The nonlinear feedback dynamics of asymmetric political polarization

Naomi Ehrich Leonard[a,1,2], Keena Lipsitz[b,1,2], Anastasia Bizyaeva[a], Alessio Franci[c], Yphtach Lelkes[d]

[a]Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544; [b]Department of Political Science, Queens College and The Graduate Center, City University of New York, Flushing, NY 11367; [c]Department of Mathematics, National Autonomous University of Mexico, 04510 Mexico City, Mexico; and [d]Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA 19104

Edited by Robert Axelrod, University of Michigan, Ann Arbor, MI, and approved August 18, 2021 (received for review March 2, 2021)

Using a general model of opinion dynamics, we conduct a systematic investigation of key mechanisms driving elite polarization in the United States. We demonstrate that the self-reinforcing nature of elite-level processes can explain this polarization, with voter preferences accounting for its asymmetric nature. Our analysis suggests that subtle differences in the frequency and amplitude with which public opinion shifts left and right over time may have a differential effect on the self-reinforcing processes of elites, causing Republicans to polarize more quickly than Democrats. We find that as self-reinforcement approaches a critical threshold, polarization speeds up. Republicans appear to have crossed that threshold while Democrats are currently approaching it.

political polarization | nonlinear dynamics | political elites | public opinion | bifurcations

American policymakers are more polarized today than any time since the end of the Civil War. After a period of bipartisanship following World War II, Republican and Democratic political elites, typically defined as legislators and other elected officials, diverged dramatically. The resulting polarization threatens the long-term stability of America and "has triggered the epidemic of norm breaking that now challenges our democracy" (ref. 1, p. 204).

Despite clear evidence of its existence, explanations for polarization, such as changes to the media environment, interest group influence, and institutional factors, are often presented in a piecemeal fashion. We offer a unified model of mass and elite polarization that subsumes many of these explanations. In particular, we focus on two elite-level positive feedback mechanisms underpinning polarization: party self-reinforcement and reflexive partisanship. Party self-reinforcement involves sources of elite polarization being driven themselves by the polarization they create. Elites engage in reflexive partisanship when they support policies simply because the other side opposes them (2). We show that the former is more consistent with historical trends in elite polarization than the latter. In addition, we demonstrate that thermostatic input from voters drives temporal and asymmetric aspects of these polarization dynamics. Doing so demonstrates that elite polarization is not, in fact, "disconnected" from public opinion (3).

We use our model as a testbed to examine processes co-occurring in a two-party democratic system between elites, as well as between elites and citizens, and to explore the temporal aspects of these processes. The model is parsimonious, and thus analytically tractable, allowing us to focus in a principled way on the essential mechanisms that drive the complex process of polarization. We use the model to systematically test and compare hypotheses and rule out those hypotheses that yield temporal dynamics that are inconsistent with historical data.

Subsequently, we offer several contributions to the polarization literature. First, we find that, among the possible explanations examined here, polarization can be best explained by a positive feedback mechanism, which, by definition, yields a pattern of increasing returns (4). Positive feedback amplifies variations in ideological position while negative feedback attenuates variations in ideological position. As positive feedback grows, it can reach a critical threshold at which point amplifying and attenuating effects are balanced. When positive feedback, in the form of party self-reinforcement or reflexive partisanship, crosses that threshold, then ideological positions can rapidly become extreme. Second, we find that elite-level self-reinforcement can explain polarization in the United States. The asymmetry in the polarization comes from asymmetry in self-reinforcement driven by the dynamics of policy mood—an aggregate measure of the public's ideology—wherein voters shift more frequently and for a longer duration to the right than to the left. Third, we rule out reflexive partisanship as a dominant mechanism since it does not explain asymmetric polarization even when driven by policy mood. The fact that reflexive partisanship is a mutual response undermines its asymmetric effect. Relatedly, we also demonstrate that the breakdown in norms of bipartisanship, i.e., the inverse of reflexive partisanship, cannot account for the rise of asymmetric polarization. Fourth, we rule out the (null) hypothesis that elites are merely responding to policy mood without a positive feedback mechanism.

Significance

Political polarization threatens democracy in America. This article helps us illuminate what drives it, as well as what factors account for its asymmetric nature. In particular, we focus on positive feedback among members of Congress as the key mechanism of polarization. We show how public opinion, which responds to the laws legislators make, in turn drives the feedback dynamics of political elites. Specifically, we find that voters' "policy mood," i.e., whether public opinion leans in a more liberal or conservative direction, drives asymmetries in elite polarization over time. Our model also demonstrates that once self-reinforcing processes among elites reach a critical threshold, polarization rapidly accelerates. By tying together elite and voter dynamics, this paper presents a unified theory of political polarization.

Author contributions: N.E.L., K.L., A.B., A.F., and Y.L. designed research; N.E.L., K.L., A.B., A.F., and Y.L. performed research; N.E.L., A.B., and A.F. contributed new reagents/analytic tools; N.E.L., K.L., A.B., A.F., and Y.L. analyzed data; N.E.L., K.L., A.B., A.F., and Y.L. wrote the paper; and N.E.L. and K.L. conceived the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]N.E.L. and K.L. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: naomi@princeton.edu or keena.lipsitz@qc.cuny.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/ doi:10.1073/pnas.2102149118/-/DCSupplemental.

Published December 6, 2021.

Asymmetric Polarization

Any model of elite polarization must offer insight into one of its more notable characteristics: its asymmetry. While elites of both parties have polarized ideologically, Fig. 1 uses dynamic, weighted, nominal three-step estimation (DW-NOMINATE) scores to illustrate that this process has been more precipitous for Republicans in the US Congress, yielding a position that is currently more extreme than the Democratic Party's (5–8). DW-NOMINATE uses roll-call votes to position legislators, based on how often they vote with or against one another, in a latent space that is most often referred to as ideology.* Most explanations for this asymmetry focus on processes occurring at the elite level both inside and outside of Congress. In Congress, researchers point to the emergence of successful conservative factions (12), as well as a particularly aggressive governing style among party leaders who use party discipline to harness members to their more extreme agendas (8).

