Supplementary Materials for

**Polarization under rising inequality and economic decline**

Alexander J. Stewart*, Nolan McCarty*, Joanna J. Bryson*

*Corresponding author. Email: alstew@sas.upenn.edu (A.J.S.); nmccarty@princeton.edu (N.M.); jjb@alum.mit.edu (J.J.B.)

Published 11 December 2020, *Sci. Adv.* 6, eabd4201 (2020)
DOI: 10.1126/sciadv.abd4201

This PDF file includes:

Sections S1 to S5
Tables S1 to S3
Figs. S1 to S11
In this supplement, we relax several of the modeling assumptions presented in the main text to demonstrate the robustness of our conclusions. In particular we vary the “functional response curves” that relate the outcome of a given social interaction to the accumulated utility of many interactions and the number of interactions among individuals. We also provide further details of the individual-based simulations presented in the main text. Note that equation numbers continue from the main text.

1 Model Parameters

The parameters associated with the model presented in Figure 1 and the main text are summarized in Table 1 below. We now discuss how variation in the parameters of the model impacts the associated evolutionary dynamics and the degree of polarization that arises across environments. We first discuss the case where the probability of success of an out-group interaction is constant and independent of the strategies adopted by other members of the population. We then discuss the case where the risk of out-group interactions depends on the strategy of the target for the interaction. We analyze both models by looking at the equilibria under both local and non-local mutations (i.e. under scenarios where players adjust their behavior either gradually or in sudden-large shifts such as may occur in response to structural change).

| Parameter | Default Simulation Value | Meaning |
|-----------|-------------------------|---------|
| $B_i$     | 0.5                     | Benefit received due to a successful in-group interaction. |
| $B_o$     | 1.0                     | Benefit received due to a successful out-group interaction. |
| $q_i$     | 1.0                     | Probability of a successful in-group interaction |
| $q_o$     | 0.6                     | Probability of a successful out-group interaction |
| $n$       | 5                       | Number of attempted interactions before strategy update |
| $N$       | 1000                    | Population size |
| $\theta$ | $[-1,1]$                | Quality of the environment (no inequality) |
| $\theta_0$ | $[0,1]$               | Average quality of the environment (with inequality) |
| $\theta_g$ | $[0,1.1]$           | Strength of inequality |
| $\pi$    | 0.1                     | Proportion of the population that is very well off |
| $h$      | 2                       | Steepness of the functional response curve |
| $\alpha$ | 0.01                    | Slope of the functional response curve |
| $\sigma$ | 10                      | Selection strength |
| $\mu$    | 0.0001                  | Mutation rate |

Table S1 – Model parameters and the default values chosen for main text individual-based simulations.
2 Case 1: Non-social payoffs

2.1 Stability and Invasibility

We first consider the case in which the probability of success of an out-group interaction is simply $q_o$, which does not depend on the strategy of the target of the interaction. Thus the payoff $w_f$ of a mutant $f$ does not depend on the background $g$ into which it is introduced. This means that a strategy that maximizes $w$ can always invade and can never be invaded, so that under an evolutionary process with non-local mutations as shown in Figure 2 (main text) the population will always arrive at the global maximum.

However, we are also interested in the behavior of the system under local mutations (or “gradual methods” as they are called in the main text). To this end we look at the selection gradient of main text Eq. 4 which gives

$$
\frac{\partial s(f,g)}{\partial f} = \sum_{k=0}^{n} \sum_{l_i=0}^{k} \sum_{l_o=0}^{n-k} \binom{n}{k} p_f^{k-1} (1 - p_f)^{n-k-1} (k - np_f) \times \\
\binom{k}{l_i} q_i^{l_i} (1 - q_i)^{k-l_i} \binom{n-k}{l_o} (q_o)^{l_o} (1 - q_o)^{n-k-l_o} \times \\
\exp[h(l_i B_i + l_o B_o + n\theta)] (1 + \exp[h(l_i B_i + l_o B_o + n\theta)])
$$

which we can evaluate numerically to calculate the points of zero selection gradient as shown in Figure S1 below. We can also use Eq. 12 to gain insight into the dynamics of polarization on display in Figure 2 of the main text.

