

Inequality, Identity, and Partisanship: How redistribution can stem the tide of mass polarization

Alexander J. Stewart^{a,1}, Joshua B. Plotkin^b, and Nolan McCarty^{c,1}

^aSchool of Mathematics and Statistics, University of St Andrews, St Andrews, KY16 9SS, United Kingdom; ^bDepartment of Biology, University of Pennsylvania, Philadelphia, PA, USA; ^cSchool of Public and International Affairs, Princeton University, Princeton, NJ, USA

This manuscript was compiled on August 23, 2021

The form of political polarization where citizens develop strongly negative attitudes towards out-party members and policies has become increasingly prominent across many democracies. Economic hardship and social inequality, as well as inter-group and racial conflict, have been identified as important contributing factors to this phenomenon known as “affective polarization.” Research shows that partisan animosities are exacerbated when these interests and identities become aligned with existing party cleavages. In this paper we use a model of cultural evolution to study how these forces combine to generate and maintain affective political polarization. We show that economic events can drive both affective polarization and the sorting of group identities along party lines, which in turn can magnify the effects of underlying inequality between those groups. But on a more optimistic note, we show that sufficiently high levels of wealth redistribution through the provision of public goods can counteract this feedback and limit the rise of polarization. We test some of our key theoretical predictions using survey data on inter-group polarization, sorting of racial groups and affective polarization in the United States over the past 50 years.

Polarization | Inequality | Cultural evolution | Risk aversion

The political polarization of citizens is increasingly a concern throughout the world, as populist movements challenge mainstream parties in efforts to disrupt established institutions and democratic norms (1). Such trends have been especially manifest in the United States, where they have culminated in political violence such as of the Unite the Right rally at Charlottesville and the storming of the Capitol during the certification of the 2020 presidential election (2) (see also (3)).

There has been extensive debate about the nature and causes of mass polarization. The earliest work, focused on the distributions of voter policy preferences, cast considerable doubt as to whether mass polarization was an important phenomenon at all. That work continues to show that the public’s attitudes on policy issues have remained stable and centrist over many decades (4) (but see (5)). This debate is reviewed in (6).

However, two other important facets of mass polarization have been rising. The first is the process of partisan sorting, where the policy preferences and group identities of a voter better align with their partisan attachments (7–9). The second is affective polarization, whereby individuals develop negative attitudes and behaviors towards members of the opposing party (10, 11).

Sorting and affective polarization appear to be strongly related to increasing inter-group conflict. The growth of inter-

group antagonism has been shown to have multiple contributing factors, including economic adversity, racial animus and a range of other socio-economic factors (12–19). Recent work focusing on the cultural evolution of polarization along identity group lines (20) has shown that a rise in economic adversity or inequality can cause polarized behavioral strategies to take hold and become entrenched in a population, even when the adverse conditions that stimulated it are reversed (21).

Despite the important link between partisan sorting and inter-group conflict, there have been few analytical efforts to examine the joint dynamics of these processes. So in this paper, we generalize the framework of Stewart et al. (20) to study the cultural evolution of group polarization and party sorting. In this model, out-group economic interactions are assumed to be more beneficial but more risky than in-group interactions (22–26); and adverse economic environments are assumed to favor risk aversion. We show that when agents attend to both group and partisan identities in choosing interaction partners this stimulates both the evolution of behavioral strategies that polarize along party lines as well as the sorting of group identities along party lines. These behaviors evolve in response to shifts in the economic environment and underlying inequality.

Efforts to mitigate risk aversion in disadvantaged groups via wealth redistribution, in the form of public goods, has the potential to counteract feedback loops that induce polarization.

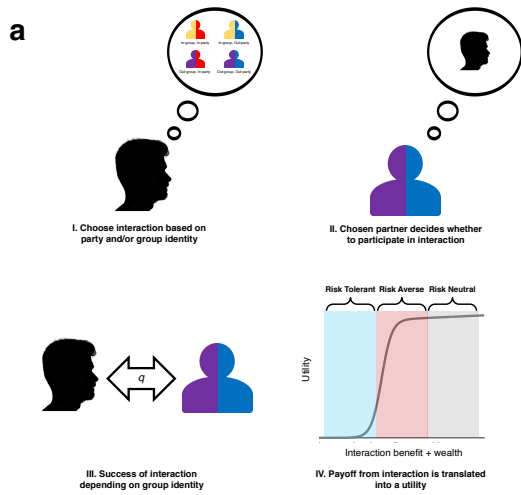
Significance Statement

Populist movements have begun to challenge mainstream political parties, disrupt established norms, and engage in violence against democratic institutions. The movements are fueled by significant support from ordinary citizens who have become increasingly politically polarized. We show that risk-averse attitudes towards other identity groups can transform into affective polarization between supporters of different political parties, through a process of cultural evolution. The economic factors that drive risk aversion can also magnify the effects of unequal wealth, creating a dangerous feedback loop between polarization and inequality. However, redistribution via public goods that reduces inequality can both prevent the onset of political polarization and make it easier for coordinated efforts to reverse entrenched polarized attitudes.

Please provide details of author contributions here.

Please declare any competing interests here.

¹To whom correspondence should be addressed. E-mail: ajs50@st-andrews.ac.uk, nmcarty@princeton.edu



Parameter	Meaning	Default value
B_I	In-group benefit	1
B_O	Out-group benefit	2
q_I	In-group success probability	1
q_O	Out-group success probability	0.6
θ_0	Baseline environment	0.5
$[h, a]$	Utility function slope	[10, 0.02]
r	Public good multiplication factor	1
α	Redistribution rate	0.5
β	Inequality	0.01
γ	Taxation loss	0
χ	Sorting	1

Fig. 1. Model of social interaction and identity. A focal individual (black) engages in beneficial economic interactions. a) I. He first chooses a target for interaction (see Table 1). In general this decision may be based on both party and group identity. II. The chosen target may then agree or not, based on the identity of the focal individual. III. If the pair interacts, the interaction is successful with probability q_I if they share the same identity group, or q_O if they belong to different identity groups. IV. The benefit of a successful interaction is translated into a level of utility that depends on the underlying economic environment (denoted “wealth”) experienced by the focal individual. Depending on that environment, the agent’s utility function may be risk-neutral (grey region), risk-averse (red region) or risk-tolerant (blue region), as described in Eq. 2. b) The table shows default parameters for our analysis, although we vary these systematically in the SI and show that our results are robust to parameter choice.

And yet, we show that low levels of redistribution can actually magnify underlying inequality and entrench polarization. But more optimistically, we also find that sufficiently high levels of redistribution can indeed reduce the impact of inequality and even prevent the emergence of polarization.

Results

To study the cultural evolution of mass political polarization, we generalize a model previously developed to study inter-group polarization and economic interactions (20).

In this model we assume that a large population of individuals is comprised of two distinct identity groups. These identities are assumed fixed, and thus correspond to a fixed feature of identity such as race, religious heritage, or socio-economic background. Although such identities are fixed in the model, the salience of the identity, and therefore its impact on behavior varies.