Others point to elite-level factors occurring outside of Congress. Intellectuals or "coalition merchants" on both the right and the left began to craft distinct conservative and liberal ideologies in the 1930s (13), but the project of those on the right was more ambitious and successful in creating ideological homogeneity among interest groups, media outlets, and think tanks that both fueled and reinforced the rightward movement of the Republican Party (7). More recently, individuals and organizations with ties to the conservative movement, such as the Koch brothers and Americans for Prosperity, have used campaign funding as a means of keeping GOP members in line (14). While Democrats have no shortage of wealthy backers, their heterogeneous policy commitments may have offered some resistance to the ideological pressure exerted by donors. Such cross-cutting pressures are largely absent for business-friendly Republicans (5).

Scholars have considered the possibility that citizens might be at least partially responsible for the asymmetric nature of elite polarization, but have largely dismissed the notion because it is not clear that Americans themselves are polarized (15, 16). Republican voters do appear to be more loyal to conservative doctrine, however, and less tolerant of compromise than Democratic voters (17).

Despite the insights these explanations offer, they are unsatisfactory for a variety of reasons. Some elite-focused explanations, such as the rise of conservative media and donor networks on the right, identify processes that began too recently to explain asymmetric polarization that began much earlier, although they almost certainly exacerbated the trend. Citizen-focused explanations, on the other hand, tend to be far too preoccupied with Republican voters and ignore almost two-thirds of the electorate, including independents and Democrats. Finally, most of these accounts do not recognize how processes occurring at the elite and citizen level are interconnected.

Self-Reinforcement, Reflexive Partisanship, and Additive Response Among Elites

Our model explains polarization through an interplay of positive feedback and negative feedback. Changes in the balance between positive and negative feedback are driven by citizen political preferences as represented by policy mood. We consider two fundamental positive feedback mechanisms that have been proposed in the literature to explain elite polarization: party self-reinforcement (hypothesis A) and reflexive partisanship (hypothesis B). We also consider the null hypothesis in which elites respond additively to citizens with no positive feedback mechanism (hypothesis C).

Self-Reinforcement. Many polarizing processes exhibit positive feedback in the form of "a powerful self-reinforcing logic" (ref. 5, p. 45). For example, Pierson and Schickler (5) argue that as the parties polarize, interest groups have an incentive to join one of the parties' coalitions. Once they do, their goal becomes to help it win at any cost. This can involve punishing defectors and eliminating moderating voices. This same logic, however, applies to any group of actors involved in the polarization process. For example, extremist party leaders can punish moderate members by backing their more extreme opponents in primaries. They can also employ party discipline to keep such members in line. Elite polarization may also trigger anger and distrust of the other side among voters, who, in turn, may elect more extreme representatives (18). When they do, it has the potential to exacerbate self-reinforcing elite polarizing processes.

Reflexive Partisanship. Political elites may also be engaging in mutual polarization or reflexive partisanship (2), wherein one party will oppose a policy merely because the other side supports it. Policymakers often find it politically useful to "exploit and deepen division rather than seeking common ground" (ref. 2, p. 193). This may yield a cycle wherein one party becomes more extreme as a response to the other party's increasing extremity.

Additive Response. We also examine the possibility that political elites respond to voter preferences without any positive feedback mechanism. That is, we consider that elites may be adjusting their ideological positions through an additive response to changing signals from voters but independently of how moderate or extreme are their own and the other party's positions.

Voter Policy Mood as Input to Elite Polarization

According to thermostatic models of public opinion, the relationship between policymaking and public opinion is dynamic (19). Citizens respond to policy outputs, and when these outputs are more liberal/conservative than their preferences, they communicate their desire for more conservative/liberal policies through surveys, through activism, and by voting out incumbents. Parties either adapt to these signals by revising their platforms, supporting different candidates, and voting for policies the public wants or they continue to lose elections (19, 20). A widely used measure of public opinion is referred to as the "policy mood," which captures where aggregate public opinion falls along a liberal–conservative dimension (21). We hypothesize that asymmetries in the thermostatic dynamics of public opinion may explain the unbalanced trends we see in Fig. 1; if conservative swings in policy mood are more substantial, numerous, or prolonged than liberal swings, they will amplify the responsive dynamics of Republican elites more substantially than those of Democratic elites.

By feeding policy mood data into a model of elite opinion dynamics, we demonstrate that such asymmetries can indeed account for the polarization trends we see in historical data if political elites are assumed to respond through a self-reinforcing positive feedback (hypothesis A [hyp. A]). This suggests citizens do not have to be polarized themselves to contribute to elite polarization; they can contribute simply by amplifying the self-reinforcement dynamics of one of the parties.

Reflexive partisanship (hyp. B) can also explain polarization, but we show that it does not do as well as party self-reinforcement in capturing the trends in historical data; most

*There is currently a debate about the extent to which ideal point estimates, such as DW-NOMINATE scores, reflect ideological or partisan conflict (9, 10) or merely voting coalitions (11). The fact that this article shows ideological swings in aggregate public opinion contribute to polarization dynamics, as reflected in DW-NOMINATE scores, suggests they do in fact capture some element of ideology. If DW-NOMINATE is, in fact, solely a measure of voting coalitions, our model is still valuable as it demonstrates that the political behavior of governmental elites is responsive to policy mood. (We are grateful to an anonymous reviewer for making this point.)

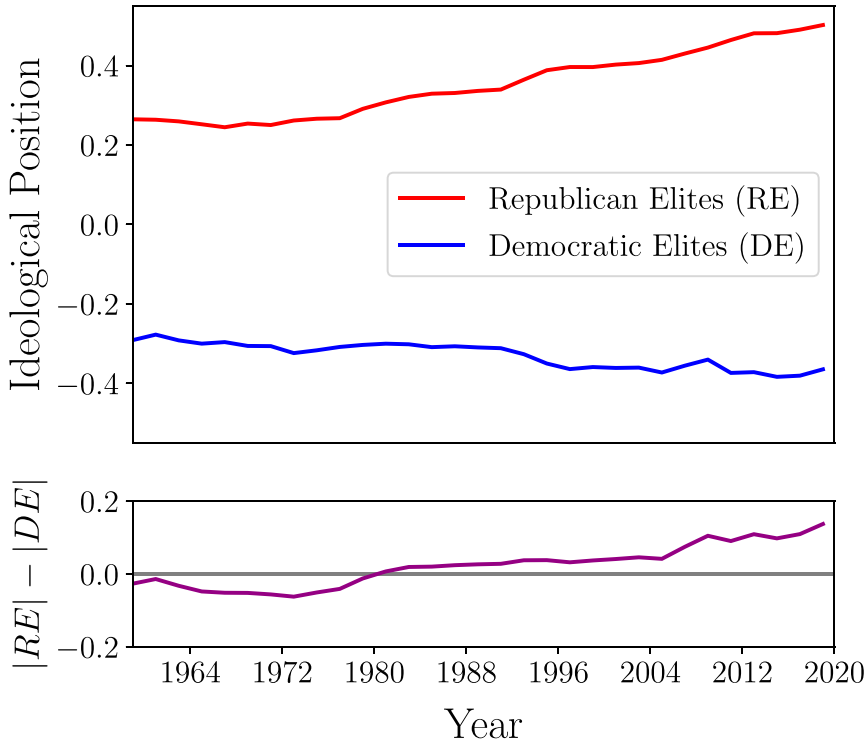


Fig. 1. DW-NOMINATE scores (first dimension) averaged across US Senate and House of Representatives.