In particular, if the environment is sufficiently bad that $l_i B_i + l_o B_o + n\theta < 0 \ \forall \ l_i, l_o$ or if it sufficiently good such that $l_i B_i + l_o B_o + n\theta > 0, \ \forall \ l_i, l_o$ then we can approximate the sigmoidal term in Eq. 12 as a constant and recover selection gradient

$$
\frac{\partial s(f,g)}{\partial p_f} = \sum_{k=0}^{n} \binom{n}{k} p_f^{k-1} (1 - p_f)^{n-k-1} (k - np_f) (1 + nr q_o B_o + kr (q_i B_i - q_o B_o)) \\
= r (q_i B_i - q_o B_o)
$$

i.e. evolution proceeds in the direction of the the strategy that increases expected fitness. However, for intermediate values of $\theta$ we can approximate the sigmoidal term as 0 for $l_i B_i + l_o B_o < -n\theta$ and as 1
otherwise. Thus Eq. 12 becomes the sum over the probability distribution conditional on the fact that
the payoff received is greater than \(-n\theta\). This cannot be calculated explicitly in most cases but note that
if \(nq_iB_i + nq_oB_o + n\theta > 0\) then terms with low values of \(l_o\) or \(l_i\) will be eliminated from the summation.
Since \(q_o < q_i\), this means that players who tend to use out-group interactions will tend to suffer more
in this regime, and thus the population becomes risk averse. In contrast, when \(nq_iB_i + nq_oB_o + n\theta < 0\)
only terms with high values of \(l_o\) or \(l_i\) will be included in the summation, which tends to favor out-group
interactions. This qualitatively captures the results shown in Figure 2.

In the following sections we systematically vary the parameters of Table 2 in order to assess the
robustness of the results presented in the main text.

Figure S1 – Pairwise invasability plot in different environments using the default parameters as given in Table
1. We see, just as in Figure 2 of the main text, that low polarization (\(p^* = 0\)) is stable in a very good or a very
bad environment (left and right plots) but that this situation is reversed in an intermediate environment (center
plot). Under both local and non-local mutations this effect is evident.
2.2 Impact of Risks and Benefits of Interactions

Figure S2 shows the strategies that maximize fitness across environments as we vary the risk parameters $q_i$ and $q_o$ and the benefit parameters $B_i$ and $B_o$. In all cases we see qualitatively similar results to those shown in Figure 2 of the main text – for intermediate environments risk aversion can lead to an increase in polarization even when the expected benefit of out-group interactions exceed those of in-group interactions.

Figure S2 – Stable equilibria as a function of environment ($\theta$, x-axis) and payoff parameters. On the y-axis we vary a) The probability of success of out-group interactions, b) the probability of failure of in-group interactions, c) the benefit of successful in-group interactions and d) the opportunity cost of a successful in-group interaction compared to a successful out-group interaction. All other parameters are set to the default values in Table 1.
We also examined the stable strategies of the model fixing the expected benefits of in- and out-group interactions $B_i q_i$ and $B_o q_o$ and varying the risk associated with out-group interactions. Once again we see a shift from stable low-polarization strategies at intermediate environments, unless the risk of out-group interactions becomes low (i.e. $q_o$ becomes sufficiently large) in which case low polarization strategies are always stable.

Figure S3 – Stable equilibria assuming fixed expected benefits from in- and out-group interactions, $q_i B_i$ and $q_o B_o$ under varying risk of out-group interactions $q_o$ (y-axis) and across environments (x-axis). All other parameters are set to the default values in Table 1.
2.3 Impact of the Number of Interactions

We examined the effect of interaction number $n$ on our results. From this we draw four qualitative conclusions as follows:

- When the increase in expected benefits per out-group interactions is high (20%) high-polarization can only take hold when the number of interactions is small ($n \in [10, 50]$ interactions, Figure S4a, S4c and S4d))

- When the increase in expected benefits per out-group interactions is low (2%) high-polarization can take hold even when each individual participates in many hundreds of interactions (Figure S4b)

- Increasing the steepness of the sigmoid function (i.e the rate of loss of fitness in a declining environment) makes high polarization more likely to take hold even when individuals participate in many interactions ($n < 50$, Figure S4c)

