to discrete time) that the movement in q_i^t is systematically lower whenever $\tilde{q}_{ji}^t < q_i^t$:

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

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

Similarly, for the poor:

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

The size of the bias in reluctant learning depends, in addition to the length of the time step, on i) the parameters r and δ; ii) the difference between $\tilde{q}_{ji}^t$ and q_i^t; and iii) whether i is rich or poor, since the rich are reluctant when the difference in ii) is negative, while the poor are reluctant when the difference is positive. To simplify notation, we let $(\tilde{q}_j^t - q_i^t)^-$ denote the negative part of $(\tilde{q}_j^t - q_i^t)$, while $(\tilde{q}_{ji}^t - q_i^t)^+$ denotes the positive part. [14] A more concise way to write the change over time in q_i^t, for rich as well as poor, is then

[14] 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.

$$(20) \quad \Delta q_i^t = \delta\left(\tilde{q}_{ji}^t - q_i^t\right)\Delta t \begin{cases} +r\delta\left(\tilde{q}_{ji}^t - q_i^t\right)^- \Delta t & if\ i \in R \\ -r\delta\left(\tilde{q}_{ji}^t - q_i^t\right)^+ \Delta t & if\ i \in P. \end{cases}$$

Let $B_{ij}^{+t} = E\left(\tilde{q}_{ji}^t - q_i^t\right)^+ > 0$. Similarly, let $B_{ij}^{-t} > 0$ be the expected negative part. Furthermore, to indicate that, for example, i is rich and j is poor and hence the expectation must be taken over $i \in R$ and $j \in P$, let us write $B_{RP}^{+t} = E(\tilde{q}_{PR}^t - q_R^t)^+$. These variables are proportional to the expected size of the learning biases in the various cases. A rich individual is only reluctant to adopt the perceived view of j if j is less status quo supportive than himself, hence only B_{RR}^{-t} and B_{RP}^{-t} matter for the rich; similarly, only B_{PP}^{+t} and B_{PR}^{+t} matter for the poor.

Now, let s_G be the share of rich individuals in group G. The probability that a rich person meets another rich person is then s_G. The dynamic of rich individuals' views – now again moving to continuous time – is thus

$$(21) \quad \dot{q}_R^t = (1 - s_G)\delta(q_P^t - q_R^t) + s_G r\delta B_{RR}^{-t} + (1 - s_G) r\delta B_{RP}^{-t},$$

where $\dot{q}_R^t$ is the change in status quo support for any $i \in R$. Similarly, for the poor, any $i \in P$, we have

$$(22) \quad \dot{q}_P^t = s_G\delta(q_R^t - q_P^t) - s_G r\delta B_{PR}^{+t} - (1 - s_G) r\delta B_{PP}^{+t}.$$

The dynamic for the group's average view $\dot{q}_G^t$ is the weighted average of the two, which gives

$$(23) \quad \dot{q}_G^t = r\delta[s_G^2 B_{RR}^{-t} + (1 - s_G)s_G(B_{RP}^{-t} - B_{PR}^{+t}) - (1 - s_G)^2 B_{PP}^{+t}].$$

To proceed, we need to specify our probability distribution for $\tilde{q}_{ji}^t$. Assumption 1 maximizes the variance of $\tilde{q}_{ji}^t$ and thus also the possibility that equilibria with limited polarization may exist:

Assumption 1. Let the probability distribution for $\tilde{q}_{ji}^t$ have support on $[0,1]$ with $E\left(\tilde{q}_{ji}^t\right) = q_j^t$. Moreover, assume that if $q_j^t \leq 0.5$, then $\tilde{q}_{ji}^t = 0$ or $2q_j^t$ with equal probability, while if $q_j^t \geq 0.5$, then $\tilde{q}_{ji}^t = 1$ or $2q_j^t - 1$ with equal probability.

We can now state the main result of this section:

Proposition 3. In a steady state, all $i \in R$ (the rich) hold the same view $q_i^t = q_R$, and all $i \in P$ (the poor) hold the same view $q_i^t = q_P$. Moreover, given Assumption 1,

I. For all values of s_G, a) $q_R = q_P = 0$ and b) $q_R = q_P = 1$ are stable states for which $\dot{q}_R = \dot{q}_P = \dot{q}_G = 0$.
II. If $s_G < \frac{1}{2}$, only a) is asymptotically stable, while for $s_G > \frac{1}{2}$, only b) is asymptotically stable.
III. For $s_G = \frac{1}{2}$, there is also a symmetric stable state with $q_R > \frac{1}{2} > q_P$ further characterized by $q_R = q_P + \frac{r}{2}$.
IV. Assuming $r < \frac{1}{2}$, there are no further stable states.

The proof of Proposition 3 is rather tedious, so we relegate the more detailed explanation and proof to Appendix 1. Here we will focus on the intuition.

Note that the sign of $\dot{q}_G^t$ depends on s_G and the relative size of the biases. Reluctance pulls the rich in the direction of becoming more status quo supportive (increasing q_i^t), while the poor are pulled towards utilitarianism (decreasing q_i^t). This drives the group average q_G^t towards zero if the majority is poor, and towards 1 if the majority is rich. If this effect dominates, there is strong polarization between groups with rich and poor majorities, respectively, in the long run: While groups with rich majorities become perfectly status quo supportive over time, groups with poor majorities become perfectly utilitarian over time.

There is also another effect, however: since all the rich within a given group are subject to the same dynamic, they become increasingly homogenous over time. The same holds for the poor. With fixed groups, substantial reluctance, and relatively similar shares of poor and rich within the group, it is conceivable that the latter effect dominates, allowing the existence of steady states in which rich and poor within the same group converge to different but less extreme views, thus limiting polarization.

7. Choosing one's peers

Let us now allow migration between social neighborhoods.

Although revising one's neighborhood affiliation is a choice, not an inference (in contrast to ethical view updating), we assume that individuals reconsider their preferred neighborhood affiliation only occasionally. [15] Social groups are thus relatively stable in the short run, making it reasonable to expect q_G^t to be a good proxy for q_G^{t+1}.

Assume now that there are only two equally large neighborhoods, A and B, and that the population consists of equally many rich and poor: there are $\frac{N}{2}$ individuals such that $Y_i \geq \bar{Y}$, and $\frac{N}{2}$ individuals such that $Y_i < \bar{Y}$. While not essential for the main mechanisms, these assumptions simplify the formalization below considerably. [16] Also, let individuals disregard the potential effect of their own migration on q_G^t in either neighborhood.

In the current framework, the only reason why social group affiliation matters to individuals is their preference for being viewed as morally responsible by their peers.[17] When revising their neighborhood status, rich individuals prefer the neighborhood with higher q_G^t, while the poor prefer the neighborhood with lower q_G^t: Since i's morally ideal contribution e_{ij}^{t*} i as judged by j is increasing in j's status quo support q_j^t if i is poor, but decreasing in q_j^t if i is rich (Proposition 2a), a rich individual i's utility is increasing in the status quo support q_G^t of i's social group, while a poor individual's utility is decreasing in q_G^t (Proposition 2c).

[15] For example, although not included in the formal model, the limited contact with people in other neighborhoods could make assessment of the views prevailing there cumbersome enough to be considered only every now and then.

[16] Since $s_{RG}^t + s_{PG}^t = 1$ for each group $G = \{A, B\}$, we only need to keep track of one of them (s_{RG}^t); when neighborhoods A and B are equally large, we also have $s_{RA}^t + s_{RB}^t = 1$, so s_{RG} denotes not only the share of group members in G who are rich, but also the share of all rich people who belongs to neighborhood G.

[17] In fact, given that social image is given by a quadratic loss function, individuals would prefer to be socially isolated, thus avoiding any social disapproval; we disregard this possibility.

The share of rich in each group can now vary over time. Let $s_{\theta G}^t$ denote the share of type $\theta \in \{R, P\}$, in group $G \in \{A, B\}$ at time t. Note that if $q_A^t > q_B^t$, only that share of the rich who are in B at time t, $1 - s_{RA}^t$, have incentives to move. Denoting $\rho > 0$ the share of individuals who revise their neighborhood affiliation in each period, and moving to continuous time by shortening period length towards zero, this can now be expressed as

$$(24) \quad \dot{s}_{RA}^t = -\dot{s}_{RB}^t = -\dot{s}_{PA}^t = \dot{s}_{PB}^t = \begin{cases} (1 - s_{RA}^t)\rho(q_A^t - q_B^t) & \text{when } q_A \geq q_B \\ s_{RA}^t \rho(q_A^t - q_B^t) & \text{when } q_A < q_B \end{cases}.$$

Eq. (24) shows that for migration to come to a rest, i.e., $\dot{s}_{RA} = 0$, we must have either $q_A = q_B$, or complete polarization: $s_{RA} = 1$ and $q_A \geq q_B$ or $s_{RA} = 0$ 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.