notably, it falls short in explaining the asymmetry in polarization. We further demonstrate that polarization is not well explained if it is assumed that political elites respond to policy mood but not through a positive feedback (hyp. C). This is to be expected since this hypothesis fails to provide a mechanism for the observed pattern of increasing returns (4).

Model

The model, illustrated in Fig. 2, describes the temporal dynamics of the ideological position of each of the two major elite populations in the US Congress. The influence of the changing position of US voters on the elites is introduced as an input to these dynamics. The model equations derive from the general model of opinion dynamics presented in refs. 22 and 23, which provides a versatile testbed for studying a wide range of behaviors in terms of a small number of parameters, even for a large number of independent decision makers forming opinions about multiple options (a specialization of this model is used in ref. 24). The versatility and analytical tractability of the model are central to our purpose: a principled and systematic investigation of key mechanisms that can help explain political polarization in the United States. The model is well suited to distinguishing among hypotheses, since it provides evidence to rule out a hypothesis that contradicts empirical data. The model can likewise be used to derive and evaluate potential strategies for depolarization.

The general model of opinion dynamics (22, 23) describes processes that are inherently nonlinear: Inputs and exchanges in the model dynamics can strengthen key factors over time, but the resulting change in behavior is exhibited only once the strength of these factors increases beyond a critical threshold. Dynamics of this kind are observed ubiquitously in physical and biological systems as well as in social systems including, notably, in political polarization in the US Congress as measured by DW-NOMINATE scores (Fig. 1). See *SI Appendix*, section S1.A for details on the general model.

Elite Dynamics

We define $x_r(t)$ and $x_d(t)$ to represent the scalar ideological position of the Republican elites and the Democratic elites, respectively, in the US Congress at time t, as measured in years. Ideological position is interpreted as center if it takes the value zero, liberal (left of center) if it takes a negative value, and conservative (right of center) if it takes a positive value.

For the Republican elite, we let $b_r(t) \geq 0$ represent an underlying conservative bias in ideological position at time t. Republican self-reinforcement level $\alpha_r(t) \geq 0$ denotes the strength, at time t, with which the Republican elite reinforces its own

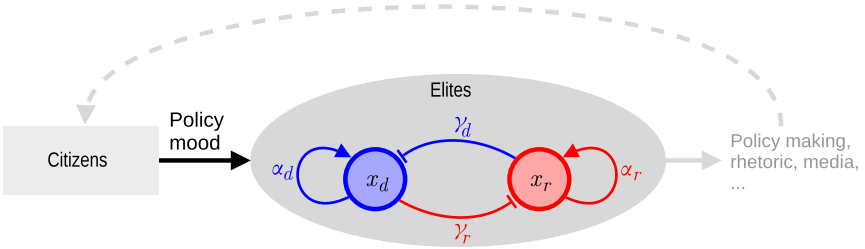


Fig. 2. Illustration of the model.

ideological position. Republican reflexive partisanship level $\gamma_r(t) \geq 0$ denotes the strength, at time t, with which the Republican elite adjusts its ideological position in direct response to the changing ideological position of the Democratic elite. For the Democratic elite, we let $b_d(t) \leq 0$ represent an underlying liberal bias in ideological position at time t. Democratic self-reinforcement level $\alpha_d(t) \geq 0$ denotes the strength, at time t, with which the Democratic elite reinforces its own ideological position. Democratic reflexive partisanship level $\gamma_d(t) \geq 0$ denotes the strength, at time t, with which the Democratic elite adjusts its ideological position in direct response to the changing ideological position of the Republican elite. The parameter $\tau_x > 0$ represents the time scale (in years) of changing elite ideological position.

The model defines the rate of change of each of the two elite populations' ideological positions as follows:

$$\tau_x \frac{dx_r}{dt} = \tanh(\alpha_r x_r - \gamma_r x_d) - x_r + b_r \qquad [1]$$

$$\tau_x \frac{dx_d}{dt} = \tanh(\alpha_d x_d - \gamma_d x_r) - x_d + b_d. \qquad [2]$$

For derivation from the general model see *SI Appendix*, section S1.A.

For $\alpha_r > 0$ and $\alpha_d > 0$, the products $\alpha_r x_r$ in **[1]** and $\alpha_d x_d$ in **[2]** represent self-reinforcement of ideological position among the Republican elites and the Democratic elites, respectively. To see that these terms provide positive feedback, assume that $x_r > 0$ and $x_d < 0$ at time t. Then $\alpha_r x_r > 0$ and drives x_r to be more positive (conservative). Likewise, $\alpha_d x_d < 0$ and drives x_d to be more negative (liberal).

For $\gamma_r > 0$ and $\gamma_d > 0$, the products $-\gamma_r x_d$ in **[1]** and $-\gamma_d x_r$ in **[2]** represent reflexive partisanship by the Republican elites and the Democratic elites, respectively. To see that these terms provide positive feedback, assume that $x_r > 0$ and $x_d < 0$ at time t. Then $-\gamma_r x_d > 0$ and this drives x_r to be more positive (conservative). Likewise, $-\gamma_d x_r < 0$ and this drives x_d to be more negative (liberal).

In the absence of a positive feedback response of either type, we have that $\alpha_r = \alpha_d = \gamma_r = \gamma_d = 0$ and the dynamics in **[1]** and **[2]** are linear. Then x_r responds additively to b_r and x_d responds additively to b_d, where we interpret $b_r(t)$ and $b_d(t)$ as input signals.