- Decreasing the steepness of the linear function has qualitatively similar effect (Figure S4d)
Figure S4 – Stable equilibria with varying numbers of interactions $n$ (y-axis) and across environments (x-axis). (top left) With a 20% increase in expected benefit from out-group interactions compared to in-group interactions high polarization only occurs for $n < 10$ (top right) with a 2% increase however high polarization can take hold even when $n = 100$ (bottom left) With a 20% increase and a steep sigmoidal function ($h = 100$) high polarization can take hold with a greater number of interactions ($n < 50$) and (bottom right) similarly for a shallower linear function $\alpha = 0.002$. All other parameters are as shown in Table 1.
We also consider the case when the number of interactions \( n \) is large such that \( n \sim N \). In this case we assume that, when using a low-polarization strategy \( p = 1 \), the number of successful interactions from \( n \gg 1 \) attempts with the out-group is normally distributed with mean \( \mu \) and variance \( \sigma \). We assume that under a low polarization strategy, \( p = 0 \), all interactions are successful. Setting \( \mu = 0.6 \) we explore the stability of in- and out-group strategies as a function of \( \sigma \) (Figure S5).

![Figure S5 – Stable equilibria with changing variance for a fixed number of interactions \( n = 100 \) (y-axis) and across environments (x-axis). Once \( \sigma > 0.15n \) high polarization can take hold, with the range of environments that favor polarization quickly becoming large once \( \sigma > 0.2n \) (blue region). All other parameters are as shown in Table 1.](image-url)
2.4 Impact of the Rate of Benefit Accumulation

Finally we varied the curvature of the benefit accumulation function. In the main text we assume a function of the form

\[
f(l_i, l_o, \theta) = \frac{\exp[h(l_i B_i + l_o B_o + n\theta)/n]}{1 + \exp[h(l_i B_i + l_o B_o + n\theta)/n]}(1 + \alpha(l_i B_i + l_o B_o))
\] (14)

where the first (sigmoidal) term captures the idea that, below a certain threshold fitness rapidly declines either, in a biological context, due to starvation or in an economic context due to in ability to meet basic financial obligations etc. The second (linear) term reflects the fact that, once above the threshold, there is still an advantage to having higher payoff, where \( h \) determines the steepness of the threshold function and \( \alpha \) the steepness of the linear function. Note that by varying \( h \) and \( \alpha \) we can produce a whole family of qualitatively different benefit functions from a purely linear function to a Heaviside step function. Finally, note that the position of the threshold above which sufficient benefit from interactions is accumulated depends on the environment \( \theta \) which describes the harshness of the environment, the cost or availability of resources depending on whether we are thinking about a biological or an human economy.

We see that increasing the steepness of the sigmoid function (Figure S6 - top row) has little effect above \( h \sim 10 \). However, below this we see an increase in high polarization strategies in good environments. In contrast, increasing the steepness of the linear function \( \alpha \) tends to reduce the range of environments in which high polarization strategies can take hold if the expected benefits of out-group interactions exceed those of in-group interactions (Figure S6, middle row).

We also explored the behavior of the model under a benefit accumulation function with constant positive or negative curvature

\[
f(l_i, l_o, \theta) = [(l_i B_i + l_o B_o + n\theta)/n]^{10^\beta}
\] (15)

where we choose the form of the exponent so that \( \beta = 0 \) corresponds to zero curvature, with negative values corresponding to a concave accumulation function. Here we see as expected that risk-averse, high-polarization strategies only arise when the accumulation function is concave. We also observe the same transition from low- to high- polarization strategies in a declining environment, but without the corresponding reverse transition as the environment continues to decline (since unlike the accumulation function of Eq. 14, Eq. 15 has a fixed direction of curvature – Figure S6, bottom row).
Figure S6 – Stable equilibria for the model with varying benefit accumulation functions. The left hand column shows how the varied parameter changes the shape of the accumulation function, while the right hand column shows the equilibria for the model as the parameter varies (y-axis) across different environments (x-axis). Top row – Increasing the steepness of the sigmoidal function has little impact above $h = 10$ however for smaller values we see an increase in polarization in good environments. Middle row – Steeper linear components to the accumulation function Eq. 14 tend to decrease the range of environments when polarization can take hold. Bottom row – Varying the curvature of the accumulation function Eq. 15 demonstrates the known result that risk aversion requires a concave utility function, where we see the same transition from low to high polarization as environments decline as described in the main text. Parameter values are as shown in Table 1, with the exception that we have set $q_o = 0.51$ to make the impact of varying these parameters more clearly visible.
3 Case 2: Social payoffs

3.1 Stability and Invasibility

Under this model the probability of success of an out-group interaction for a player $h$ interacting with another player $g$ is $q_o(1 - p_g)$ where the term $(1 - p_g)$ accounts for the willingness of $g$ to engage in an out-group interaction. We assume that players are always willing to engage in an in-group interaction if initiated by another player (which captures the idea that players are always willing to share ideas etc with members of their group. This may not be the case if such interactions are intrinsically costly).