We assume that members of the population choose to interact with one another, using a one-dimensional strategy p , that describes the probability of choosing an in-group member for an economic interaction, whereas the probability of choosing an out-group member is $1 - p$ (Figure 1 and Table 1). A large literature documents the prevalence of economic discrimination on the basis of group identities. For a recent review and critique, see (27). That difference in partisan identifications can lead to economic discrimination is consistent with the experimental findings in (28, 29).

We assume that out-group interactions are more risky than in-group interactions. In particular, an in-group interaction has success probability q_I , while an out-group interaction has success probability $q_O < q_I$. Successful in-group interactions generate benefit B_I , whereas out-group interactions generate benefit B_O , such that the expected benefit of out-group interactions exceeds that of in-group interactions – i.e. $q_I B_I < q_O B_O$.

This asymmetry, in which out-group interactions generate greater potential benefits than in-group interactions ($B_O > B_I$) may reflect the benefits of cultural or ideological diversity on the quality of joint enterprise, as seen for example in the quality of Wikipedia articles produced by an ideologically diverse team (26). These asymmetries may also be motivated by the expansion of economic gains and opportunities associated with the expansion of markets when groups are willing to trade and work with one another (30).

Finally, we assume that the state of the underlying economic environment, θ , determines the risk profile experienced by individuals as the benefits of their social interactions are translated into utility. The expected utility for a player i is given by

$$w_i = p_i q_I F(B_I + \theta) + p_i (1 - q_I) F(\theta) + (1 - p_i) (1 - \bar{p}_O) q_O F(B_O + \theta) + (1 - p_i) (\bar{p}_O q_O + (1 - q_O)) F(\theta) \quad [1]$$

where \bar{p}_O is the average strategy among out-group members. We have assumed that out-group interactions are only possible if both players are willing to interact with out-group members, whereas in-group interactions are always available (see Table 1 below and Methods). The function F defines an individual’s utility as a function of material payoff x and has the form

$$F(x) = \frac{\exp[hx]}{1 + \exp[hx]} (1 + ax). \quad [2]$$

Here h controls the steepness of the non-linear sigmoid component of the curve and a controls the gradient of the linear component of the curve. This modified ‘S’ shaped utility function allows us to capture changes to risk aversion experienced by individuals as a function of the underlying economic environment, θ . Assuming $a \ll 1$, the utility function F is maximally concave (risk-averse) when $hx \approx \ln \left[\frac{\sqrt{3}+1}{\sqrt{3}-1} \right]$, and is maximally convex (risk-tolerant) when $hx \approx \ln \left[\frac{\sqrt{3}-1}{\sqrt{3}+1} \right]$, and it becomes linear (risk neutral) when $x \gg 0$.

Intuitively our utility function implies there is risk-aversion when the underlying economic environment parameter θ is small but positive. In this regime, which may be thought of as analogous to a risk profile of an individual close to poverty,

failures of economic interactions result in very sharp declines in utility – and so in-group interactions are preferable to the more risky out-group interactions. But when the underlying economic environment is very good ($\theta \gg 0$), risk aversion declines and out-group interactions, which have greater expected returns, are preferable. Finally when the underlying environment is so bad ($\theta < 0$) that a successful economic interaction produces a sharp increase in utility, then risky out-group interactions become strongly preferred.

A version of this model of inter-group interaction has already shown (20) that both high polarization ($p = 1$) and low polarization ($p = 0$) are stable outcomes when the economic environment is strong; but only high-polarization is stable as risk aversion increases. As a result, populations tend to become polarized when the underlying economic environment exogenously declines, and they remain polarized even if the economic environment subsequently improves.

We now generalize the framework to study mass political polarization, in which individuals have a fixed group identity and a sticky, but more malleable, partisan identity. We study how polarization along party lines can emerge as a consequence of risk aversion, as well as the extent to which group identities sort along party lines. We further allow for feedback between individual economic interactions and the overall state of the economic environment. Thus, the environmental dynamic is not exogenous, but rather coupled to partisan identification and individual decisions about economic interactions. This coupling leads to a runaway process that accelerates the rise of polarization and also exacerbates economic inequality through its impact on inter-group interactions.

Model of party identity and social decisions: In order to generalize the model outlined above and to capture the dynamics of mass political polarization, we assume that the population is composed of two identity groups and two political parties. Each individual has both a group identity and a party identity. In general we assume an individual's partisan identity can change while their group identity is fixed. While this is a reasonable assumption for many group identities, others such as religion and ethnicity may switch to match one's partisanship (31, 32). On the flip side, partisanship in the U.S. has been shown to be quite stable at the individual level except in exceptional circumstances such as the realignment of white southerners following Civil Rights (33). But we can interpret party switching in our model as driven by generational change. The risks and benefits of economic interactions between individuals vary by group identity, but we assume that they are independent of party identity.

We now consider two additional decision processes beyond that described above. In these versions, an individual's strategy p depends both on party and group identity. First, we describe a case in which the decision to interact with another player depends on their party identity alone, and second we consider a case in which both group identity and party identity are salient to interaction choices. Importantly, we assume that differences across these models are due to differences in the saliences of group and party identification, not the ability of agents to observe group or party identification. Future work may consider the case where partisanship must be inferred

	Interaction type				
	Same group, same party	Same group, other party	Other group, same party	Other group, other party	
Probability of i choosing interaction	p_i	p_i	$1 - p_i$	$1 - p_i$	} Only Group
Probability of j agreeing to interaction	1	1	$1 - p_j$	$1 - p_j$	
Probability of interaction success	q_I	q_I	q_O	q_O	
Benefit from successful interaction	B_I	B_I	B_O	B_O	} Only Party
Probability of i choosing interaction	p_i	$1 - p_i$	p_i	$1 - p_i$	
Probability of j agreeing to interaction	1	$1 - p_j$	1	$1 - p_j$	
Probability of interaction success	q_I	q_I	q_O	q_O	} Group OR Party
Benefit from successful interaction	B_I	B_I	B_O	B_O	
Probability of i choosing interaction	$2p_i(1 - p_i) + p_i^2$	$p_i + (1 - p_i)^2$	$p_i + (1 - p_i)^2$	$1 - p_i^2$	
Probability of j agreeing to interaction	1	$1 - p_j$	$1 - p_j$	$1 - p_j^2$	
Probability of interaction success	q_I	q_I	q_O	q_O	
Benefit from successful interaction	B_I	B_I	B_O	B_O	

Table 1: Summary of decision processes. We consider three decision processes for a focal individual i choosing an economic interaction. For each decision process, there is a probability of interaction given the identity of a potential target, and a probability of a target j consenting to interaction. If an interaction takes place, the probability of success, and the benefit generated, depend on the identity groups of the pair.

from group identity, or vice versa.

Table 1 summarizes the parameters and interaction probabilities in all three cases. A detailed description of the mathematical model is given in the Methods section below, and further details of its analysis can be found in the SI.