The hyperbolic tangent function "tanh" provides a smooth, nonlinear "saturating" bound (between -1 and $+1$) on the value of its argument, which in **[1]** and **[2]** is the sum of the positive feedback terms from self-reinforcement and reflexive partisanship. This saturating function slows down the acceleration of ideological position away from center when the positive feedback gets very large. Without the tanh function, the ideological positions could diverge without bound and at rates that are inconsistent with empirical data. We note further that saturating functions on positive feedback terms are applied in a wide range of models of natural phenomena, including political polarization (25).

The terms $-x_r$ in **[1]** and $-x_d$ in **[2]** provide a negative feedback that represents a nominal resistance to changing ideological position away from center. Negative feedback and positive feedback counteract each other, and there is a critical threshold where they are of equal magnitude. When positive feedback is smaller than negative feedback, ideological position is regulated close to center (tracking the bias value). When positive feedback is larger than negative feedback, ideological position can diverge significantly away from the center (and the bias value). A rigorous description and proof of these behaviors using bifurcation theory are provided in *SI Appendix*, section S1.B; see also *SI Appendix*, Fig. S1.

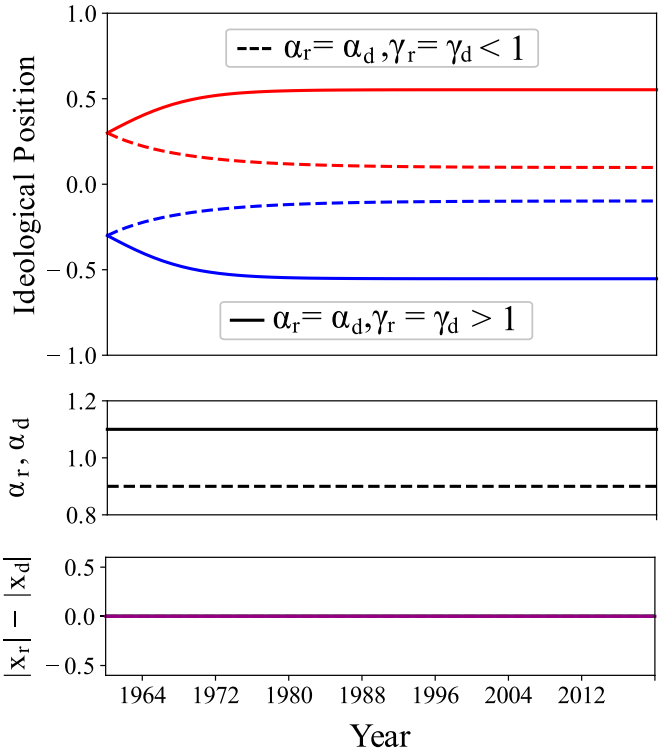


Fig. 3. Comparison of below-threshold and above-threshold positive feedback illustrates the interplay of positive and negative feedback mechanisms in the nonlinear model dynamics (**[1]** and **[2]**). Plotted as a function of time in years starting in 1959 are (*Top*) x_r (red) and x_d (blue); (*Middle*) $\alpha_r, \alpha_d, \gamma_r, \gamma_d$; and (*Bottom*) $|x_r| - |x_d|$. Initial conditions are $x_r(1959) = 0.3$ and $x_d(1959) = -0.3$. Parameters are $\tau_x = 1$ y, $b_r = 0.01$, $b_d = -0.01$. Dashed lines correspond to $\alpha_r = \alpha_d = \alpha = 0.9$ and $\gamma_r = \gamma_d = 0$ or $\gamma_r = \gamma_d = \gamma = 0.9$ and $\alpha_r = \alpha_d = 0$. Because $\alpha < 1$ or $\gamma < 1$, negative feedback dominates and x_r and x_d are regulated toward the center. Solid lines correspond to $\alpha_r = \alpha_d = \alpha = 1.1$ and $\gamma_r = \gamma_d = 0$ or $\gamma_r = \gamma_d = \gamma = 1.1$ and $\alpha_r = \alpha_d = 0$. Because $\alpha > 1$ or $\gamma > 1$, positive feedback dominates and x_r and x_d diverge significantly away from center.

In Fig. 3, we illustrate how the modeled temporal dynamics of ideological position differ for positive feedback below the critical threshold compared to above the critical threshold. We use symmetric initial conditions ($x_r = 0.3$ and $x_d = -0.3$ in 1959) and symmetric constant biases ($b_r = 0.01$ and $b_d = -0.01$). We set $\tau_x = 1$ y as a default for comparative purposes only. We let $\alpha_r = \alpha_d = \alpha \neq 0$ and $\gamma_r = \gamma_d = 0$. The critical value $\alpha = 1$ corresponds to equal positive and negative feedback terms. When $\alpha = 0.9 < 1$, negative feedback dominates and x_r and x_d converge to near-center values (dashed lines). When $\alpha = 1.1 > 1$, positive feedback dominates and x_r and x_d diverge significantly away from the center (solid lines). Because of the symmetry in parameters, the plots are identical for $\gamma_r = \gamma_d = \gamma \neq 0$ and $\alpha_r = \alpha_d = 0$ if $\gamma = 0.9 < 1$ (dashed lines) and $\gamma = 1.1 > 1$ (solid lines).

In Fig. 4, we compare the modeled temporal dynamics in **[1]** and **[2]** for the three hypotheses on the response mechanism for political elites. We illustrate how x_r and x_d evolve over time with a linear increase of party self-reinforcement levels α_r, α_d (hyp. A); reflexive partisanship levels γ_r, γ_d (hyp. B); and additive inputs b_r, b_d (hyp. C), in Fig. 4 *A*–*C*, respectively.