Pairwise invasibility plots for the model are shown in Figure S7-8 below.

Because we are assuming a population in which $N \gg n$, the fitness of the resident $g$ is independent of the mutant $f$, such that the selection gradient only depends on $w_f$, which gives us Eq. 6 in the main text. Calculating this gradient explicitly by differentiating Eq. 6 of the main text gives

$$\frac{\partial s(f, g)}{\partial p_f} = \frac{n}{k} \sum_{k=0}^{n} \left(\frac{n}{k}\right) p_f^{k-1}(1 - p_f)^{n-k-1}(k - np_f) \times$$

$$\left(\frac{k}{l_i}\right) q_i^{l_i} (1 - q_i)^{k-l_i} \left(\frac{n - k}{l_o}\right) (q_o(1 - p_g))^{l_o}(1 - q_o(1 - p_g))^{n-k-l_o} \times$$

$$\frac{\exp[h(l_iB_i + l_oB_o + n\theta)]}{1 + \exp[h(l_iB_i + l_oB_o + n\theta)]}(1 + r(l_iB_i + l_oB_o))$$

which once again we can evaluate numerically to calculate the points of zero selection gradient as shown in Figure 3 of the main text and below. However we can also evaluate Eq. 16 in the special case where the environment is sufficiently bad that $l_iB_i + l_oB_o + n\theta < 0, \forall l_i, l_o$ or sufficiently good $l_iB_i + l_oB_o + n\theta > 0, \forall l_i, l_o$ so that we can approximate the sigmoidal term in Eq. 17 as constant and recover selection gradient

$$\frac{\partial s(f, g)}{\partial p_f} = \sum_{k=0}^{n} \left(\frac{n}{k}\right) p_f^{k-1}(1 - p_f)^{n-k-1}(k - np_f) (1 + nrq_oB_o + kr(q_iB_i - q_oB_o))$$

$$= r(q_iB_i - q_o(1 - p_f)B_o)$$

which, when evaluated at $p_f = p_g$ means that the invasion success of the mutant depends on the resident strategy. In particular there is an equilibrium at $p^* = 1 - \frac{n_pB_o}{q_oB_o}$, which is always a viable strategy provided $q_oB_o > q_iB_i$ i.e out-group interactions have higher intrinsic expected payoff than in-group interactions.
We can evaluate the stability of this equilibrium by taking the second derivative (see Eq. 11 above) which gives

$$\frac{\partial^2 s(f, g)}{\partial f^2} = rq_o B_o.$$  \hspace{1cm} (18)

We see that the equilibrium is always unstable. In addition we note that at the upper boundary, when \(p_f = p_g = 1\) Eq. 17 reduces to \(rq_B\) which is always positive, indicating maximum polarization is always stable in extreme environments under this model. Similarly, at the lower boundary when \(p_f = p_g = 0\) Eq. 17 reduces to \(q_iB_i - q_oB_o\) which is always negative if the expected intrinsic payoff from out-group interactions is greater than from in-group interactions.

Finally in the case of intermediate environments Eq. 17 cannot be analyzed explicitly, although it can be explored numerically as shown in Figure 3 of the main text and in Figure S7 below.