8. Total dynamics

Let us now bring the elements above together in a complete dynamic model. Eq. (24) above describes the dynamic development in the share of rich and poor in each social neighborhood. Eq. (23) describes the dynamics of ethical views caused by social learning in fixed groups, but without taking into account the direct effect of migration on the average status quo support in each neighborhood.

Note first that by writing eq. (23) 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), we can establish a compact notation R_G^t for the overall effect of reluctance:

$$(25) \quad \dot{q}_A^t = r\delta[s_A^2 B_{RR}^{-t} + (1 - s_A)s_A(B_{RP}^{-t} - B_{PR}^{+t}) - (1 - s_A)^2 B_{PP}^{+t}] = R_A^t$$

$$(26) \quad \dot{q_B} = r\delta[s_B^2 B_{RR}^{-t} + (1 - s_B)s_B(B_{RP}^{-t} - B_{PR}^{+t}) - (1 - s_B)^2 B_{PP}^{+t}] = R_B^t .$$

Here we have returned to the notation from Section 6 where s_G denotes the share of rich individuals in group G (since $s_{RG}^t = 1 - s_{PG}^t$, we only need to keep track of the share of rich in each group).

The set of equations (24) - (26) has one interior solution, $q_A = q_B$ and $s_G = \frac{1}{2}$, which is unstable: a slight deviation causing the status quo support in the two neighborhoods to differ, say $q_A > q_B$, would attract the rich to group A and the poor to B. Thus, if ignoring the direct effects of migration on q_G, reluctance would pull views gradually towards a higher q_A (since the rich attracted to A are reluctant to adopt more utilitarian 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 income; the rich would be strict status quo supporters, while the pure would be strict utilitarians.

Migration increases the status quo support in social neighborhood A directly to the extent that the rich moving from B to A are more status quo supportive than the poor migrating in the other direction. Thus, to consider the full effects of migration, an extra term $(q_{PB}^t - q_{RA}^t)\dot{s}_A$ must be added

to expression (25), where $q_{\theta G}^t$ denotes the average status quo support among members of income group $\theta = P, R$ in neighborhood $G = A, B$. A detailed explanation of why is provided in Appendix 2.

Inserting for $\dot{s}_A$ from Eq. (24) in the case where the rich group is A, and hence $q_A \geq q_B$, then gives

$$(27) \quad \dot{q}_A = R_A^t - \rho(1 - s_A^t)(q_A^t - q_B^t)(q_{PA}^t - q_{RB}^t)$$

Similarly, for the poor group, B,

$$(28) \quad \dot{q_B} = R_B^t + \rho s_B^t(q_A^t - q_B^t)(q_{PA}^t - q_{RB}^t).$$

These additional terms do not affect the equilibrium, however, because $\dot{s}_A = (q_{PB} - q_{RA}) = 0$ when the dynamic process has come to a rest (eq. (24)), and similarly for $\dot{s}_B$. Note further that close to the steady state, $1 - s_A \approx 0$ and $s_B \approx 0$, and as $s_A \to 1$ and $s_B \to 0$ by eq. (24), the migration terms will eventually be negligible and hence not affect the asymptotic stability of the equilibrium.

Outside of the steady state, the term $(q_{PA} - q_{RB})$ can in general be either positive or negative, depending on whether the poor in A are more or less status quo supportive than the rich in B. Since we have not imposed any restrictions on the relationship between individuals' initial status quo support and their income, it is conceivable that migration temporarily contributes to reductions in q_A and increases in q_B. Nevertheless, over time, reluctance makes the poor gradually more utilitarian and the rich gradually more status quo supportive (see Appendix 1), so such "reverse" movements cannot persist over time.

Intuitively, the average status quo support 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 share of status quo support to one of its boundaries, 0 or 1, and thus cannot push it any further. That is, all the rich flock together in one group (here, A), being entirely status quo supportive, while all the poor have self-selected into the other group and become completely utilitarian.

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

Proposition 4. Given Assumption 1, there are only two asymptotically stable states. In both states, all $i \in R$ (the rich) are located in one neighborhood and hold a completely status quo supportive ethical view ($q_i^t = q_R = 1$), while all $i \in P$ (the poor) are located in the other neighborhood and hold a completely utilitarian view ($q_i^t = q_P = 0$). Since a given neighborhood can either be the rich one or the poor one, there are two asymptotically stable states.

Proof: See Appendix 2.

Note that a modest but strictly positive migration cost would hardly change our main conclusions substantially. Only individuals whose social image benefits of moving exceed the value of the (presumably one-time) moving cost would then choose to migrate. While some marginal individuals (i.e., with incomes near $\bar{Y}$) would then choose to stay, influencing who is the marginal mover, reluctant learning among everyone else would still drive the groups' average views to the extremes.

9. Minimal contributions

As demonstrated above, morally motivated individuals tend to be driven towards strong segregation and polarization when there is reluctant social learning and migration between social neighborhoods: those with high incomes socialize with each other, finding the status quo income distribution fair, whereas those with low incomes also seek each other's company while supporting utilitarian views. In such a situation, everyone considers their own moral responsibility for contributing to the public good to be modest; moreover, everyone is surrounded by peers confirming this view. For example, while a morally motivated but poor individual may consider climate change a major social problem, she may still not contribute much, thinking that it is primarily the rich who ought to solve the problem. A morally motivated rich person, on the other hand, may agree that climate change is a major social problem, but would still contribute a rather limited amount because she finds no moral reason for the rich to contribute proportionally more than the poor.

Indeed, the long-run equilibrium not only represents the strongest possible segregation and polarization in our model, but also an absolute contribution minimum in the sense defined below.

Definition. A combination of sorting and ethical views is an *absolute contribution minimum* if, for given individual incomes, there is no other combination of ethical views and sorting into social groups that would yield strictly lower total contributions to the public good.

Proposition 5. With reluctance, the long-term equilibrium described in Proposition 4 is an absolute contribution minimum.

Proof: Proposition 4 shows that with reluctance, $q_i^t = 1$ for all $i \in R$ and $q_i = 0$ for all $i \in P$ in equilibrium. Proposition 2a) shows that the morally ideal contribution e_{ij}^{t*} for i prescribed by j's ethical view q_j^t is decreasing in q_j^t for $i \in R$, and increasing in q_j for $i \in P$. Proposition 1a-b), together with eqs. (6), (6') and (6''), show that contributions e_i^t are increasing in e_{ii}^{t*} as well as e_{iG}^{t*} (the relation may be weak for some individuals, ref. eqs. (6') and (6'')). Thus, since e_i^t for $i \in R$ is decreasing in q_i^t as well as in q_G^t, the minimum e_i^t from a rich individual $i \in R$ occurs when $q_i^t = q_G^t = 1$. Similarly, since e_i^t for $i \in P$ is increasing in q_i^t as well as in q_G^t, the minimum e_i^t from a poor individual $i \in P$ occurs when $q_i^t = q_G^t = 0$. Thus, in the steady state all individuals are at their lowest possible actual contribution levels e_i^t. Hence the equilibrium is an absolute contribution minimum. ■

10. Reducing polarization through contact across neighborhoods

A key driver of the polarization result above is the assumption that individuals learn their ethical views from peers in their own neighborhood. If there is contact between social groups, however, some learning of ethical views is likely to take place even between neighborhoods, limiting the ethical polarization. Thus, policies stimulating social contact between groups – for example, encouraging their joint participation and encounters in public debate, making kids from diverse social neighborhoods attend the same schools, or encouraging attendance in shared cultural experiences – could help reduce polarization. This mirrors the conclusion of Benabou et al. (2018) that more mixed interaction reduce polarization and raise prosociality (since extreme polarization yields minimal public good contributions).

As a point of departure, consider the equilibrium above with complete polarization and segregation. Assume that a policy stimulating contact between neighborhoods is introduced in this situation. This will not have any direct effect on migration, but will cause social learning of ethical views to partly take place between neighborhoods.