The key differences between the various decision processes summarized in Table 1 involve different probabilities that a player chooses a particular type of interaction, and different probabilities of that interaction being accepted by the other agent. When group identity alone is salient to choice of interaction partners, then out-group members may reject an interaction. When only party identity is salient, then out-party members may reject an interaction. When both group identity and party identity are salient, either an out-group or an out-party member may reject an interaction.

We show that for all three decision processes the dynamics of cultural evolution lead to bi-stability when the underlying economic environment makes individuals risk-neutral or risk-tolerant, with both a high-polarization and a low-polarization equilibrium as stable outcomes, and that this bistability is robust to the choice of parameters (see SI). However, if the environment becomes sufficiently risk-averse, only the high-polarization equilibrium is stable. Thus a population faced with a sufficiently risk-averse environment moves towards a state of high polarization and remains there, even if the underlying environment subsequently improves and risk aversion declines (20). We explore the consequences of these dynamics for sorting of identity groups along party lines, and in the presence of redistribution via public goods.

Sorting: In general, political parties and identity groups may be different in size. However, we make the simplifying assumption that both groups and parties are equal in size. If we denote the proportion of group 1 in party 1 as x_1 and the proportion of group 2 in party 2 as x_2 , the assumption of equal sized groups and parties and groups means $x_1 = x_2 = x$. Under this assumption we can define the degree of sorting of identity groups along partisan lines via the simple expression

$$\chi = 2x - 1 \quad [3]$$

such that $\chi = 0$ corresponds to identity groups distributed

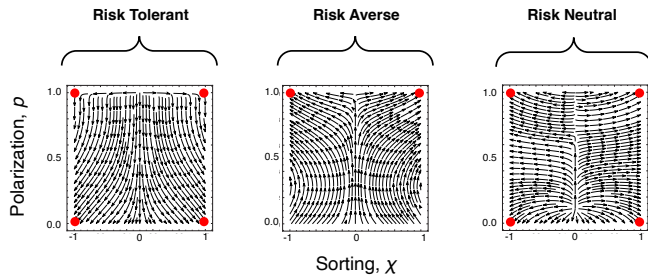


Fig. 2. Polarization and sorting. Phase portraits illustrate the dynamics of polarization p and degree of sorting χ under our model of economic interactions and party switching, with fixed identity groups. Arrows indicate the average selection gradient experienced by a local mutant against a monomorphic background (see Methods). Red dots indicate stable equilibria. (left) When the decision process for social interactions considers group or party identity, both high and low polarization states are stable, but sorting is always high $|\chi| = 1$. (center) However, when the environment is risk-averse, only high polarization and high sorting are stable. (right) And finally when the environment is risk neutral, the system returns to bistable polarization with high sorting. These plots show dynamics for $B_I = 1$, $B_O = 2$, $q_I = 1.0$, $q_O = 0.6$, $h = 10$ and $a = 0.02$. The phase portraits here show the selection gradient experienced by a monomorphic population in which parties and groups are of equal size. Dynamics under alternate decision processes (Table 1) and different parameter choices are shown in the SI.

for example higher-paying jobs or the ability to invest their gains. Alternatively, the wealthier group can be modelled as experiencing a better baseline economic environment as in (20).

Joint dynamics of sorting and polarized attitudes: We first consider the interdependence of sorting and polarization. We keep party size fixed such that a small change in x can be thought of as two members of different identity groups and parties swapping parties. We explore this interdependence for both decision processes that account for group and party and those that account for party alone.

We find that for both types of decision processes, low polarization favors high sorting – so that people change parties until parties are aligned with identity group. This is in contrast to a decision process that takes account only of group identity, where sorting has no effect on utility and therefore does not tend to evolve (see SI). Under the mixed scenario, in which interaction strategies attend to both group and party identity, high sorting evolves in all environments (Figure 2). And so the model predicts, in general, that a shift from individuals paying attention to only group identity, to individuals also paying attention to party identity, will lead to sorting.

For interaction strategies that account only for party identity, however, high polarization favors low sorting (Figure S2) when the environment does not induce risk aversion. This arises because, when players focus only on party identity and only interact with their own party, expected payoffs can be maximized by making the parties well mixed with respect to identity group, without any risk of failed interactions due to out-group members refusing to participate in an interaction (Table 1). However when group identity is visible and salient for interactions, this is not possible.

Identity group structure: We explored the coevolution of sorting and affective polarization as a function of the size and number of identity groups (see SI section 3). We find that, when there is pressure to keep parties of equal size, sorting evolves alongside polarization (Figure S9). We also explored the impact of group size and structure on the evolution of polarization and sorting (SI section 3). We find that as the number of identity groups increases, the degree of polarization and sorting declines, because interacting exclusively with an in-group becomes increasingly difficult (Figure S10). We also find that when one there are two identity group of different sizes the larger identity group becomes less sorted than the smaller group (Figure S11). Finally, we explored the effect of having one large identity group and multiple smaller identity groups. Under this scenario we find that sorting and polarization remain high, with one party becoming heterogeneous (i.e. made up of many small identity groups) and the other remaining largely homogeneous (Figure S11).

Attention to party: So far we have described the evolution of affective polarization when people pay attention to both group and party identity, and when people pay attention to group identity alone. Next we considered the effects of varying attention to party identity on the dynamics of sorting and polarization (SI section S3). We find that, whereas polarization in response to risk aversion evolves regardless of the degree of attention paid to party identity,

equally among the parties, while $\chi = 1$ corresponds to party 1 perfectly aligned with group 1 and $\chi = -1$ corresponds to party 2 perfectly aligned with group 1.

Inequality and redistribution: In our model, successful economic interactions not only benefit the pair of interacting individuals, but they also generate a contribution to a public good that benefits the population at large. To capture this public goods provision we assume that the current economic environment θ is a linear function of the benefits generated by successful interactions:

$$\theta = \alpha(1 - \gamma\alpha)r\bar{B} - \theta_0 \quad [4]$$

where α is the tax rate on wealth, γ captures the deadweight loss due to taxation, r is the benefit multiplication factor of the public good, and θ_0 is the baseline economic environment when no economic interactions occur. Here \bar{B} denotes the average benefit from economic interactions across the population. The “after tax” payoff received by an individual who generates benefit B from economic interactions is thus $(1 - \alpha)B + \theta$. This model is motivated by political economic models of linear taxation (34). The consequences of alternative forms of taxation and public goods production offer a natural direction for future work.