In Fig. 4*A*, party self-reinforcement levels α_r and α_d increase linearly in time with reflexive partisanship levels $\gamma_r = \gamma_d = 0$ and biases $b_r = 0.01$ and $b_d = -0.01$. When the positive feedback is below the critical threshold for either party, i.e., $\alpha_r < 1$ (respectively, $\alpha_d < 1$), then x_r (respectively, x_d) remains close to the

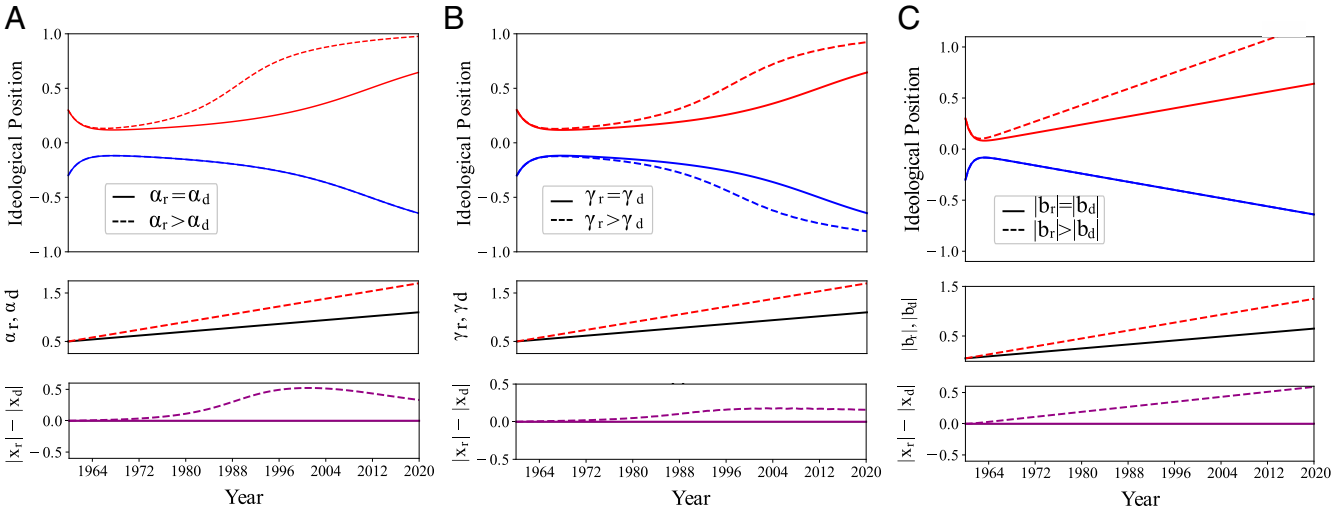


Fig. 4. Comparison of the model dynamics (**[1]** and **[2]**) for the three hypotheses on political elite response mechanism. Plotted as a function of time in years starting in 1959 are (*A–C*, *Top*) x_r (red) and x_d (blue); (*A–C*, *Middle*) $\alpha_r, \alpha_d, \gamma_r, \gamma_d$, $|b_r|$, $|b_d|$ (black is overlay of solid red, solid blue, and dashed blue); and (*A–C*, *Bottom*) $|x_r| - |x_d|$ (purple). For all simulations, $x_r(1959) = 0.3$, $x_d(1959) = -0.3$, and $\tau_x = 1$ y. (*A*) Positive feedback from self-reinforcement (hyp. A): comparison of symmetric versus asymmetric increase in α_r and α_d over time. $b_r = 0.05$, $b_d = -0.05$, $\gamma_r = \gamma_d = 0$, and $\alpha_d(t) = (0.01)(t - 1959) + 0.3$. Solid lines correspond to the symmetric conditions $\alpha_r(t) = \alpha_d(t) = \alpha(t)$. While $\alpha(t) < 1$, x_r and x_d remain near the center, and when $\alpha(t) > 1$, they diverge symmetrically away from the center. Dashed lines correspond to the asymmetric conditions $\alpha_r(t) = (0.02)(t - 1959) + 0.3$. In this case x_r diverges sooner and ultimately more significantly than x_d since $\alpha_r > 1$ sooner than $\alpha_d > 1$. (*B*) Positive feedback from reflexive partisanship (hyp. B): comparison of symmetric versus asymmetric increase in γ_r and γ_d over time. Conditions are identical to those for *A* except that the roles of α_r and γ_r are swapped and the roles of α_d and γ_d are swapped. For the symmetric conditions (solid lines), x_r and x_d diverge symmetrically away from the center. For the asymmetric conditions (dashed lines), the asymmetry in the divergence of x_r and x_d is not as significant as in *A*. (*C*) Additive response (hyp. C): comparison of symmetric versus asymmetric increase in $|b_r|$ and $|b_d|$ over time. $\alpha_r = \alpha_d = \gamma_r = \gamma_d = 0$, and $b_d(t) = -\Big((0.01)(t - 1959) + 0.05\Big)$. Solid lines correspond to the symmetric conditions $b_r(t) = |b_d(t)|$ and dashed lines to $b_r(t) = (0.02)(t - 1959) + 0.05$. x_r and x_d linearly track b_r and b_d, respectively.

small bias; whereas after the positive feedback crosses above the critical threshold, i.e., $\alpha_r > 1$ (respectively, $\alpha_d > 1$), then x_r (respectively, magnitude of x_d) grows significantly. It follows then that when α_r and α_d change symmetrically (solid lines), i.e., when $\alpha_r(t) = \alpha_d(t)$, polarization is symmetric, and when α_r and α_d change asymmetrically (dashed lines), e.g., when α_r increases at twice the rate as α_d, polarization is asymmetric. We can also observe in Fig. 4*A* that x_r (respectively, x_d) is very sensitive to changes in α_r (respectively, α_d) near the critical value $\alpha_r = 1$ (respectively, $\alpha_d = 1$) and fairly insensitive to changes away from the critical value.

Fig. 4*B* shows the analogous simulation results for reflexive partisanship levels γ_r, γ_d increasing linearly in time with self-reinforcement levels $\alpha_r = \alpha_d = 0$ and biases $b_r = 0.01$ and $b_d = -0.01$. When γ_r and γ_d change symmetrically (solid lines), i.e., when $\gamma_r(t) = \gamma_d(t)$, polarization is symmetric, as in Fig. 4*A*. However, when γ_r and γ_d change asymmetrically (dashed lines), e.g., when γ_r increases at twice the rate as γ_d, polarization is still much less asymmetric than in Fig. 4*A*. This lack of asymmetry in the temporal evolution of x_r and x_d can be attributed to the mutual influence of one position on the other. Even in the case that γ_d is significantly smaller than γ_r, if x_r grows large, then the positive feedback term $-\gamma_d x_r$ that drives up the magnitude of x_d in **[2]** can get large enough to dominate the negative feedback and cause x_d to diverge. This explains how x_d (dashed blue line) in Fig. 4*B* starts to polarize even while $\gamma_d < 1$ and as a result the asymmetry in the polarization is relatively small. We show in *SI Appendix*, Fig. S2 that γ_r must be larger, by several orders of magnitude, than γ_d to get the same magnitude of asymmetry in polarization as observed in Fig. 4*A*.