Recall that in Eq. (20), we modelled the change over time in an individual's ethical view q_i^t as the sum of two parts: the unbiased learning part $\delta\left(\tilde{q}_{ji}^t - q_i^t\right)$, which is proportional to the difference between one's own view and the perceived view of the random group member one meets, plus a term reflecting reluctance. Now, assume instead that with probability κ, the other individual j is from the other neighborhood ($G_i \neq G_j$), while with probability $(1-\kappa)$ the other is from one's own neighborhood ($G_i = G_j$). This affects the dynamics of social learning, giving instead

$$(29) \quad \dot{q}_A^t = \delta([(1-\kappa)q_A^t + \kappa q_B^t)] - q_A^t) + R_A^t.$$

Let A be the rich social group. Then, when moving to continuous time, incorporating reluctance, and adding the direct effect of migration on average ethical views in the group (see eq. (27) and the discussion thereof), the equilibrium condition becomes

$$(30) \quad \dot{q_A^t} = \delta\kappa(q_B^t - q_A^t) + R_A^t \ = 0.$$

In equilibrium, with A as the rich group, $s_A^t = 1$. Thus, in equilibrium

$$(31) \quad \kappa\delta(q_A^t - q_B^t) = rR_A^t \,.$$

This rules out $q_A^t = 1$ and $q_B^t = 0$, since if $q_A^t = 1$ we must have $R_A^t = 0$. Consequently, we no longer get complete polarization. The strength of reluctance approaches zero as $q_A^t \to 1$; hence, at some point before we get to $q_A^t = 1$, the effect of meeting people in the other group will cancel out the effect of reluctance. As a result, equilibrium polarization is limited.[18] We summarize this as a Proposition.

Proposition 6. If social learning takes place partly between neighborhoods, and the share of social learning from the other neighborhood is κ, then polarization is incomplete in equilibrium: $q_A - q_B < 1$.

If equilibrium polarization is indeed incomplete, the equilibrium no longer represents an absolute contribution minimum. Intuitively, with social learning even between neighborhoods, individuals occasionally meet people who demand more of them than their own peers do, making them more prone to contribute somewhat higher amounts to the public good.

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

11. Government responsibility

Above, public good provision was assumed to be a private responsibility: there was no public sector, nor any other type of collective action. Since the segregation and polarization processes we describe are triggered by the burden of moral responsibility, similar results do not necessarily arise if public good provision is left to the government.

To see this, consider first the case where – partly outside of our model – individuals view contributing voluntarily to be a moral responsibility only as long as government public good supply falls short of the provision level individuals themselves consider socially optimal. In Section 5, we saw that everyone agrees that the socially optimal total public good supply is $E^* = (\sum_{j=1}^{N} Y_j) - \frac{1}{N\gamma}$. Assume now that government policies reflect this unanimous view in the sense that government provision equals E^*. In this case, neither segregation nor polarization would arise: If individuals do not consider private public good provision a moral requirement when public good provision is already socially optimal, then for every i and j, we would have $e_{ii}^{t*} = e_{ji}^{t*} = 0$. As a result, there would be no reason for reluctance: no change in ethical view would increase one's moral burden. Similarly, there would be no reason for social migration: no peers would impose a strictly positive moral burden on anyone else. Hence, no segregation process would take place; polarization would be limited to the fact that within each initial social group, individuals' ethical views would gradually converge to the initial group average, but no force would be present pulling ethical views in more extreme directions.

Keeping strictly to our model, however, we may in fact have that for some i and j, $e_{ji}^{t*} \neq 0$ even when government provision equals E^*. The reason is that if public good provision is financed by taxes, some individuals may consider their own tax payments unfairly low, thus finding that contributing even in excess of E^* is more socially responsible than spending all their income on own consumption. This would result in segregation according to income, but only a very limited degree of polarization.

To analyze this more formally, let us now replace eq. (1) by

$$(32) \quad Y_i = c_i^t + T_i^t + \epsilon_i^t$$

where T_i^t is the tax levied on individual i in period t, wheras $\epsilon_i^t \geq 0$ is i's voluntary contribution to the public good over and above her tax payments. Further, replace eq. (2) by

$$(33) \quad E^t = E^{Pol,t} + \sum_{i=1}^{N} \epsilon_i^t$$

where $E^{Pol,t}$ is the government supply of the public good in period t. Let the government budget be balanced in each period:

$$(34) \quad E^{Pol,t} = \sum_{j=1}^{N} T_i^t$$

Disregard the possibility of side payments between individuals, as well as any administrative or efficiency costs of taxation. Assume, furthermore, that the political process is such that actual policies reflect an ethical view q_m^t, corresponding to the degree of status quo support of an actual or hypothetical voter m (for example, m may be the median voter). If voting is costless and no voter expects to be pivotal, there is no self-interest incentive to vote in a particular way, in which case it seems plausible that individuals would vote according to their ethical views (rather than narrowly self-interested preferences). This yields

$$E^{Pol,t} = E^* = (\sum_{j=1}^{N} Y_j) - \frac{1}{N\gamma} \tag{35}$$

and

$$T_i^t = e_{im}^{t*}. \tag{36}$$

Here, e_{im}^{t*} is the voluntary contribution that would have been morally ideal for individual i in the absence of public provision, as judged by voter m in period t.

The morally ideal voluntary contributions, according to i, is now found by maximizing W_i^t, as given by eq. (12), with respect to ϵ_j^t, $j = 1, \dots, N$, given $u(c_i^t) = \ln c_i^t$ and eqs. (32) – (36), considering $E^{Pol,t}$ and the tax distribution exogenous. Inserting in eq. (12) yields

$$W_i^t = \sum_{j=1}^{N} \mu_{ij}^t \left(\ln(Y_j - \epsilon_j^t - T_j^t) + \gamma[(\sum_{j=1}^{N} Y_j) - \frac{1}{N\gamma} + \sum_{i=1}^{N} \epsilon_j^t] \right).$$

Maximizing this with respect to ϵ_j^t, using that $\sum_{j=1}^{N} \mu_{ij}^t = 1$, gives the following first order condition for an interior welfare maximum:

$$\frac{\partial W_i^t}{\partial \epsilon_j^t} = -\frac{\mu_{ij}^t}{Y_j - \epsilon_j^t - T_j^t} + \gamma N = 0$$

Rearranging, and letting the notation reflect that the result of this exercise is the morally ideal voluntary contribution for j as judged by i, we get

$$\epsilon_{ij}^{t*} = Y_j - \frac{\mu_{ij}^t}{\gamma N} - T_j^t.$$

Now, inserting $T_i^t = e_{im}^{t*}$ from eq. (36), and $e_{jm}^{t*} = Y_j - \frac{\mu_{im}^t}{N\gamma}$ from eq. (13), we get

$$\epsilon_{ij}^{t*} = \frac{\mu_{im}^t - \mu_{ij}^t}{\gamma N}.$$

This is strictly positive whenever $\mu_{im}^t > \mu_{ij}^t$, and zero otherwise (due to the assumption that $\epsilon_i^t \geq 0$).

Note first that with $j = i$, this implies that if voter m's ethical view gives i less weight than i does herself ($\mu_{im}^t < \mu_{ii}^t$), i finds it ethically appropriate not to contribute more than her taxes. Similarly, with $j \neq i$, it follows that i's social group will only think that i should contribute beyond her taxes if $\mu_{im}^t > \mu_{iG}^t$. This yields the following proposition:

Proposition 7. *Any state where all the rich are in one group and all the poor are in the other, and in which all the rich hold the same view q_R while all the poor hold the same view q_P, with $q_R > q_m > q_P$, is a steady state.*

Proof: Recall from eqs. (27) and (28) that for the rich (assumed to be in group A, and thus $q_A = q_R$, and using (24)),

$$\dot{q}_A = R_A^t - (q_{PB}^t - q_{RA}^t)\dot{s}_A. \tag{37}$$

Similarly, for the poor group, B,

$$\dot{q_B} = R_B^t + (q_{PA}^t - q_{RB}^t)\dot{s}_B. \tag{38}$$

With no migration the last terms would vanish, because $\dot{s}_G = 0$. Furthermore, without reluctance $R_A^t = R_B^t = 0$. Thus, with no reluctance and no migration we would have a steady state with $\dot{q}_A = \dot{q}_B = \dot{s}_A = \dot{s}_B = 0$.