Redistribution of public goods is particularly important in the presence of pre-existing wealth inequality. A large body of empirical and theoretical work has demonstrated that inequality and polarization correlate and are likely causally linked (35, 36). To capture the effects of pre-existing inequality we assume that one identity group receives benefits from social interactions scaled by a factor $2\beta B$ while the other group receives $2(1 - \beta)B$. Thus when $\beta = 0.5$ there is no baseline inequality, whereas when $\beta = 0.01$ the wealthier group is around 100 times better off as a result of economic interactions than the poorer group, reflecting

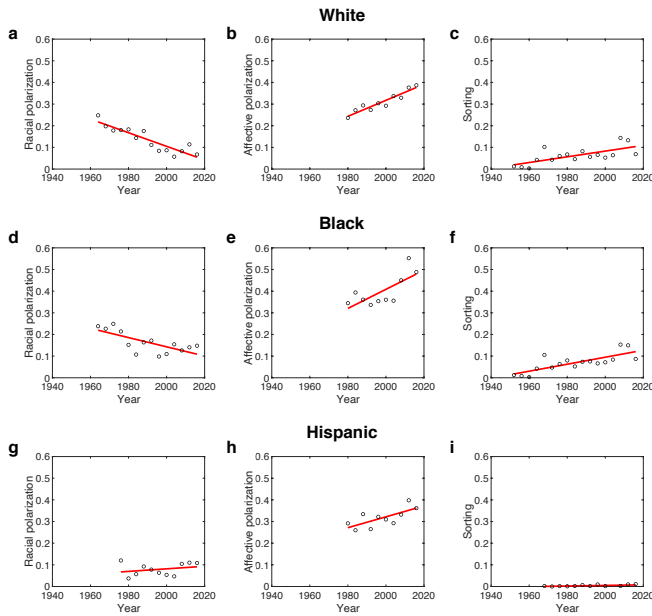


Fig. 3. Increasing salience of party identity in the United States. We looked of in-group favorability (left column) and affective polarization (center column) among white, black and Hispanic respondents in ANES data at each presidential election over past decades. While in-group favorability towards racial in-group versus racial out-group declines over time among (a) white ($p < 0.01$, $t = 7.3$) and (d) black ($p < 0.01$, $t = 3.5$) respondents, there is no significant change among Hispanic respondents (g). Affective polarization increases among white (b) ($p < 0.01$, $t = 8.2$), black (e) ($p = 0.016$, $t = 3.1$) and Hispanic (h) ($p = 0.021$, $t = 2.9$) respondents indicating a relative decline in the salience of racial identity and an increase in party identity among both the white and the black groups, and a correspondingly weaker change among the Hispanic group. Sorting of racial groups along party lines has increased among white (c) ($p < 0.01$, $t = 3.7$) and black (f) ($p < 0.01$, $t = 4.7$) respondents, but shows a much weaker change among Hispanic respondents (i) ($p = 0.047$, $t = 2.2$). Sorting is measured by the variance in party preference explained by racial identity.

the most significant effect among white and black respondents. A corresponding increase is also seen in the degree of sorting among white and black respondents, with Hispanic sorting showing only a slight increase. This suggests a shift in which individuals pay relatively greater attention to party identity over time, and also become more sorted with respect to racial identity. This pattern is consistent the predictions of our model (Figure 2) – namely, that as attention is increasingly paid to party, this will induce sorting of group identities along party lines.

Redistribution, inequality and polarization: According to our model increased polarization arises as a result of risk aversion in a poor economic environment. In general, however, different identity groups may experience different economic environments. In particular, when there is inequality such that some groups possess less wealth than others, they are more likely to be risk-averse and thus become polarized. Such inequality can lead to the evolution of polarization in the whole population (20). However, redistribution via public goods can reduce inequality, and might improve the overall economic environment.

We use Monte Carlo simulations to explore the impact of such redistribution on the dynamics of mass polarization in the presence of inequality. In particular we focus on situations in which the range of θ (Eq. 4) encompasses different economic environments, ranging from risk-neutral through risk-averse and risk-tolerant.

Figure 4 shows the effect of redistribution and inequality on the dynamics of polarization. We see that sufficient redistribution can reduce both inequality and polarization, although a high degree of redistribution is required to prevent polarization. This effect holds when public goods are purely redistributive ($r = 1$ in Eq. 4), and when public goods increase the overall wealth of the population ($r > 1$) (Figure S13). When taxation produces a deadweight loss ($\gamma > 0$, Eq. 4), it becomes harder to reduce polarization via redistribution (Figure S3).

Although sufficient redistribution can reduce inequality and polarization, it is also important to note that the effect of feedback between individual economic interactions and the overall economic environment that arises as a result of redistribution can facilitate the evolution of polarization compared to a high quality stable environment (i.e fixed θ – Figure 4c) in which polarization does not evolve. Thus introducing feedback between individual interactions and the environment through low or intermediate levels of redistribution can make things worse, by both failing to reduce inequality and facilitating the evolution of polarization (Figure 4 c-d).

We exogenously varied the amount of sorting, χ . Sorting tends to increase polarization, but it can have complex effects on levels of inequality and population average utility. This is because intermediate levels of polarization tend to result in lower levels of utility. Where reducing sorting can also reduce polarization to low levels, it has a beneficial effect in reducing inequality and increasing population average utility (see SI).

Inequality reduces average utility: We also explored the impact of inequality on population average utility. In general, underlying inequality resulting from lower income from economic interactions, tends to reduce the population average utility compared to the case where average income

sorting increases rapidly with attention to party identity (Figure S9), suggesting a feedback loop in which increasingly salient party identities lead to sorting, which further increases the salience of party identity.

Race, party and sorting: Our model predicts that individuals who take into account both group and party identity when making economic decisions will tend to evolve to a state in which group and party identity align. To examine this theoretical prediction we looked at the salience of racial identity, party identity, and sorting of people identifying as white, black or Hispanic in US presidential elections between 1964 and 2016. Using American National Election Survey (ANES) data we calculated the affective polarization and in-group favorability towards own racial group versus people from other racial groups, as well as the degree of sorting, measured by the variance in party preference explained by racial identity (Figure 3 and SI) (37).

We find that the salience of racial identity (measured by the in-group favorability among respondents) has declined over time among white and black respondents, and remained stable among Hispanic respondents. At the same time the salience of party has increased (measured by affective polarization among white respondents) among all groups, with