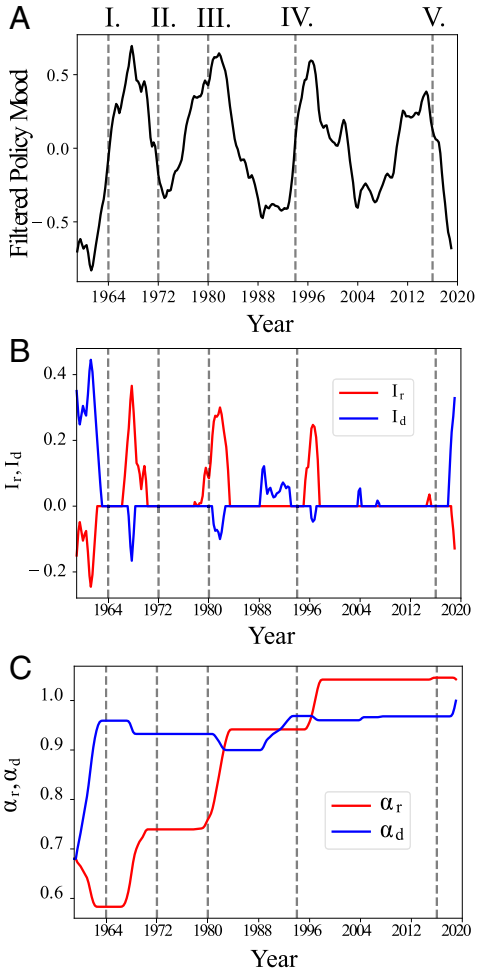
Fig. 4*C* shows simulation results for the magnitude of inputs $|b_r(t)|$ and $|b_d(t)|$ increasing linearly in time with self-reinforcement and reflexive partisanship levels $\alpha_r = \alpha_d = \gamma_r = \gamma_d = 0$. For this hypothesis, x_r and x_d track the linear trajectories of b_r and b_d, respectively. There is no positive feedback and thus no critical threshold or increasing returns.

The model in **[1]** and **[2]** can also be derived as the population-average model reduction of an agent-based model that describes the evolution over time of each individual elite member's ideological position. This agent-based model is described in *SI Appendix*, section S4. In *SI Appendix*, Figs. S9–S12, we provide simulations of the agent-based model with 50 Republican elites and 50 Democratic elites and show how our low-dimensional model of a Republican elite population and a Democratic elite population well represents a model of all 100 individuals, even when small individual differences in the single-agent dynamics are present in each group.

Voters and Policy Mood Input

To determine the role that aggregate public opinion preferences have played in driving asymmetric elite polarization, we seed our model of opinion dynamics with James Stimson's (26) annual policy mood data. To create this measure, Stimson collected a large number of domestic policy survey questions and used a dyad ratio algorithm to extract latent dimensions of public opinion. The policy mood measure used here captures the first dimension and is commonly referred to as a left–right measure of ideology (21).

In modeling how members of the US Congress adapt their ideological positions to policy mood (PM) variations (20), it is natural to assume that they ignore minute jumps in public opinion. To model the effective PM input from voters to elite, we thus pass the PM data through a cascade of two filters (see *SI Appendix*, section S3 for details). The first one is a first-order high-pass filter, which extracts PM variations and removes any PM offset, i.e., puts the "zero" level at the average of the PM over time. The second one is a first-order low-pass filter, which



I. Civil Rights Act; **II.** Watergate Scandal; **III.** Reagan Elected; **IV.** Gingrich and Republicans take House; **V.** Trump Elected

Fig. 5. Illustration of how policy mood is used as an input to the model of elite dynamics (**[1]** and **[2]**). Plotted as a function of time in years are (*A*) the filtered policy mood PM_f; (*B*) policy mood inputs I_r and I_d, each of which is a function f (**[5]**) of the filtered policy mood PM_f; and (*C*) the response of α_r and α_d to the policy mood input as modeled by **[6]** and **[7]** with $U_r = L_r = U_d = L_d = 0.45$. Vertical lines mark five notable events in recent US history. Observe that α_r crosses the critical threshold of 1 just after event IV (Gingrich and Republicans take the House). Initial conditions and model parameters: $I_0 = 0.1$, $\alpha_r(t_0) = \alpha_d(t_0) = 0.68$, $k_\alpha = 0.25$, and $t_0 = 1959$.

removes high-frequency PM jitter. The PM data and the resulting filtered PM signal, $PM_f(t)$, are shown in *SI Appendix*, Fig. S8.

$PM_f(t)$ is also plotted in Fig. 5*A*. While a higher policy mood score usually indicates a more liberal mood, here we reverse the scale so it is consistent with the DW-NOMINATE scale. Observe that peaks (conservative swings) and valleys (liberal swings) of the filtered policy mood signal in Fig. 5*A* capture relevant historical events remarkably well. Conservative peaks occur just after passage of the 1964 Civil Rights Act, during the Reagan and Republican "Revolutions" of the 1980s and 1990s, and just prior to the election of President Trump. Liberal valleys occur in the middle of the Civil Rights Movement, following Watergate, in the waning years of the Reagan and George W. Bush presidencies, and recently as we approached the end of Trump's single term in office. This is no surprise, as scholars have noted that policy mood peaks and valleys often coincide with a new party assuming the presidency (20, 27).

The policy mood inputs $I_r(t)$ and $I_d(t)$ to Republican and Democratic elites are obtained by transforming $PM_f(t)$ through a generic nonlinearity

$$I_r(t) = f\Big(PM_f(t) + I_0\Big), \qquad [3]$$

$$I_d(t) = f\Big(-\Big(PM_f(t) - I_0\Big)\Big), \qquad [4]$$

where f is a function such that $f(0) = 0$ and $I_0 > 0$ is the basal elite ideological drive.