Note now that if we have a strict inequality $\mu_{ii}^t > \mu_{im}^t$, then $\epsilon_{ii}^{t*} = 0$ also after a slight change in i's ethical view: A marginal change in μ_{ii}^t will not change the fact that $\epsilon_{ii}^{t*} = 0$, irrespectively of the direction of the change. Thus, in this case there is no cost in terms of a heavier moral responsibility – and thus no reason for reluctance – related to changing one's ethical view.

If $\mu_{ii}^t < \mu_{im}^t$, on the other hand, $\epsilon_{ii}^{t*} > 0$. However, since by assumption $q_R > q_m > q_P$, $\mu_{ii}^t > \mu_{im}^t$ holds, in fact, for everyone: For the rich, μ_{ii}^t is increasing in q_i, while the reverse holds for the poor.

Hence there is no reluctance in the steady states described in the Proposition. Also, there is no migration: For each i, whether rich or poor, $\epsilon_{iG}^{t*} = 0$ is i's own group while $\epsilon_{iG}^{t*} > 0$ in the opposite group.

Note that there is a complication in interpreting m as the *actual* median voter: if the two groups are equally large, the median voter is not well defined. Furthermore, since the distribution of voter preferences will be discrete, then if some point in the voter distribution is defined as the median, we would either have $q_R = q_m$ or $q_m = q_P$, while the Proposition requires $q_R > q_m > q_P$. The situation where $q_R > q_m > q_P$ could still result from several mechanisms: First, just like i observes j' view imperfectly, politicians may also have an imperfect assessment of the electorate, perceiving the view of group A as distributed continuously with mean q_A, and similarly for group B. With $q_A > q_B$, the *perceived* median voter will then be in-between the view held in each group $q_A > q_m^t > q_B$, with strict inequalities. Alternatively, just like voters have ethical views they care about, the same may hold for politicians/political parties: for example, one party may prefer the position of the rich social group, q_R, while the other party might prefer the position of the poor group, q_P. If parties care partly about keeping to their preferred ethical position but also partly about the probability of being elected, competing political parties may commit to policies deviating from their preferred positions towards the median of the voter distribution. If so, the winning party will have a position q_m such that $q_R > q_m > q_P$.

The implication of Proposition 7 is that government supply of public goods limits the drive towards ethical polarization. Imagine an initial situation where ethical views and income are uncorrelated. If one group, say B, is slightly more egalitarian than the other, the poor will migrate toward B, and social learning in B goes slightly more towards egalitarianism than social learning in A. Moreover, some poor may initially hold less egalitarian views, inducing them to make positive voluntary contributions, making their moral responsibility a burden. Their reluctance will then also pull them towards a more egalitarian view. Similar processes apply to the rich. But as soon as the poor have moved to social group B and have become on average more egalitarian than q_m^t, corresponding to the current policy, the process stops, and there is no further movement in the groups' average ethical views. Thus, the difference in ethical views between poor and rich can be quite small.

Note finally that with small differences in ethical views between the groups, the incentives to migrate are small. If we introduced an additional term in the utility function allowing individuals to have a small exogenous but heterogeneous preference for one of the groups, then segregation might be incomplete, and the ethical view difference between groups would be even smaller.

12. Conclusions

Moral concerns can motivate people to contribute voluntarily to public goods, such as environmental quality, even when doing so is not in their narrow self-interest. As we have demonstrated above, however, the burden of moral responsibility may trigger long-run social processes that not only undermine individuals' willingness to contribute, but also make them self-select into strongly income segregated and highly polarized social environments. In the steady state of our model, all individuals are convinced that their own fair public good contribution is relatively low – and are surrounded by peers confirming this view. Consequently, voluntary contributions to the public good are minimal.

These results arise although we assume throughout that individuals are morally motivated, and that they cannot simply choose the ethical views that fit their interests best. Furthermore, we abstract from several mechanisms potentially causing segregation according to income, such as property prices, status-seeking and conspicuous consumption. It is, rather, the burden of moral responsibility itself that tends to trigger social mechanisms that partly relieve individuals of this burden – while causing segregation and polarization as by-products in the process.

Nevertheless, we also find that the ethical polarization, and thus the downwards drive of contributions, may be limited through policies increasing contact between social groups. Perhaps more surprisingly, we show that if public good provision is left to the government, the drive towards polarization can be considerably weaker.

It seems reasonable to expect strong polarization and segregation to increase the risk of social problems such as mistrust, miscommunication, social unrest, crime, and violent conflict. Hence, although evolution may possibly have endowed humans with a preference for moral behavior (Alger and Weibull 2016; Alger et al. 2020), it is interesting and somewhat disturbing to note that even in a population consisting entirely of morally motivated individuals, active policies may be required to prevent a divided and conflict-stricken society with minimal public good supply.

References

Adler, M.D, and O.F. Norheim (Eds.) (2022): *Prioritarianism in Practice*. Cambridge, UK: Cambridge University Press.

Alesina, A., and G.-M. Angeletos (2005): Fairness and Redistribution, *American Economic Review* 95(4), 960-980.

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.

Alger, I., and J.W. Weibull (2016): Evolution and Kantian morality, *Games and Economic Behavior* 98, 56-67.

Alger, I., J.W. Weibull, and L. Lehmann (2020): Evolution of preferences in structured populations: Genes, guns, and culture, *Journal of Economic Theory* 185, 104951.

Amasino, D., D. Pace, and J. van der Weele (2024): Fair Shares and Selective Attention, *American Economic Journal: Microeconomics*, forthcoming.

Aronson, E., T.D. Wilson, R.M. Akert (2005): *Social Psychology* (fifth edition), New Jersey: Pearson Education International.

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 (2006a): Incentives and prosocial behavior, *American Economic Review* 96 (5), 1652-1678.

Benabou, R., and J. Tirole (2006b): Belief in a Just World and Redistributive Politics, Quarterly Journal of Economics 121(2), 699-746.

Bergstrom, T., Blume, L. and Varian, H. (1986): On the private provision of public goods. *Journal of Public Economics* 29, 25–49.

Besley, T., and T. Persson (2023): The Political Economics of Green Transitions, *Quarterly Journal of Economics*, qjad006, https://doi.org/10.1093/qje/qjad006.

Bilodeau, M., and N. Gravel (2004): Voluntary provision of a public good and individual morality, *Journal of Public Economics* 88, 645–666.

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

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., Lurås, H., Nyborg, K. (1996): Allowing Disagreement in Evaluations of Social Welfare, *Journal of Economics* 63, 303-324.

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.

Bruvoll, A. and K. Nyborg (2004): The Cold Shiver of Not Giving Enough: On the Social Cost of Recycling Campaigns, *Land Economics* 80 (4), 539-549.

Chetty, R., J.N. Friedman, E. Saez, N. Turner, and D. Yagan (2020): Income Segregation and Intergenerational Mobility Across Colleges in the United States, *Quarterly Journal of Economics* 135(3), 1567–1633.

Chetty, R., Jackson, M.O., Kuchler, T. et al. (2022): Social capital I: measurement and associations with economic mobility. *Nature* 608, 108–121.

Cohn, A., L.J. Jessen, M. Klašnja, P. Smeets (2023): Wealthy Americans and redistribution: The role of fairness preferences, *Journal of Public Economics* 225(C).

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

Di Tella, Rafael, Sebastian Galiant, and Ernesto Schargrodsky (2007): The Formation of Beliefs: Evidence from the Allocation of Land Titles to Squatters, Quarterly Journal of Economics 122 (1), 209–241.

Eika, L., M. Mogstad, and B. Zafar (2019): Educational Assortative Mating and Household Income Inequality, *Journal of Political Economy* 127 (6), 2795-2835.

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

Fehr, D., and M. Vollman (2022): Misperceiving Economic Success: Experimental Evidence on Meritocratic Beliefs and Inequality Acceptance, CESifo Working Paper 9983/2022.

Festinger, L. (1957): *A Theory of Cognitive Dissonance*. Stanford University Press, Stanford, CA.

Gangadharan, L., N. Nikiforakis, M. C. Villeval (2017): Normative Conflict and the Limits of Self-Governance in Heterogeneous Populations, *European Economic Review* 100, 143-156.

Harsanyi, J.C. (1955): Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility, *Journal of Political Economy* 63 (4), 309-321.