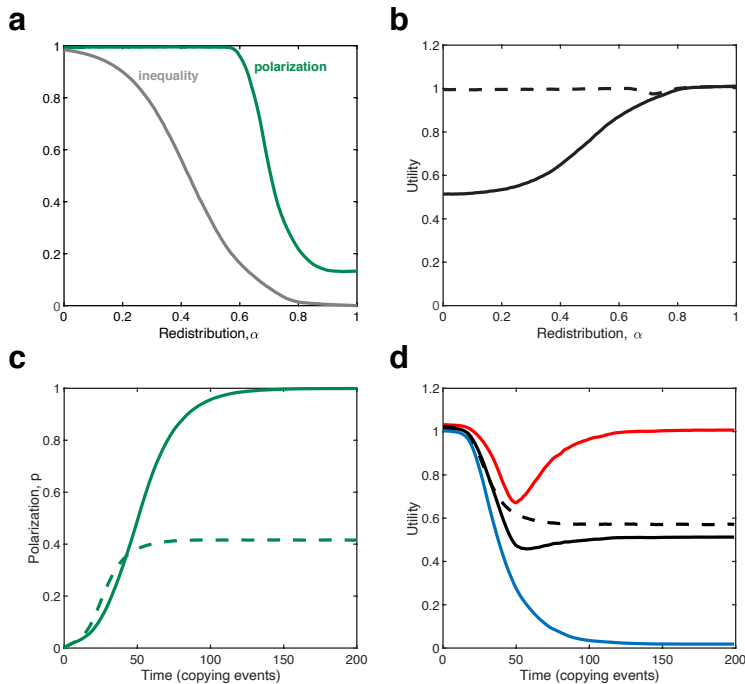


Fig. 4. Redistribution and inequality. Ensemble mean equilibria (panels a and b) and time trajectories (panels c and d) for a population initialized in a low polarization state, from individual-based simulations in the presence of wealth redistribution (Eq. 4). We show results in the case of no underlying economic inequality, $\beta = 0.5$ (dashed lines), as well the case of high underlying inequality, $\beta = 0.01$ (solid lines). Results shown here arise from a decision process that attends to group or party identity, and sorting is fixed exogenously at $\chi = 1$. (a) When public goods are not multiplicative ($r = 1$ and $\theta_0 = 0.5$), and redistribution is absent ($\alpha = 0$) overall inequality (gray line) and polarization (green line) are high. With increasing rates of redistribution, first overall inequality and then polarization decline to zero. b) Increasing redistribution increases overall utility towards the level achieved when underlying inequality is absent. c) When public goods increase overall utility ($r = 10$ and $\theta_0 = 5.4$), redistribution (here $\alpha = 0.5$) can act to magnify both polarization and inequality over time, compared to a fixed environment without feedback via redistribution. d) Overall inequality initially has a relatively small impact on population mean utility in a population initialized in a low-polarization state (black lines), but redistribution is seen to magnify the effects of underlying inequality, with the richer group (red line) experiencing a transient decline in utility as polarization evolves before returning to a value close to maximum, while the poorer group (blue line) suffer an irreversible decline towards utility close to zero. Plots show ensemble mean values across 10^4 replicate simulations, for groups of 1000 individuals each. Success probabilities and benefits are fixed at $B_I = 1$, $B_O = 2$, $q_I = 1.0$, $q_O = 0.6$ with $h = 10$ and $a = 0.02$, while $\gamma = 0$. Evolution occurs via the copying process (see methods) with selection strength $\sigma = 10$, mutation rate $\mu = 10^{-3}$ and mutation size $\Delta = 0.01$.

remains the same, but there is no inequality (Figure 4 and Figure S12-S13). This is because the poorer group tends to experience the risk-averse environment in which failed interactions produce a sharp decline in utility.

Recovery from polarization: For a wide range of parameter values the system is bistable, with both high- and low-polarization equilibria maintained unless the environment is risk-averse (see SI). Until this point we have focused on the conditions under which a population will evolve from a low- to a high-polarization state – i.e. the conditions under which the low-polarization equilibrium is lost. However, recovering low polarization once high polarization has evolved requires a switch from one equilibrium to another. In practice this may occur as the result of an environmental shock (see SI) or as a result of coordinated action in which a sufficient number of individuals simultaneously adopt a low-polarization strategy to move the population to a state that is then attracted to a low-polarization equilibrium. The threshold frequency of individuals required to achieve this transitions is determined by the size of basin of attraction for the high-polarization equilibrium.

We calculated the frequency required for escape – the proportion of the population that must simultaneously adopt a low-polarization behavior to escape the high-polarization equilibrium (Figure 5) as a function of the baseline economic environment (θ_0) and the degree of sorting (χ). We find that, for many environments, escape is easiest following an economic shock (i.e. a poor environment) in which the population becomes risk-tolerant. However, we also find that, once the economic environment is sufficiently advantageous, such that risk aversion is lessened, escape from polarization is feasible,

especially if the degree of sorting is also low. Thus, addressing polarization, absent an economic catastrophe, requires first improving the economic environment, minimizing sorting, and then engaging in sufficient coordinated action for a subset of the population to adopt low-polarization behavior, followed by the spread of low-polarization behavior by social contagion.

Discussion

Our model provides a framework for connecting the effects of inter-group animus, economic adversity and mass political polarization through the lens of cultural evolution (38, 39). We focus on polarization expressed through loss of positive social interactions with members of an out-party (10, 11), and sorting of identity groups along party lines (7–9). We show that attending to party identity when deciding who to interact with is sufficient to translate inter-group polarization stimulated by adverse economic conditions into political polarization between members of opposing parties (Figure 2). We further show that if party identity is able to evolve alongside behavioral strategies, this can also lead to sorting of identity groups along party lines (Figure 3). We then show that feedback between individuals’ economic interactions and the overall economic environment can lead to increased polarization and amplify the effects of underlying inequality between groups (Figure 4). These effects are magnified when identity groups are sorted along party lines, but can be mitigated or prevented entirely if redistribution of wealth via public goods is put in place to combat inequality (Figure 4-5).

In performing simulations on redistribution (Figure 4) we

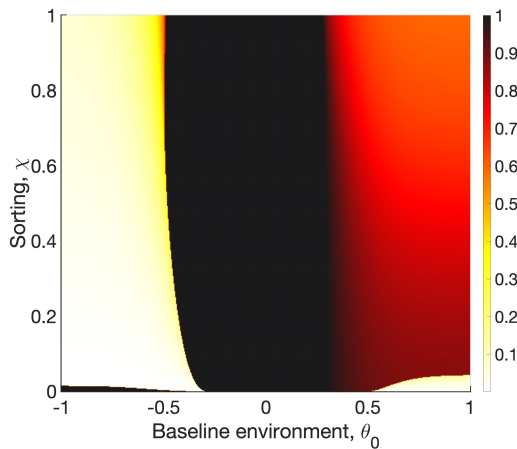


Fig. 5. Recovering from polarization. We numerically calculated the size of the basin of attraction for the high-polarization state, which determines the escape frequency – that is, the proportion of the population that must simultaneously adopt a low-polarization behavior to escape the high-polarization equilibrium. We show the escape frequency as a function of the baseline environment, θ_0 and the (fixed) degree of sorting, χ . For intermediate values of theta, risk aversion means that it becomes increasingly difficult to reverse polarization (higher escape frequency, darker colors). However when the environment is very bad, and risk tolerance dominates, or if the environment is very good, it becomes possible to reverse polarization through the coordinated behavior of small frequency of low-polarization individuals. When the environment is good and sorting is low, polarization is easiest to reverse without entering a highly deleterious environment. Other parameters are set to $B_I = 1$, $B_O = 2$, $q_I = 1.0$, $q_O = 0.6$ with $h = 10$ and $a = 0.02$, while $\gamma = 0$, $\beta = 0.5$, $\theta_0 = 0.5$ and $r = 1$ unless otherwise stated.

S5) that an increase in the salience of party identity can be beneficial under some circumstances, however it is important to note that such changes are likely driven by other factors as well. For example social desirability bias can lead to reduced racial polarization (40) while allowing inter-party animus to remain high. In addition, there is evidence (41) that increased affective polarization is partly driven by exogenous factors such as media environment and lack of contact opportunity with out-group members resulting in induced homophily (42). The interaction between such structural forces and the dynamics of cultural evolution described is an important topic for future research.