![Figure S7– Model of polarization under which the success of an out-group interaction depends on both the intrinsic probability of success, \(q_o\) and the willingness of other players to engage in out-group interactions, \(1 - p\) i.e. on the strategy of other members of the group or population, as shown in Figure 3 of the main text, with in-group interaction benefit \(B_i = 0.2\). (left) Under the framework of adaptive dynamics, we calculate the selection gradient for invasion by a rare, local mutant in a monomorphic population, We calculate the selection gradient as a function of environment quality \(\theta\) and of resident population strategy \(p\). The direction of the selection gradient and the consequent evolutionary dynamics are shown by the blue (increasing polarization) and red (decreasing polarization) regions with arrows indicating the direction of evolutionary change in \(p\) for a given environment \(\theta\) (right). We show that under these parameters, the system is bistable across all environments b) As a result, a population initialized at a low or high polarization state tends to remain there (black line) regardless of the environment (purple dashed line) tends Each individual is assigned to one of two groups so that all individuals have an in- and an out-group of 500 individuals under the assumptions of the adaptive dynamics model described above (see main text). Innovations, in which individuals try out novel strategies, occur at rate \(\mu = 0.001\) per copying event and new strategies occurred via a deviation around the current strategy of size \(\Delta = 0.01\), plus boundary conditions to ensure strategies remain in the physical range \([0, 1]\). Model parameters and visualization are otherwise as per Figure 2b.

However the stability at the boundaries can be assessed. Taking \(p_g = 1\) we recover
\[
\frac{\partial s(f, g)}{\partial p_f} = nq_i^n \exp[h(nB_i + n\theta)] \frac{1}{1 + \exp[h(nB_i + n\theta)]} (1 + r_nB_i) + \sum_{l_i=0}^{n-1} \left( \frac{n^2(1 - q_i)}{n - l_i} - 1 \right) q_i^n (1 - q_i)^{n-1-l_i} \exp[h(l_iB_i + n\theta)] \frac{1}{1 + \exp[h(l_iB_i + n\theta)]} (1 + rl_iB_i)
\]

Now note that if this quantity is positive when the sigmoidal term is constant it is always positive as the sigmoidal term will always reduce the contribution of terms \( l_i < n(1 - n(1 - q_i)) \) that contribute negative weight to the summation more than it reduces terms \( l_i > n(1 - n(1 - q_i)) \) thus we can assess the stability of the upper boundary by settling the sigmoid constant and equal to 1. We then find from Eq. 18

\[
\frac{\partial s(f, g)}{\partial p_f} = rq_iB_i
\]

as given above. Thus the upper boundary is always stable except in the limit \( q_iB_i \to 0 \) in which case the upper boundary converges with the unstable point \( p^* \).

Finally we consider the stability of the lower boundary \( p_g = 0 \) at which point we find selection gradient

\[
\frac{\partial s(f, g)}{\partial p_g} = \sum_{l_o=0}^{n} \left( \frac{n}{l_o} \right) q_o^n (1 - q_o)^{n-l_o} \exp[h(l_oB_o + n\theta)] \frac{1}{1 + \exp[h(l_oB_o + n\theta)]} (1 + r(l_oB_o)) - \sum_{l_o=0}^{n} \left( \frac{n}{l_o} \right) q_o^n (1 - q_o)^{n-l_o} \exp[h(l_oB_o + n\theta)] \frac{1}{1 + \exp[h(l_oB_o + n\theta)]} (1 + r(l_oB_o))
\]

which can be both positive and negative as illustrated in Figure 3 of the main text and below. Finally we also note that the form of Eq. 17 permits the existence of equilibria for non-boundary values of \( p_g \), such that the system contains multiple stable equilibria. We given an example of such a case below.
3.2 Bi-stability and Multi-stability

The qualitative difference between the first (Case 1) and second (Case 2) models presented in Figure 2 of the main text is the ability of the Case 2 model to sustain multiple stable equilibria. This is shown for the case of local mutations under the framework of adaptive dynamics. However pairwise invasion plots for the same parameter values used to produce Figure 3 of the main text reveal that both equilibria are in fact stable against all invaders (Figure S8). This leads to our conclusion that Case 2 can produce irreversible loss of low-polarization behavior absent coordinated behavioral shifts that bypass the disadvantage faced by rare invaders.

Figure S8 – Pairwise invasability plot in different environments using the default parameters as given in Table 1. We see, just as in Figure 3 of the main text, that low polarization ($p^* = 0$) and high polarization ($p^* = 1$) are both stable in a very good or a very bad environment (left and right plots) but that only high polarization is stable for this choice of parameters in an intermediate environment, which can lead to the irreversible loss of low polarization behavior in a shifting environment. As shown in above, the high polarization equilibrium is never lost, although its basin of attraction can become arbitrarily small.