Because elites are sensitive only to policy mood variations, and thus insensitive to any policy mood offset in either direction, we have filtered the PM so that the signal $PM_f(t)$ has (close to) zero average. Doing so, however, implies that, in the long run, a linear function f in the definitions **3** and **4** of the policy mood inputs to elites, $I_r(t)$ and $I_d(t)$, cannot lead to any marked asymmetry in polarization. We assume that elites possess a threshold above which they become sensitive to policy mood swings and below which they are insensitive to them. This dead zone is akin to an "electoral blind spot" or "the policy region over which aggregate electorates do not enforce their preferences" (ref. 28, p. 577). Party elites know that voters will not enforce their preferences in this zone because they do not know enough about policy to be able to tell relatively moderate policy proposals apart (28) and because many issues are not salient enough to attract voters' attention in the first place (29). Thus, only more extreme swings in policy mood will force elites to take notice and respond.

By assuming the existence of this dead zone, we define f to be the nonlinear function

$$f(x;U,L) = \begin{cases} x - U, & \text{if} x \geq U \\ 0, & \text{if} -L < x < U \\ x + L, & \text{if} x \leq -L, \end{cases} \qquad [5]$$

where $U \geq 0$ and $L \geq 0$ are upper and lower sensitivity thresholds, respectively. To allow for differences between parties we let U_r and L_r be the sensitivity thresholds for the Republicans and U_d and L_d those for the Democrats. Then, the policy mood input $I_r(t)$ is as defined in **[3]** with $f(x) = f(x;U_r,L_r)$ and $I_d(t)$ as in **[4]** with $f(x) = f(x;U_d,L_d)$.

Policy Mood Input Drives Elite Self-Reinforcement Dynamics. We first consider hyp. A. To account for the thermostatic adaptation of policymakers to policy mood, we let the policy mood inputs $I_r(t)$ and $I_d(t)$ affect elite dynamics through the self-reinforcing levels α_r and α_d. Elites cannot respond to changes in policy mood instantaneously. Their response must take the form of either adaptation, which involves advocating and enacting new policies, or selection, which involves electing new representatives. Both processes take time, which is why responsiveness is usually gradual and incremental (30).

To model the overall sensitivity of elite dynamics to PM, the rates of change of α_r and α_d are proportional to the policy mood inputs with the same proportionality constant k_α:

$$\frac{d\alpha_r}{dt} = k_\alpha I_r(t) \qquad [6]$$

$$\frac{d\alpha_d}{dt} = k_\alpha I_d(t). \qquad [7]$$

Motivated by the empirical data (20, 27), the model captures a conservative/liberal shift of the public leading to a conservative/liberal shift of both parties of Congress through modulation of their ideological self-reinforcement.

The central hypothesis we propose with our modeling assumptions is that a large conservative swing in policy mood $[PM_f(t) > 0]$ leads to an increase in the Republican elite ideological self-reinforcement if $PM_f(t) + I_0 \geq U_r$ and a decrease in

the Democratic elite ideological self-reinforcement if $PM_f(t) - I_0 \geq L_d$. Likewise, a large liberal swing in policy mood $[PM_f(t) < 0]$ leads to an increase in the Democratic elite ideological self-reinforcement if $-(PM_f(t) - I_0) \geq U_d$ and a decrease in the Republican elite ideological self-reinforcement if $-(PM_f(t) + I_0) > L_r$. And this can lead to a (possibly asymmetric) increase of the net polarizing positive feedback.

Policy Mood Input Drives Elite Reflexive Partisanship Dynamics. To test hyp. B we consider our central hypothesis applied to reflexive partisanship levels γ_r and γ_d rather than self-reinforcement levels. To model this we apply $I_r(t)$ and $I_d(t)$ as input to dynamic changes in γ_r and γ_d, analogous to (and instead of) **[6]** and **[7]** with $\alpha_d = \alpha_r = 0$ and $|\, b_d \,| = |\, b_r \,|$ small.

Policy Mood Input Drives Elite Additive Input Dynamics. To test hyp. C we consider our central hypothesis applied to additive inputs b_r and b_d. To model this we apply $I_r(t)$ and $I_d(t)$ as input to dynamic changes in $|\, b_r \,|$ and $|\, b_d \,|$, analogous to (and instead of) **[6]** and **[7]** with $\alpha_r = \alpha_d = \gamma_r = \gamma_d = 0$.

The simulations in the rest of this paper assume hyp. A where we let the time scale associated with the evolution of elite ideological position be $\tau_x = 4$ y. We also explore other times scales and their implications in *SI Appendix* (*SI Appendix*, Fig. S3). Simulations for hyp. B and hyp. C are presented in *SI Appendix*, along with details on our rigorous and robust analysis and comparison of how well each hypothesis captures the trends in the historical data on polarization and its asymmetry.

Results

Nonlinear Elite Response to Policy Mood Underlies Asymmetric Elite Dynamics. The policy mood inputs I_r and I_d to the Republican and Democratic elites, respectively, are plotted as a function of time in Fig. 5*B* in the case that all sensitivity thresholds are the same: $U_r = L_r = U_d = L_d = 0.45$. The peaks in Fig. 5*B* reveal that swings in policy mood become smaller over time, at least until the one occurring in response to the Trump presidency. This could suggest that policy mood is becoming more stable, but it also may reflect that the issue preferences of Americans have been polarizing to a certain degree and becoming more rigid.

As can be observed in the plots of I_r and I_d, the nonlinearity in f highlights the alignment of policy mood swings with key historical events. It also highlights that Republican policy mood swings are larger in magnitude and more prolonged than Democratic policy mood swings. This is the key difference that drives asymmetric elite polarization in our model.

We consider sensitivity thresholds that are symmetric between Republican and Democratic elites. In *SI Appendix*, Fig. S4, we explore the possibility of asymmetric thresholds—specifically that Republican elites might be less responsive to leftward swings than Democratic elites to rightward swings. The simulation results suggest Republicans would have crossed the critical threshold for "runaway" polarization much earlier (closer to 1980 than 1990) and that the asymmetric nature of elite polarization overall would have been much worse.

Policy Mood Input to Elite Self-Reinforcement Drives Asymmetric Polarization. Each Republican/Democratic policy mood swing, as reflected in the input $I_r(t)$ /$I_d(t)$, leads to a sharp increase in Republican/Democratic ideological self-reinforcement and a modest decrease in Democratic/Republican ideological self-reinforcement. The asymmetry in conservative versus liberal policy mood swings translates into asymmetric temporal evolution of the modeled ideological self-reinforcement levels $\alpha_r(t)$ and $\alpha_d(t)$ (Fig. 5*C*).