Our work focuses on affective polarization and sorting with respect to identity groups among the electorate. However attempts to prevent and reverse polarization must also take account of the mechanisms that enable elite (43, 44) and ideological polarization (44, 45), and must account for the role of factors such as geography (46) and population and social network structure (47–49) in producing mass polarization, in addition to the inter-group and economic factors studied here. We must also remain alert to the circumstances under which polarization can provide benefits (50, 51) (e.g. Figure S12 in which increased sorting can increase polarization but reduce inequality).

The impact of underlying inequality on the evolution of polarization, and the amplification of the effects of inequality via economic feedback, illustrate the need to think carefully about mass political polarization in the context of inter-group conflict and the economic environment (2, 35). This is particularly true when assessing ways to prevent or reverse mass polarization. The success of redistribution at stemming the tide of polarization in our model is striking, and it suggests a possible path for preventing such attitudes from taking hold in future. We emphasize though that this strategy is only possible if implemented in a population that is not already polarized, in an environment that supports low polarization. Once polarization sets in, it typically remains stable under individual-level evolutionary dynamics, even when the economic environment improves or inequality is reversed. The only remedy for reversing a polarized state, under our analysis, requires either a shock (Figure S14) or a sufficiently good economic environment coupled with collective action by a portion of the population who change strategies simultaneously.

Materials and Methods

In this section we describe the decision process, the calculation of utility and selection gradient, and the copying process used in simulations. Further analysis of the model can be found in the SI.

Measure of inequality: Throughout we adopt a simple measure of inequality: the difference in relative utility between the two groups i.e. $\frac{w_{HIGH} - w_{LOW}}{w_{HIGH} + w_{LOW}}$ where w_{HIGH} is the average utility of the richer group and w_{LOW} the utility of the poorer group

Decision process: Table 1 gives the probability for a focal player i choosing to interact with a given player j based on the identity of j and the decision process adopted by i . In order to calculate the utility of i given a decision process, we must calculate the probability distribution for the next interaction i participates in, conditional on an interaction occurring. That is, we must weight the probability of interactions given in Table 1 by the number of individuals in each group, and normalize the distribution. This corresponds to a process in which the focal player randomly draws an individual from the population and then decides to pursue an

used fixed levels of sorting, on the basis that changes to party identity can be assumed slow compared to changes in interaction strategy. However if the evolution of sorting occurs rapidly, situations may arise in which members of different parties experience divergent selection pressures, raising the possibility that different levels of polarization may arise in some subsets of the population and not others. We find (Figure S11) that when a population is made up of one large and many small identity groups, the dynamics of sorting tend to make the large group form one homogeneous party, and the smaller groups bands together to form a more diverse party, with both parties exhibiting high levels polarization – a situation reminiscent of the major US political parties today. Exploring how effects such as these interact with inequality is a natural direction for future work.

In our model we specify a mechanism, based on pairwise economic interactions, such as trade, employment decisions, and social cooperation, that leads to the risk aversion that drives polarization. The risk aversion we describe depends on the rapid decline in utility (Eq. 2), which may reflect either perceived or actual loss of income. Although one might expect, on the basis of our model, that richer groups should be less polarized and more heterogeneous in their interactions, we find here (Figure 4) and in previous work (20) that polarization rapidly spreads to the whole population, including to risk-tolerant individuals, once it takes hold among a risk-averse subset.

Empirical trends of racial and affective polarization and sorting are consistent with our model (Figure 3), but the drivers of increased party salience are complex. We find (Figure

563 interaction with that individual based on the probabilities given in
 564 Table 1. These normalized distributions are given below for the
 565 decision process that takes account of only party identity, and for
 566 the decision process that takes account group or party identity.
 567 Note that if the decision process takes account of group identity
 568 only, no normalization is required, since the degree of sorting does
 569 not impact the probability of interaction.

570
 571 **Only party identity:** Under this decision process the
 572 probability of an individual i belonging to group 1 and party 1
 573 choosing to interact with an individual with identity kl is $\pi_{kl}(x)$
 574 where k indexes the group identity I or O and l indexes the party
 575 identity and x is the frequency of individuals from group 1 in party
 576 1 (and, by symmetry, the number of individuals from group 2 in
 577 party 2). We then have

$$\begin{aligned}
 578 \quad \pi_{II}(x) &= \frac{p_i x}{p_i x + (1 - p_i)(1 - x) + p_i(1 - x) + (1 - p_i)x} \\
 579 \quad \pi_{IO}(x) &= \frac{(1 - p_i)(1 - x)}{p_i x + (1 - p_i)(1 - x) + p_i(1 - x) + (1 - p_i)x} \\
 580 \quad \pi_{OI}(x) &= \frac{p_i(1 - x)}{p_i x + (1 - p_i)(1 - x) + p_i(1 - x) + (1 - p_i)x} \\
 581 \quad \pi_{OO}(x) &= \frac{(1 - p_i)x}{p_i x + (1 - p_i)(1 - x) + p_i(1 - x) + (1 - p_i)x}
 \end{aligned} \tag{5}$$

582 where x is the proportion of identity group k that also belong to
 583 party k .

584
 585 **Party OR Group Identity:** Under this decision process
 586 the probability of an individual i who belongs to group 1 and party
 587 1, choosing to interact with an individual with identity kl is $\phi_{kl}(x)$.
 588 We then have

$$\begin{aligned}
 591 \quad \phi_{II}(x) &= \frac{(2p_i - p_i^2)x}{(2p_i - p_i^2)x + 2(p_i + (1 - p_i)^2)(1 - x) + (1 - p_i^2)x} \\
 592 \quad \phi_{IO}(x) &= \frac{(p_i + (1 - p_i)^2)(1 - x)}{(2p_i - p_i^2)x + 2(p_i + (1 - p_i)^2)(1 - x) + (1 - p_i^2)x} \\
 593 \quad \phi_{OI}(x) &= \frac{(p_i + (1 - p_i)^2)(1 - x)}{(2p_i - p_i^2)x + 2(p_i + (1 - p_i)^2)(1 - x) + (1 - p_i^2)x} \\
 594 \quad \phi_{OO}(x) &= \frac{(1 - p_i^2)x}{(2p_i - p_i^2)x + 2(p_i + (1 - p_i)^2)(1 - x) + (1 - p_i^2)x}
 \end{aligned} \tag{6}$$

595
 596 This decision strategy reflects a situation in which an individual
 597 sees someone as a member of their in-group if they share either the
 598 same group or the same party identity, and weights both of those
 599 dimensions of identity equally. We explore an AND type decision
 600 process in the SI.