We also note that Eq. 17 permits the possibility of equilibria that lay in the interior of strategy space i.e for values $0 < p_g < 1$ of the resident strategy. We illustrate the existence of such a stable interior equilibrium in Figure S9 below, and also note its vulnerability to environmental shifts which disrupt the equilibrium and lead to invasion by high or low polarization strategies.
Figure S9– Pairwise invasability plot showing three stable equilibria. Two are globally stable, at $p^* = 0$ (low polarization) and $p^* \approx 0.88$ (high intermediate polarization) while the high polarization equilibrium is seen to be locally stable. The parameters shown are as given in Table 1, with the alteration that a steep threshold ($h = 100$) is required to generate the internal equilibrium.
3.3 Rising inequality and economic decline

To supplement main text Figure 3 we ran simulations in which a smaller proportion of the population, \( \pi = 0.01 \) and \( \pi = 0.001 \), are wealthy. We also ran simulations assuming that the average environment \( \theta_0 \) was increasing while the level of wealth inequality \( \theta_g \) was decreasing (and vice versa). As shown in Figure S10 below, both cases lead to the emergence and entrenchment of polarization as in Figure 3.

Figure S10 – (top row) We ran individual-based simulations in an economic environment with exogenously changing levels of inequality \( \theta_g \) with \( \theta_0 \) fixed and varying \( \pi \). Here we assume sinusoidally varying inequality with \( \theta_0 = 1.0 \) and \( \theta_g \in [0, 1.0] \) and different values of \( \pi \) as shown in the legend. All other parameters are those provided in main text Figure 2c-d (top left) in all three cases the population behavior is qualitatively the same, with polarization emerging and becoming entrenched as (top right) inequality \( \theta_g \) increases. (bottom row) We ran individual-based simulations in an economic environment with exogenously changing levels of inequality \( \theta_g \) and \( \theta_0 \) with fixed \( \pi = 0.1 \). Here we assume sinusoidally varying environment such that \( \theta_0, \theta_g \in [0, 1.0] \) and with \( \text{theta}_g \) increasing as \( \theta_0 \) decreases. All other parameters are those provided in main text Figure 2c-d (bottom left) The population behavior is qualitatively the same as in the other cases, with polarization emerging and becoming entrenched as (bottom right) the environment of the wealthy \( \theta_+ \) increases (dashed line) while the environment of the non-wealthy \( \theta_- \) decreases (solid line) and vice versa.
4 Individual-based Simulations

We performed individual based simulations to test the analytical and numerical predictions from the model presented in the main text and in this supplement. Simulations were performed under the copying process using populations of \( N = 1000 \) individuals with the mean trajectories determined from an ensemble of \( 10^4 \) sample paths. Simulations were run for \( 100N - 200N \) copying events and environmental shifts were simulated by allowing \( \theta \) to change sinusoidally with a period of \( 100N - 200N \) copying events. Fitness was calculated for each individual by randomly assigning all members of the population to one of two groups. To simulate Case 2 out-group interactions were then determined for a given focal individual by randomly choosing a player not from their group with the success of the interaction determined by the chosen player’s strategy and the intrinsic success rate \( q_o \). Mutations were assumed to occur at a rate \( 0.1N \) per copying event, with the target of the mutation chosen randomly from the population. For the non-social model (Case 1) simulations, we allowed global mutations such that the mutating player was assigned a new strategy \( p^\dagger \in [0,1] \). For the Case 2 simulations, we used local mutations such that the target of the mutation had their strategy perturbed by \( \Delta = \pm 0.01 \) with mutations that increase and decrease \( p \) equally likely, and we impose the appropriate boundary conditions to ensure strategies were physical.

5 Data analysis

Here we provide additional analysis and robustness checks for the empirical results presented in the main text (Figure 4). We consider the relationship between affective polarization and inequality at the state level in the USA, across the last three presidential election cycles, 2008-2016. We use data from the ANES and CCES election surveys to measure affective polarization (????) along with data from the Census Bureau and the Federal Elections Commission on Gini coefficients, unemployment rates and vote share at the time of each election in each state (??).