Hart, William, Dolores Albarracín, Alice H. Eagly, Inge Brechan, Matthew J. Lindberg, Lisa Merrill (2009): *Psychological Bulletin*, 135 (4), 555–588.

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

Laffont, J.J. (1975): Macroeconomic constraints, economic efficiency and ethics: An introduction to Kantian economics, *Economica* 42 (168), 430– 437.

Lobeck, M., and M.N. Støstad (2023): The Consequences of Inequality: Beliefs and Redistributive Preferences. CESifo Working Paper No. 10710.

Koch, C., N. Nikiforakis, C.N. Noussair (2021): Covenants Before the Swords: The Limits to Efficient Cooperation in Heterogeneous Groups, *Journal of Economic Behavior and Organization* 188, 307-321.

Konow, J. (2000): Fair shares: Accountability and cognitive dissonance in allocation decisions, *American Economic Review* 90 (4), 1072–1091.

Negishi, T. (1960): Welfare economics and existence of an equilibrium for a competitive economy, *Metroeconomica* 12 (2-3), 92- 97.

Nikiforakis, N., C.N. Noussair, T. Wilkening (2012): Normative Conflict and Feuds: The Limits of Self-Enforcement, *Journal of Public Economics* 96, 797-807.

Nyborg, K. (2000): Homo Economicus and Homo Politicus: Interpretation and Aggregation of Environmental Values, *Journal of Economic Behavior and Organization* 42/3, 305-322.

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

Rawls, J. (1971): *A theory of justice. Cambridge, MA: Belknap Press of Harvard University Press.*

Roemer, J.E. (2015): Kantian optimization: A microfoundation for cooperation, *Journal of Public* Economics 127, 45-57.

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.

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^t for the rich while decreasing it for the poor. This drives the group average q_G^t towards zero if the majority is poor but towards 1 if the majority is rich. However, since all the rich (poor) within a given group are subject to the same dynamic, they become increasingly homogenous over time. The latter force could, if the biases of the rich and the poor are sufficiently different, imply the existence of steady states in which the rich and poor within a fixed group hold different views. The purpose of the present Appendix is to explore the conditions under which such steady states may exist.

Let us simplify notation by suppressing the explicit notation of social group G and time t, writing $s_G = s$ for the share of rich individuals in the group. Without loss of generality, let $\delta = 1$, which only affect the speed of convergence but not the direction or state to which it converges. Furthermore, let us now make the following more specific assumption on the probability distribution of $\tilde{q}_{ji}$:

Assumption A1. If $q_j \leq 0.5$, we assume that $\tilde{q}_{ji} = 0 \; or \; 2q_j$ with equal probability. Similarly, if $q_j \geq 0.5$, $\tilde{q}_{ji} = 1 \; or \; 2q_j - 1$ with equal probability.

That is, although individuals are capable of judging whether another leans towards utilitarianism or status quo support, they are less capable of judging precisely how extreme the other's view is. Assumption A1 maximizes the variance of $\tilde{q}_{ji}$ given the assumptions already made about this distribution (Section 6), i.e., that is has support on [0,1] and is unbiased. Although the formal proof below only holds for this specific distribution, note that by maximizing the variance, we allow biases and thus also differences in biases to become substantial, which is what drives the potential for steady states in which $q_R \neq q_P$. In Proposition 3 below, we show that given Assumption A1, then as long as $r < \frac{1}{2}$ and $s \neq \frac{1}{2}$, there are no steady states for which $q_R \neq q_P$ within a fixed group.

To see this, note first that over time, the views of the rich within each group converge, and the same is true for the poor. From eq. (20), if $i \in R$ meets an individual j, then

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

Consider now two different individuals $i \in R, i' \in R$, with $q_i^t > q_{i'}^t$, meeting the same j. Then

$$(A2) \quad \Delta(q_i - q_{i'}) = \left(\tilde{q}_{ji} - q_i\right)\Delta t + r\left(\tilde{q}_{ji} - q_i\right)^{-}\Delta t - \left(\tilde{q}_{ji'} - q_{i'}\right)\Delta t - r\left(\tilde{q}_{ji'} - q_{i'}\right)^{-}\Delta t).$$

Using $(A2)$ and that the probability distribution for $\tilde{q}_{ji}^t$ is the same for all i (Assumption A1), and moving to continuous time and infinite population (cancelling out the randomness in who meets who), we are left with

$$(A3) \quad |\dot{q}_i - \dot{q}_{i'}| < -(1 - r)|(q_i - q_{i'})|.$$

Since the same argument applies to poor individuals, we have the following result:

Lemma A1: *Over time, q_i converges to a common view q_R for all rich (all $i \in R$) within a fixed group, and to a common view q_P for all the poor (all $i \in P$) within the group.*

To consider asymptotically stable states, we can thus limit attention to the case where all the poor within a given group hold exactly the same view q_P, while all the rich hold the same view q_R. Under this assumption, the dynamic of the view of rich and poor, respectively, are (eqs. (21) and (22) in the main text):

$$(A4) \quad \dot{q}_R = (1-s)(q_P - q_R) + srB_{RR}^{-} + (1-s)rB_{RP}^{-}$$

and

$$(A5) \quad \dot{q}_P = s(q_R - q_P) - srB_{PR}^{+} - (1-s)rB_{PP}^{+}.$$

To fix ideas, note that if the biases introduced by reluctance were of equal strength for rich and poor, q would be increasing if $s > \frac{1}{2}$, implying that the only asymptotically stable state would be $q = 1$. Similarly, if $s < \frac{1}{2}$, q would gradually decrease, making $q = 0$ the only asymptotically stable state. However, given Assumption A1, the biases are unlikely to be equally strong. We thus need to explore whether differences in the strength of the biases for the rich and the poor, respectively, can potentially change the above simple conclusion.

The following Proposition summarizes the main result on this. Note that part of the proof is demonstrated in the more detailed analysis further down, considering several individual special cases.

Proposition 3. *Given Assumption A1, the following holds:*

- *V.* *For all values of s, a) $q_R = q_P = 0$ and b) $q_R = q_P = 1$ are stable states where $\dot{q}_R = \dot{q}_P = \dot{q}_G = 0$.*
- *VI.* *If $s < \frac{1}{2}$, only a) is asymptotically stable, while for $s > \frac{1}{2}$, only b) is asymptotically stable.*
- *VII.* *For $s = \frac{1}{2}$, there is also a symmetric stable state with $q_R > \frac{1}{2} > q_P$ further characterized by $q_R = q_P + \frac{r}{2}$.*
- *VIII.* *Assuming $r < \frac{1}{2}$, there are no further stable states.*

Proof: Part I is trivial, as there will be no misperception of views in this case, thus no bias and hence no change in views. For $s < \frac{1}{2}$, claim II is Proposition A1 below. The claim for $s > \frac{1}{2}$ follows by symmetry, as we can define $s' = 1 - s$ as the share of poor, and $q' = 1 - q$ as the degree of utilitarianism and get similar expressions for the case $s' < \frac{1}{2}$. Claim III is Proposition A2 below. The expression for the different biases depends on the level of q_R and q_P, with several different cases. To prove the last claim, IV, we need to discuss each possible case. This is done below.

Note that the specification of the probability distribution depends on q_j being above or below 0.5. Further, the distribution has two points, and we need to consider separately the cases where both or only one invoke reluctance. This leaves us with several different cases we need to consider separately.

The first case: $0 < q_P < q_R < 0.5$.

Consider first the learning process of reluctant poor individuals, determining the development over time in q_P. Note that, given Assumption A1 and that $0 < q_P < q_R < 0.5$, the view of another person j will be perceived as 0 with 50% probability, regardless of whether j is rich or poor. This will not invoke reluctance for a poor observer i. Only the other case invokes reluctance, hence

$$B_{PR}^{+} = \frac{1}{2}(2q_R - q_P)$$

$$B_{PP}^{+} = 2q_P - q_P = \frac{1}{2}q_P$$

When we insert this in the equation (A.5) we get

$$(A6) \quad \dot{q}_P = s(q_R - q_P) - sr\tfrac{1}{2}(2q_R - q_P) - (1-s)\tfrac{1}{2}rq_P.$$

To have Note that $\dot{q}_P$ is declining in q_P, hence the larger q_P, the stronger is the pull toward 0. On the other hand, $\dot{q}_P$ is increasing in q_R: the more different q_P and q_R are, the stronger is the pull to make them more equal.