601
 602 **Expected utility:** In order to explore the evolutionary dy-
 603 namics of polarization, we calculate the expected utility of a mutant
 604 strategy p_i , which deviates by a small amount from the resident
 605 strategy p employed by the rest of the population. Using Eq. 5 and
 606 Eq. 6 above we can now write down the expected fitness for such a
 607 mutant under a given decision process. When players only attend
 608 to party identity the utility of such a mutant is

$$\begin{aligned}
 610 \quad w_i(x) &= [\pi_{II}(x) + \pi_{IO}(x)(1 - p)]q_I F(B_I + \theta) + \\
 611 \quad & [(\pi_{II}(x) + \pi_{IO}(x))(1 - q_I) + \pi_{IO}(x)pq_I]F(\theta) + \\
 612 \quad & [\pi_{OI}(x) + \pi_{OO}(x)(1 - p)]q_O F(B_O + \theta) + \\
 613 \quad & [(\pi_{OI}(x) + \pi_{OO}(x))(1 - q_O) + \pi_{OO}(x)pq_O]F(\theta)
 \end{aligned} \tag{7}$$

614
 615 whereas the utility of a mutant when players attend to party or
 616 group is

$$\begin{aligned}
 w_i(x) &= [\phi_{II}(x) + \phi_{IO}(x)(1 - p)]q_I F(B_I + \theta) + \\
 & [(\phi_{II}(x) + \phi_{IO}(x))(1 - q_I) + \phi_{IO}(x)pq_I]F(\theta) = \\
 & [\phi_{OI}(x)(1 - p) + \phi_{OO}(x)(1 - p^2)]q_O F(B_O + \theta) = \\
 & [(\phi_{OI}(x) + \phi_{OO}(x))(1 - q_O) + (\phi_{OI}(x)p + \phi_{OO}(x)p^2)q_O]F(\theta)
 \end{aligned} \tag{8}$$

620
 621 **Selection gradient:** We can now calculate the average selection
 622 gradient (52, 53) experienced by the mutant p_i , which is given by

$$s_p = \left. \frac{\partial[xw_i(x)]}{\partial p_i} \right|_{p_i=p} + \left. \frac{\partial[(1 - x)w_i(1 - x)]}{\partial p_i} \right|_{p_i=p} \tag{9}$$

623
 624 When the selection gradient is positive the mutant has an advantage
 625 over the resident strategy on average. Note however that when
 626 $0 < |\chi| < 1$ different individuals experience different effects from
 627 the same mutation. This issue is discussed in more detail in the SI.

628 We can also calculate the effect of a small change to the degree
 629 of sorting in the population by calculating the gradient

$$s_x = \frac{\partial[xw_i(x)]}{\partial x} + \frac{\partial[(1 - x)w_i(1 - x)]}{\partial x} \tag{10}$$

630
 631 When s_x is positive, the average effect of an increase in sorting is
 632 to increase the average utility of the population. It is Eqs. 5-10
 633 that are used to produce Figures 2-3 (see also SI).

634
 635 **Evolutionary simulations:** In order to simulate the evo-
 636 lutionary dynamics of this system we consider a population evolving
 637 under a “copying process” (54) in which individuals are able to
 638 observe the utility of other individuals and compare it to their own.
 639 The dynamics of the model are as follows: An individual i is chosen
 640 at random from a population of fixed size N . A second individual j
 641 is then chosen at random for her to “observe”. If i has utility w_i
 642 and j has utility w_j then i chooses to copy the strategy of j with
 643 probability $1/(1 + \exp[\sigma(w_j - w_i)])$, where σ scales the “strength
 644 of selection” of the evolutionary process. Note that if $w_j \gg w_i$ the
 645 probability of i copying the behavior of j is close to 1, whereas if
 646 $w_j \ll w_i$ the probability is close to 0.

647 Individual based simulations used to produce Figures 4-5 were
 648 performed under the copying process using populations composed
 649 of two identity groups of size $N = 1000$ individuals, with sorting
 650 of groups among parties fixed. Mean trajectories were determined
 651 from an ensemble of 10^4 sample paths. Simulations were run for
 652 $100N - 200N$ copying events to find equilibria. Mutations were
 653 assumed to occur at a rate $1/N$ per copying event, with the target of
 654 the mutation chosen randomly from the population. Mutations were
 655 assumed to be local such that the target of the mutation had their
 656 strategy perturbed by $\Delta = \pm 0.01$ with mutations that increase and
 657 decrease p equally likely, and we impose the appropriate boundary
 658 conditions to ensure strategies were physical.

659
 660 **ACKNOWLEDGMENTS.** We thank participants of the PNAS
 661 Conference on the Dynamics of Political Polarization, held (virtually)
 662 in January 2021, and especially Stuart Feldman for feedback
 663 and suggestions for research.

- 664 1. C Muddy, CR Kaltwasser, *Populism: A very short introduction*. (Oxford University Press), (2017).
- 665 2. DJ Ahler, The group theory of parties: Identity politics, party stereotypes, and polarization in the 21st century. *The Forum* **16**, 3–22 (2018).
- 666 3. L Mason, NP Kalmoe, What you need to know about how many americans condone political violence — and why. *The Wash. Post* (2021).
- 667 4. MPWSJAJCP Fiorina, *Culture War? The Myth of a Polarized America*. (Pearson Longman New York), (2005).
- 668 5. A Abramowitz, *The Disappearing Center: Engaged Citizens, Polarization, and American Democracy*. (Yale University Press), (2010).
- 669 6. N McCarty, *Polarization: What Everyone Needs to Know*. (Oxford University Press, New York), (2019).
- 670 7. MS Levendusky, *The Partisan Sort: How Liberals Became Democrats and Conservatives Became Republican*. (University of Chicago Press), (2009).