Affective polarization is defined at the individual level as the difference in “warmth”, measured via a “feelings thermometer”, between the major party an individual most identifies with (Republican or Democrat) and the out-party, i.e. the major party the respondent least identifies with. Feelings thermometers have long been a standard part of election surveys, and are administered on a 100 point scale, with 0 corresponding to strong negative feelings towards a party, and 100 corresponding to strong positive feelings. Intuitively, if an individual gives a high score to one party and a low score to another, this indicates a high degree of affective polarization i.e. a large net positive feeling towards a preferred party.

Recent research in political science has stressed that partisanship is a salient social identity as well
as a marker for the various other group identities that have become associated with the major parties (????). Thus, affective polarization is a measure well-suited to evaluating our group-based model.

Individual-level affective polarization scores can be used to calculate an average at the state or national level, which provides a measure of the degree to which the electorate is polarized in their attitudes towards the parties. The in- or out-group favoring behavioral strategies described by our model can be understood as corresponding to this individual-level affective polarization. Since we predict (Figure 3) that inequality can act as a driver of in-group favoring attitudes, we can seek support for this prediction in the affective polarization data. To this end we present three analyses which support the hypothesis that inequality can act as a driver for affective polarization.

5.1 Pooled data

In Figure 4 (main text) we show the correlation between the average polarization at the state level and the state-level Gini coefficient. Results shown are for the pooled data across all three election cycles (50 states + DC for each of three election cycles, for a total of 153 data points). This correlation is significant under a two-way fixed effects model with a single intercept (two-tailed t-test, $t = 5.2$, $p < 0.01$). The state-level data for each election cycle is displayed in Table S2 below, and the code to reproduce the analysis is available via github.
Table S2 – Pooled data for each state during each presidential election cycle 2008-2016. Affective polarization is calculated as the average across the respondents within the state during the election cycle and normalized to produce a value in the range [0,1].
5.2 Election-specific intercepts

Next we consider the pooled data under a two-way fixed effects model with election-specific intercepts, which accounts for the systematic change in Gini coefficient across states over the 8-year period 2008-2016. This correlation is significant when all data is included (two-tailed t-test, $t = 5.7$, $p < 0.01$). When we exclude states with fewer than 10 respondents in a given election cycle the correlation remains significant (two-tailed t-test, $t = 4.4$, $p < 0.01$ – this is the case shown in Figure S11) and the result continues to hold when states with fewer than 50 respondents are excluded (two-tailed t-test, $t = 2.1$, $p < 0.05$).

![Figure S11](image)

Figure S11 – We show the correlation for the pooled data across all three election cycles between state-level affective polarization estimated from (???) and state-level Gini coefficient taken from (?), under a fixed-effects model with election-specific intercepts. Data shown exclude states with fewer than 10 respondents.

5.3 Individual-level data

Finally we look at the correlation between Gini and affective polarization at the level of the individual respondent, controlling for the effect of race and education of the individual and unemployment rate, and partisan lean (measured by the percentage vote received by Republicans vs Democrats in the corresponding presidential election) of the state where they reside. Race is encoded as a dummy variable (white vs non-white) while education is measured on a 6-point scale under the standard election survey. Gini, unemployment rate and partisanship for each individual are given according to the value
for the state at the time of the election. The results are shown in Table S3 with Gini and education showing a significant positive individual-level correlation with affective polarization, and race showing a significant negative correlation. Race and education both display strong correlations, but are also individual-specific measures. This indicates that the background level of inequality an individual experiences is positively correlated with the degree of polarization they express, but this effect is strongly modulated by individual-level variations in race and education.

| Independent variables | Correlation     | p-value  |
|-----------------------|-----------------|----------|
| GINI                  | 0.58* (0.16)    | 0.0003   |
| Unemployment rate     | 0.0002 (0.001)  | 0.87     |
| Vote share (R/D)      | 0.05 (0.03)     | 0.11     |
| Race (white/non-white)| -0.04* (0.005)  | 7x10^-16 |
| Education             | 0.009* (0.003)  | 0.0002   |
| N                     | 13045           | -        |
| Adjusted $R^2$        | 0.00706         | -        |

Table S3 – Individual-level correlations. The second column shows the correlation coefficient along with the standard error in parentheses. Correlations significant with $p < 0.01$ are indicated with an asterisk, while the calculated p-values are shown in the third column.