Using eq. A6, requiring $\dot{q}_P = 0$ and collecting terms for the rich and poor, we get

$$(A7) \quad 2s(1-r)q_R = 2s(1-r)q_P + rq_P \Longrightarrow q_R = \left(1 + \tfrac{r}{2s(1-r)}\right)q_P.$$

Next, we turn to the learning dynamics for the rich. Consider first $q_R < 2q_P$.

Note that Assumption A1 and $q_R < 2q_P$ imply that meeting a poor j and perceive their view as $2q_P$ does not produce reluctance for a rich i. Hence the rich are reluctant only if they perceive the person they meet to hold a view of $q_j = 0$. This happens with probability 50%, thus $B_{RR}^{-} = B_{RP}^{-} = \frac{1}{2}q_R$. Thus

$$(A8) \quad \dot{q}_R = (1-s)(q_P - q_R) + r\tfrac{1}{2}q_R = (1-s)q_P - \left(1 - s - \tfrac{r}{2}\right)q_R.$$

Note also in this case that $\dot{q}_R$ is declining in q_R and increasing in q_P, so the pull toward zero is stronger the higher q_R is.

Combining (A7) and (A8), we see that when $\dot{q}_P = 0$ then

$$(A9) \quad \dot{q}_R = \left(1 - s - \left(1 - s - \tfrac{r}{2}\right)\left(1 + \tfrac{r}{2s(1-r)}\right)\right)q_P = \left(\tfrac{r}{2}\left(1 + \tfrac{r}{2s(1-r)}\right) - \tfrac{r(1-s)}{2s(1-r)}\right)q_P$$

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

We note that for $q_P > 0$, then $\dot{q}_R < 0$ if $s(1-r) < 1 - s - \frac{r}{2}$ or equivalently $2s < 1 + \left(s - \frac{1}{2}\right)r$. Collecting terms with s on the left hand side we get $(2-r)s < \frac{1}{2}(2-r)$ or simply $s < \frac{1}{2}$, which is true by assumption. Thus, if $\dot{q}_P = 0$ then $\dot{q}_R < 0$, so there cannot be any stable state with $q_P > 0$.

Finally, we need to check that the premise $\dot{q}_P = 0$ is consistent with the assumption that $q_R < 2q_P$. From (A7) we get that $\dot{q}_P = 0$ when $q_R = \left(1 + \frac{r}{2s(1-r)}\right) q_P$ and $q_R < 2q_P$ are both true when

$$q_R = \left(1 + \frac{r}{2s(1-r)}\right) q_P < 2q_P \Leftrightarrow s > \frac{r}{2(1-r)}$$

We summarize this as a lemma:

Lemma A2: If $s > \frac{r}{2(1-r)}$ then $\dot{q}_R < 0$ when $\dot{q}_P = 0$. $\dot{q}_R$ is declining in q_R and increasing in q_P, while the opposite holds for $\dot{q}_P$.

Proof: This follows from the calculations above. Note that the constraint on $2s$ follows from the assumption $2q_P \geq q_R$.

The lemma shows that there cannot be a case where the reluctance of rich and poor keep each other in balance when $2q_P > q_R$.

Next we turn to the dynamics for the rich when $q_R > 2q_P$.

Note that Assumption A1 and $q_R > 2q_P$ imply that rich are always reluctant if they meet a poor. They are also reluctant when they meet a rich and perceive this person to hold a view of 0. This happens with probability 50%, thus $B_{RR}^{-} = \frac{1}{2} q_R$, while $B_{RP}^{-} = (q_R - q_P)$. Thus

$$(A10) \quad \dot{q}_R = (1-s)(1-r)(q_P - q_R) + sr\frac{1}{2}q_R = (1-s)q_P - \left(1 - s - \frac{r'}{2}\right) q_R$$

where $r' = \frac{sr}{1-r}$. Note that the expression for $\dot{q}_R$ is identical to the case $q_R > 2q_P$ except with r' rather than r. We can thus use the calculations above where we found $\dot{q}_R = (1-s)q_P - \left(1 - s - \frac{r'}{2}\right) q_R$. Next, we combine this with the condition (A7) for $\dot{q}_P = 0$, stating that $q_R = \left(1 + \frac{r}{2s(1-r)}\right) q_P$. This yields:

$$(A11) \quad \dot{q}_R = \left(1 - s - \left(1 - s - \frac{r'}{2}\right)\left(1 + \frac{r}{2s(1-r)}\right)\right) q_P = \left(\frac{rs}{2(1-r)}\left(1 + \frac{r}{2s(1-r)}\right) - \frac{r(1-s)}{2s(1-r)}\right) q_P$$

$$= \frac{r}{2(1-r)}\left[s + \frac{r}{2(1-r)} - \frac{(1-s)}{s}\right] q_P$$

We note that $\dot{q}_R < 0$ if $1 - s > s^2 + \frac{sr}{2(1-r)}$.

Next, we check that the premise $\dot{q}_P = 0$ is consistent with the assumption that $q_R > 2q_P$. As above we find that this is possible only for $s < \frac{r}{2(1-r)}$. Using this condition, we see that the last parentheses in (A11) satisfy

$$s + \frac{r}{2(1-r)} - \frac{(1-s)}{s} < 2s - \frac{(1-s)}{s}$$

which is negative for all $s < \frac{1}{2}$, thus for $s < \frac{r}{2(1-r)}$ in particular.

Lemma A3: If $s < \frac{r}{2(1-r)}$, then $\dot{q}_R < 0$ when $\dot{q}_P = 0$.

We have now covered all cases with $0 < q_P < q_R < 0.5$, establishing that $\dot{q}_R < 0$ when $\dot{q}_P = 0$.

Note that the dynamic equations are linear in q_P and q_R. Hence they can be written in the form $\dot{q}_P = a\Delta q - bq_P$ and $\dot{q}_R = -c\Delta q + dq_P$, where $\Delta q = q_R - q_P$. When Δq is sufficiently large, q_P will increase and q_R will decline, hence reducing Δq. By Lemma A2 and A3, q_R is decreasing when Δq is decreasing even at the point where $\dot{q}_P = 0$. Hence Δq will decrease until $\dot{q}_P < 0$ and the system will approach the stable state $q_P = q_R = 0$. We conclude that

Proposition A1: For $s < \frac{1}{2}$, $q_P = q_R = 0$ is an asymptotically stable state.

In the remainder we consider the possibility of other stable states with $0.5 < q_R$.

The second case: $0 < q_P < 0.5 < q_R$.

Consider first the learning process of the poor. Now reluctance when a poor meets a rich depends on whether $q_P < \tilde{q}_{ji}$ in both possible realizations of the perception, that is whether $2q_R - 1 > q_P$ or not.

$$(A12)\quad B_{PR}^{+} = \begin{cases} \frac{1}{2}(1-q_P) & if\ 2q_R - 1 < q_P \\ (q_R - q_P) & if\ 2q_R - 1 > q_P \end{cases}$$

$$(A13)\quad B_{PP}^{+} = 2q_P - q_P = \frac{1}{2}q_P$$

Consider the case $2q_R - 1 < q_P$. When we insert the first case of (A12) and (A13) in eq. (A5) we get

$$(A14)\quad \dot{q}_P = s(q_R - q_P) - s\frac{r}{2}(1-q_P) - (1-s)\frac{r}{2}q_P = s(q_R - q_P) - s\frac{r}{2} - \frac{1}{2}(1-2s)rq_P.$$

We first consider the conditions for no movement in the view of the poor, that is $\dot{q}_P = 0$. Collecting terms for rich and poor we get

$$(A15)\quad 2sq_R + sr = 2s(1-r)q_P + (1-s)rq_P \Longrightarrow q_R = \left(1 + \frac{r(1-2s)}{2s}\right)q_P + \frac{r}{2}$$

Then consider the subcase $2q_P > q_R$.

In this case the rich are only reluctant if meeting poor whom they perceive to hold the view $q_P = 0$. Thus $B_{RP}^{-} = \frac{1}{2}q_R$ and $B_{RR}^{-} = \frac{1}{2}(1-q_R)$. Inserting in (A4) gives

$$(A16)\quad \dot{q}_R = (1-s)(q_P - q_R) + s\frac{r}{2}q_R + \frac{(1-s)r}{2}(1-q_R).$$

Stability in the view of the rich thus requires:

$$(1-s)q_P = (1-s)q_R - s\frac{r}{2}q_R - \frac{(1-s)r}{2}(1-q_R) = (1-s)q_R - srq_R + \frac{r}{2}q_R - \frac{r}{2}(1-s).$$

So we get

$$(A17)\quad q_P = \left(1 + \frac{r(1-2s)}{2(1-s)}\right)q_R - \frac{r}{2}$$

and from (A15):

(A18) $q_R = \left(1 + \frac{r(1-2s)}{2s}\right) q_P + \frac{r}{2}$.