- 682 8. L Mason, 'i disrespectfully agree': The differential effects of partisan sorting on social and
683 issue polarization. *Am. J. Polit. Sci.* **59**, 128–145 (2015).
- 684 9. L Mason, J Wronski, One tribe to bind them all: How our social group attachments strengthen
685 partisanship. *Polit. Psychol.* **39**, 257–277 (2018).
- 686 10. S Iyengar, G Sood, Y Lelkes, Affect, Not Ideology: A Social Identity Perspective on Polariza-
687 tion. *Public Opin. Q.* **76**, 405–431 (2012).
- 688 11. L Mason, A cross-cutting calm: How social sorting drives affective polarization. *Public Opin.*
689 *Q.* **80**, 351–377 (2016).
- 690 12. BF Schaffner, M MacWilliams, T Nteta, Explaining white polarization in the 2016 vote for
691 president: The sobering role of racism and sexism in *Conference on the US Elections of*
692 *2016*. (2016).
- 693 13. MD Luttwig, CM Federico, H Lavine, Supporters and opponents of donald trump respond dif-
694 ferently to racial cues: An experimental analysis. *Res. & Polit.* **4**, 2053168017737411 (2017).
- 695 14. J Sides, M Tesler, L Vavreck, How trump lost and won. *J. Democr.* **28**, 34–44 (2017).
- 696 15. R Inglehart, P Norris, Trump, brexit, and the rise of populism: Economic have-nots and cul-
697 tural backlash. *HKS Fac. Res. Work. Pap. Ser. RWP16-026* (2016).
- 698 16. M Tesler, In a trump-clinton match-up, racial prejudice makes a striking difference. *The Wash.*
699 *Post* (2016).
- 700 17. A Arnorsson, G Zoega, On the causes of brexit. *CESifo Work. Pap. Ser. No. 6056* (2016).
- 701 18. J Kolko, Trump was stronger where the economy is weaker. *FiveThirtyEight* (2016).
- 702 19. EC Mitrea, M Mühlböck, J Warmuth, Extreme pessimists? expected socioeconomic down-
703 ward mobility and the political attitudes of young adults. *Polit. Behav.* (2020).
- 704 20. AJ Stewart, N McCarty, JJ Bryson, Polarization under rising inequality and economic decline.
705 *Sci. Adv.* **6** (2020).
- 706 21. BK Szymanska, M Maa, DR Tabina, J Gaoa, MW Macy, Polarization and tipping points.
707 *PNAS Conf. on Dyn. Polit. Polariz.* (2021).
- 708 22. M Ruef, Strong ties, weak ties and islands: structural and cultural predictors of organizational
709 innovation. *Ind. Corp. Chang.* **11**, 427–449 (2002).
- 710 23. AW Woolley, CF Chabris, A Pentland, N Hashmi, TW Malone, Evidence for a collective intel-
711 ligence factor in the performance of human groups. *Science* **330**, 686–688 (2010).
- 712 24. MA Martinez, HE Aldrich, Networking strategies for entrepreneurs: balancing cohesion and
713 diversity. *Int. J. Entrepreneurial Behav. & Res.* **17**, 7–38 (2011).
- 714 25. P Carruthers, PK Smith, eds., *Theories of theories of mind*. (Cambridge University Press),
715 (1996).
- 716 26. F Shi, M Teplitskiy, E Duede, JA Evans, The wisdom of polarized crowds. *Nat. Hum. Behav.*
717 **3**, 329–336 (2019).
- 718 27. KK Charles, J Guryan, Studying discrimination: Fundamental challenges and recent progress.
719 *Annu. Rev. Econ.* **3**, 479–511 (2011).
- 720 28. C McConnell, Y Margalit, N Malhotra, M Levendusky, The economic consequences of parti-
721 sanship in a polarized era. *Am. J. Polit. Sci.* **62**, 5–18 (2018).
- 722 29. S Iyengar, SJ Westwood, Fear and loathing across party lines: New evidence on group
723 polarization. *Am. J. Polit. Sci.* **59**, 690–707 (2015).
- 724 30. GS Becker, *The economics of discrimination*. (University of Chicago press), (2010).
- 725 31. PJ Egan, Identity as dependent variable: How americans shift their identities to align with
726 their politics. *Am. J. Polit. Sci.* **64**, 699–716 (2020).
- 727 32. MF Margolis, How politics affects religion: Partisanship, socialization, and religiosity in amer-
728 ica. *The J. Polit.* **80**, 30–43 (2018).
- 729 33. D Green, B Palmquist, E Schickler, *Partisan hearts and minds*. (Yale University Press),
730 (2008).
- 731 34. AH Meltzer, SF Richard, A rational theory of the size of government. *J. political Econ.* **89**,
732 914–927 (1981).
- 733 35. NM McCarty, KT Poole, H Rosenthal, *Polarized America: The dance of ideology and unequal*
734 *riches*. (MIT Press, Cambridge, MA), second edition, (2016).
- 735 36. J Voorheis, N McCarty, B Shor, Unequal incomes, ideology and gridlock: How rising inequal-
736 ity increases political polarization. Available at SSRN: <https://ssrn.com/abstract=2649215>
737 (2015).
- 738 37. The American National Election Studies (www.electionstudies.org). These materials are
739 based on work supported by the National Science Foundation under grant numbers SES
740 1444721, 2014-2017, the University of Michigan, and Stanford University. (year?).
- 741 38. R Boyd, PJ Richerson, *Culture and the Evolutionary Process*. (University of Chicago Press),
742 (1985).
- 743 39. LL Cavalli-Sforza, MW Feldman, *Cultural transmission and evolution: a quantitative approach*.
744 (Princeton University Press) No. 16, (1981).
- 745 40. A Khalid, How white liberals became woke, radically changing their outlook on race. *NPR*
746 (2019).
- 747 41. S Iyengar, Y Lelkes, M Levendusky, N Malhotra, SJ Westwood, The origins and conse-
748 quences of affective polarization in the united states. *Annu. Rev. Polit. Sci.* **22**, 129–146
749 (2019).
- 750 42. G Kossinets, D Watts, Origins of homophily in an evolving social network. *Am. J. Sociol.* **115**,
751 405–450 (2009).
- 752 43. S Feldman, et al., Emergence of new influencer elites challenges traditional polarization anal-
753 ysis. *PNAS Conf. on Dyn. Polit. Polariz.* (2021).
- 754 44. NE Leonard, K Lipsitz, A Bizyaeva, A Franci, Y Lelkes, The nonlinear feedback dynamics of
755 asymmetric political polarization. *PNAS Conf. on Dyn. Polit. Polariz.* (2021).
- 756 45. R Axelrod, JJ Daymude, S Forrest, Preventing extreme polarization of political attitudes.
757 *PNAS Conf. on Dyn. Polit. Polariz.* (2021).
- 758 46. OJ Chu, M Wiedermann, JF Donges, G Pop-Eleches, GB Robertson, The micro-dynamics
759 of geographic polarization: a model and an application to survey data from ukraine. *PNAS*
760 *Conf. on Dyn. Polit. Polariz.* (2021).
- 761 47. AJ Stewart, et al., Information gerrymandering and undemocratic decisions. *Nature* **573**,
762 117–121 (2019).
- 763 48. FP Santos, Y Lelkes, SA Levin, Social matching algorithms and dynamics of polarization in
764 online social networks. *PNAS Conf. on Dyn. Polit. Polariz.* (2021).
- 765 49. CK Tokita, A M.Guess, CE Tarnita, How polarized social networks emerge from organic infor-
766 mation cascades. *PNAS Conf. on Dyn. Polit. Polariz.* (2021).
- 767 50. M Kawakatsu, Y Lelkes, SA Levin, CE Tarnita, Revisiting madison's cure for the "mischiefs
768 of faction": Can cooperation and polarization evolve together? *PNAS Conf. on Dyn. Polit.*
769 *Polariz.* (2021).
- 770 51. VV Vasconcelos, et al., Successful coordination in polarized societies. *PNAS Conf. on Dyn.*
771 *Polit. Polariz.* (2021).
- 772 52. C Mullon, L Keller, L Lehmann, Evolutionary stability of jointly evolving traits in subdivided
773 populations. *Am Nat* **188**, 175–95 (2016).
- 774 53. O Leimar, Multidimensional convergence stability. *Evol. Ecol. Res.* **11**, 191–208 (2009).
- 775 54. A Traulsen, MA Nowak, JM Pacheco, Stochastic dynamics of invasion and fixation. *Phys Rev*
776 *E Stat Nonlin Soft Matter Phys* **74**, 011909 (2006).