Note first that for $s = \frac{1}{2}$ then both equations imply $q_R = q_P + \frac{r}{2}$. That is, in a group with equally many rich and poor, there is a stable state where $q_P < 0.5 < q_R$.

Proposition A2: If $s = \frac{1}{2}$ there is a stable state with $q_P < 0.5 < q_R$ and $q_R = q_P + \frac{r}{2}$.

Note that in this stable state, if r is small, then $q_R \approx q_P \approx \frac{1}{2}$. Hence in a group with equal share of rich and poor, everyone will hold a view approximately in the middle, with the poor tending slightly toward a more egalitarian and the rich slightly towards status quo seeking.

Does a similar stable state also exist when the majority is poor? Adding (A12) and (A14), we get

$$q = s\dot{q}_R + (1-s)\,\dot{q}_P$$

$$= s(1-s)(q_P - q_R) + s^2\frac{r}{2}q_R + \frac{s(1-s)r}{2}(1-q_R)$$

$$+ (1-s)s(q_R - q_P) - s(1-s)\frac{r}{2}(1-q_P) - (1-s)^2\frac{r}{2}q_P$$

$$= \frac{r}{2}\left((s^2 q_R - (1-s)^2 q_P) - (s(1-s)(q_R - q_P)\right)$$

$$\leq \frac{r}{2}\left((s(1-s)q_R - s(1-s)q_P) - (s(1-s)(q_R - q_P)\right) = 0.$$

Here the inequality is strict when $s < \frac{1}{2}$, hence under these conditions the average q will be declining, so there can be no stable state.

Consider now the subcase $2q_P < q_R$.

In this case, the rich are always reluctant when meeting someone poor. Thus

$$B^-_{RP} = (q_R - q_P)$$

$$B^-_{RR} = \frac{1}{2}(1-q_R).$$

Inserting in (A4), we get

(A19) $\dot{q}_R = (1-s)(q_P - q_R) + sr(q_R - q_P) + \frac{(1-s)r}{2}(1-q_R)$.

Stability in the view of the rich thus requires:

$$(1-s-sr)q_P = (1-s-sr)q_R - \frac{(1-s)r}{2}(1-q_R)$$

$$q_P = \left(1 + \frac{(1-s)r}{2(1-s-sr)}\right)q_R - \frac{(1-s)r}{2}.$$

Once again, we use the condition from above

$$q_R = \left(1 + \frac{r(1-2s)}{2s}\right) q_P + \frac{r}{2}$$

And combine to

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

We successively simplify

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

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

Again, as $2 + \frac{2s(1-s)}{(1-s-sr)(1-2s)} + \left(\frac{(1-s)r}{(1-s-sr)}\right) > 2$ and $(1-s)r < 1$, this is inconsistent with the assumption that $q_R > \frac{1}{2}$.

As we have checked both possible subcases, we can conclude that

Lemma A4: There is no stable state with $2q_R - 1 < q_P$ and $q_P < 0.5 < q_R$.

We then consider the subcase $2q_R - 1 > q_P$.

Now q_R is so large that the poor are always reluctant when meeting a rich person, hence $B_{PR}^{-} = (q_R - q_P)$. Using (A5) once again, we get

$$(A20) \quad \dot{q}_P = s(q_R - q_P) - srB_{PR}^{+} - (1 - s_G)rB_{PP}^{+} = s(1-r)(q_R - q_P) - \frac{(1-s)r}{2} q_P.$$

Further, consider the subcase $2q_P < q_R$. This is the parallel case where rich are always reluctant when meeting poor. Thus

$$(A21) \quad \dot{q}_R = (1-s)(q_P - q_R) + srB_{RR}^{-} + (1-s)rB_{RP}^{-} = (1-s)(1-r)(q_P - q_R) + \frac{sr}{2}(1 - q_R).$$

Computing the aggregate dynamics, we get

$$q = s\dot{q}_R + (1-s)\dot{q}_P = \frac{r}{2}(s^2(1-q_R) - (1-s)^2 q_P)$$

and thus in a stable state we must have

(A22) $\quad (1-s)^2 q_P = s^2(1-q_R)$

Next set $\dot{q}_R = 0$ in (A19):

(A23) $\quad (q_R - q_P) = \frac{sr}{2(1-s)(1-r)}(1-q_R)$

Using (A22) to eliminate q_P, we get

$$q_R - \frac{s^2}{(1-s)^2}(1-q_R) = \frac{sr}{2(1-s)(1-r)}(1-q_R)$$

$$\left(1 + \frac{s^2}{(1-s)^2} + \frac{sr}{2(1-s)(1-r)}\right) q_R = \left(\frac{s^2}{(1-s)^2} + \frac{sr}{2(1-s)(1-r)}\right)$$

$$q_R = \frac{\left(\frac{s^2}{(1-s)^2} + \frac{sr}{2(1-s)(1-r)}\right)}{\left(1 + \frac{s^2}{(1-s)^2} + \frac{sr}{2(1-s)(1-r)}\right)}.$$

Note that we derive this on the assumption that $q_R > \frac{1}{2}$, thus we need $\frac{s^2}{(1-s)^2} + \frac{sr}{2(1-s)(1-r)} > 1$. We needed $r < \frac{1}{2}$ above, and then $\frac{sr}{2(1-s)(1-r)} < \frac{1}{2}$, hence we require $\frac{s^2}{(1-s)^2} > \frac{1}{2}$,

Note also that we can write

$$2q_P = \frac{2s^2}{(1-s)^2}(1-q_R) = \frac{2\frac{s^2}{(1-s)^2}}{\left(1 + \frac{s^2}{(1-s)^2} + \frac{sr}{2(1-s)(1-r)}\right)}$$

Hence the requirement that $2q_P < q_R$ implies that

$$2\frac{s^2}{(1-s)^2} < \frac{s^2}{(1-s)^2} + \frac{sr}{2(1-s)(1-r)}$$

which is equivalent to

$$\frac{s^2}{(1-s)^2} < \frac{sr}{2(1-s)(1-r)} < \frac{1}{2},$$

which is in turn inconsistent with the requirement $\frac{s^2}{(1-s)^2} > \frac{1}{2}$ we found above.

Thus this case cannot be a stable state.

How about the subcase $2q_P > q_R$? Now, rich are reluctant when meeting poor and perceiving them to hold the view $q_P = 0$. Thus we get no change in the view of rich if

$$(A26)\quad \dot{q}_R = (1-s)(q_P - q_R) + srB_{RR}^{-} + (1-s)rB_{RP}^{-}$$

$$= (1-s)(q_P - q_R) + \tfrac{(1-s)r}{2} q_R + \tfrac{sr}{2}(1-q_R) = 0.$$

Combine this with the condition for stability in the view of the poor:

$$s(1-r)(q_R - q_P) - \frac{(1-s)r}{2} q_P = 0$$

This gives

$$\frac{(1-s)r}{2s(1-r)}q_P = \frac{(1-s)r}{2}q_R + \frac{sr}{2}(1-q_R)$$

which simplifies to

$$q_P = \frac{(1-2s)s(1-r)}{1-s}q_R + s^2(1-r)$$

We need to check that this is consistent with $2q_P > q_R$ and $2q_R - 1 > q_P$. Note that $2q_P \leq \frac{2(1-r)}{3}q_R < \frac{2}{3}q_R$. Thus we must have $s^2 > \frac{2}{3}q_R > \frac{1}{3}$, which is inconsistent with the requirement that $s < \frac{1}{2}$.

The last case: $0.5 < q_P < q_R < 1$.

This case is, by symmetry, similar to the case $q_P < q_R < 0.5$, but with $s > \frac{1}{2}$. This is because we could change variables to $q' = 1 - q$ and $s' = 1 - s$, that is, q' measuring the degree of utilitarianism and s' being the share of poor. All the equations would be the same, except with prime on the variables. We can thus use the same equations and arguments. It follows that if $s > \frac{1}{2}$ then $\dot{q}_R > 0$ whenever $\dot{q}_S = 0$. Thus, there are no stable states. Moreover, it follows that the average group view eventually will be growing forever, using the same argument as above. By symmetry it follows that there are no stable states with $0.5 < q_P < q_R < 1$ when $s < \frac{1}{2}$. Moreover, the average view will eventually be declining. Thus, there is no stable state for $0.5 < q_P < q_R < 1$, and the stable state $q_P = q_R = 1$ is not asymptotically stable.

Appendix 2: on asymptotically stable states in the case with migration

Equations (25) and (26) represent a continuous-time version of the discrete-time eq. (23) (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. (23)). What we are missing is an expression for the change from t'' to t''', reflecting migration between A and B.

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

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

where $q_{\theta G}^t$ is the average status quo support 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 average status quo support differs between A and B. Assume that $q_A^t > q_B^t$. The poor in A who revise their neighborhood affiliation, i.e., $\rho(1 - s_{RA}^{t-1})$, will now move to B; the rich in B who revise, i.e., $\rho(1 - s_{RA}^{t-1})$, will move to A. (Recall that $s_{RA}^{t-1} = 1 - s_{RB}^{t-1} = 1 - s_{PA}^{t-1}$, thus $s_{RB}^{t-1} = s_{PA}^{t-1}$.) Since the ethical updating has already been done, the remaining individuals change neither their ethical views nor their neighborhood affiliation between t'' and t'''. Thus, the change in average ethical views in A between t'' and t''', entirely due to the direct effect of migration, is

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

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

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

Stating this as differential equations and adding the relevant expressions to the direct effect of ethical updating as specified in eqs. (25) and (26), 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:

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

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

We are now equipped to prove Proposition 4:

Proposition 4 (*Extreme segregation and polarization in the long run).*
There are only two asymptotically stable states. In both states, all the rich are located in one neighborhood and hold a completely status quo supportive ethical view, while all the poor are located in the other neighborhood and hold a completely utilitarian view. Since a given neighborhood can either be the rich one or the poor one, there are two asymptotically stable states.

Proof: In a stable state, $\dot{s}_{RA}^t = -\dot{s}_{RB}^t = -\dot{s}_{PA}^t = \dot{s}_{PB}^t = 0$ and $\dot{q}_A^t = \dot{q}_B^t = 0$. Migration adds a term $(q_{PB} - q_{RA})\dot{s}_{RA}$ to (25) and similarly to (26). 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 3 shows that, without migration, assuming A is a group with a poor majority, there is only one stable state: $q_A = 0$. In this case B would be a group with a rich majority, with $q_B = 1$ as the only stable state. Since $q_A = 0$ and $q_B = 1$, then when we allow migration the rich in A will migrate to B, while the poor in B migrate in the other direction. Thus, the only stable state is when all rich are in one group and all poor in the other. Exactly the same argument applies with neighborhoods A and B interchanged. If A has an equal share of rich and poor, there is an additional stable state as discussed in Proposition 3, Part III, in which rich and poor 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: On moral responsibility with government provision

Let us now replace eq. (1) by

$$(A1) \quad Y_i = c_i^t + \epsilon_i^t + T_i^t$$

where ϵ_i^t is i's voluntary contribution to the public good over and above her tax payments. Further, replace eq. (2) by $(A2)$:

$$(A2) \quad E^t = E^{Pol,t} + \sum_{i=1}^{N} \epsilon_i^t$$

where $E^{Pol,t}$ is the government supply of the public good in period t. Assume that both parties X and Y commit to providing E^{t*}, the agreed socially optimal level of the public good (as derived in Section 5):

$$(A3) \quad E^{Pol,t} = E^{t*} = (\sum_{j=1}^{N} Y_j) - \frac{1}{N\gamma}$$

Furthermore, each party commits to a tax financing scheme with a balanced budget in each period,

$$(A4) \quad E^{Pol,t} = \sum_{j=1}^{N} T_i^t$$

where the tax levied on each individual i in period t, T_i^t, is given by

$$(A5) \quad T_i^t = e_{iPol}^{t*}$$

where e_{iPol}^{t*} is the morally ideal contribution by i according to the social welfare function ascribed to by party $Pol = X, Y$, or, in other words, the voluntary contribution that would have been morally ideal for individual i in the absence of public provision, as judged by a j such that $q_j^t = q_{Pol}$.

Now, while every i agrees that E^{t*} is the socially optimal provision of the public good, this result was derived under the assumption that everyone else also contributes *their* morally ideal contributions, given i's subjective social welfare function. Could it be the case that although $E^{Pol,t} = E^{t*}$, individual i still judges the morally ideal voluntary contribution for herself or others, ϵ_{ij}^{t*}, to be strictly positive, since i finds the prevailing tax distribution socially suboptimal?

The morally ideal voluntary contributions, according to i, is now found by maximizing W_i^t, as given by eq. (12), with respect to $\epsilon_j^t, j = 1, \dots, N$, given $u(c_i^t) = \ln c_i^t$ and eqs. $(A1) - (A5)$, considering $E^{Pol,t}$ and the tax distribution exogenous.

Inserting in eq. (12) yields

$$W_i^t = \sum_{j=1}^{N} \mu_{ij}^t \left(\ln(Y_j - \epsilon_j^t - T_j^t) + \gamma[(\sum_{j=1}^{N} Y_j) - \frac{1}{N\gamma} + \sum_{i=1}^{N} \epsilon_j^t] \right).$$

Maximizing this wrt ϵ_j^t, using that $\sum_{j=1}^{N} \mu_{ij}^t = 1$, gives the following first order condition for an interior welfare maximum:

$$\frac{\partial W_i^t}{\partial \epsilon_j^t} = -\frac{\mu_{ij}^t}{Y_j - \epsilon_j^t - T_j^t} + \gamma N = 0$$

Rearranging, and letting the notation reflect that the result of this exercise is the morally ideal voluntary contribution for j as judged by i, we get

$$\epsilon_{ij}^{t*} = Y_j - \frac{\mu_{ij}^t}{\gamma N} - T_j^t.$$

Now, inserting $T_i^t = e_{iPol}^{t*}$ from eq. $(A5)$, and $e_{jPol}^{t*} = Y_j - \frac{\mu_{iPol}^t}{N\gamma}$ from eq. (13), we get

$$\epsilon_{ij}^{t*} = \frac{\mu_{iPol}^t - \mu_{ij}^t}{\gamma N}.$$

This is strictly positive whenever $\mu_{iPol}^t > \mu_{ij}^t$. Thus, for any individual whose interests are given a larger weight by the government' ethical view than the weight implied by i's own view, there is still a perception that the individual should contribute more than her taxes. To understand why, recall that public good benefits are assumed to be linear (to avoid complicating other model calculations). Hence, the (gross) marginal social benefit of public good provision equals γN regardless of the supply level. The marginal utility of consumption, however, is declining in the consumption level.

Thus, if the individual disagrees with the prevailing tax distribution but considers it exogenously given, she cannot help those who (in her opinion) pay too high taxes. For those who pay too little in terms of taxes, however, the social value of their marginal consumption is less than the social benefits of increasing E^t marginally even beyond E^{t*}, and thus i finds that these individuals ought, ideally, to make strictly positive voluntary contributions.

We will not go through the full analysis for this case. Nevertheless, it should be clear that with these assumptions – unlike the simple case of no moral responsibility ascription whenever $E^{Pol} \geq E^{t*}$, used in Section 11 of the main text – segregation and polarization forces will indeed be at play. There will be a reason for social learning to be reluctant, since some ethical views can imply a heavier moral burden for a given individual than others, and individuals will prefer peers who do not demand too much of them. These forces will now be limited, however, by individuals' contributions through the tax system. For example, as long as party X stays in power, a rich person i's moral burden ϵ_{ii}^{t*} does not increase by adopting a more utilitarian ethical view if, even after the change, $q_i^t \geq q_X$ (since i will then still think she contributes too much through tax payments, leaving no moral obligation to contribute voluntarily). Similarly, for a rich voter i, any peer j for whom $q_j^t \geq q_X$ will find that $\epsilon_{ij}^{t*} = 0$ and thus be equally preferred by i. Note that since these limits to the segregation and polarization process depend on the ethical view q_{Pol} of the ruling party, it can now also matter whether and how parties' views change over time. We leave the full analysis of possible steady states of this model for future work.