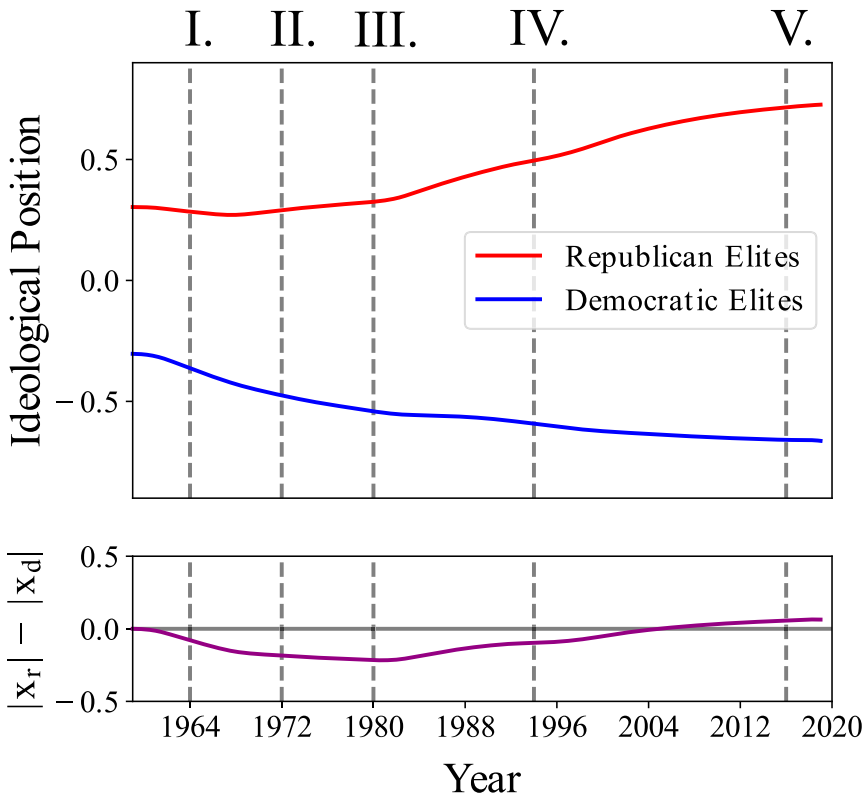


Fig. 6. Simulation results of model dynamics (**[1]** and **[2]**) with self-reinforcement levels α_r and α_d changing as shown in Fig. 5*C* in response to policy mood input I_r and I_d shown in Fig. 5*B* according to dynamics (**[6]** and **[7]**) with nonlinear function f defined by **[5]** with $U_r = L_r = U_d = L_d = 0.45$. Initial conditions and model parameters: $x_r(t_0) = 0.3$, $x_d(t_0) = -0.3$, $b_r = 0.1$, $b_d = -0.1$, $\tau_x = 4$, and $t_0 = 1959$.

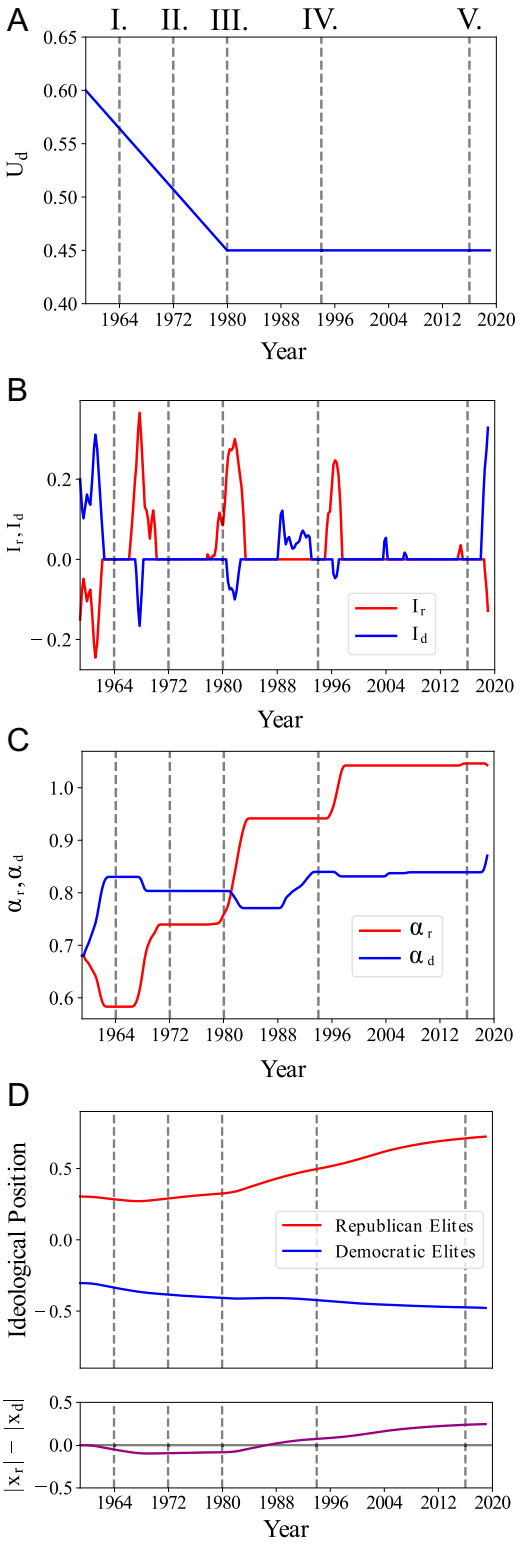


Fig. 7. Increasing Democratic responsiveness to liberal policy mood swings. Democratic elites gradually become more responsive over time to policy mood swings throughout the 1960s and 1970s, as modeled by the time-varying U_d plotted in *A*. The other sensitivity thresholds are $U_r = L_r = L_d = 0.45$, as in Fig. 6. All other parameters and initial conditions are the same as in Fig. 6. (*B* and *C*) The resulting policy mood inputs I_r and I_d (*B*) and response of α_r and α_d (*C*) to these policy mood inputs. (*D*) The qualitative trends in ideological positions and the asymmetry in the polarization can be compared to the DW-NOMINATE scores in Fig. 1; see Fig. 8.

Both $\alpha_r(t)$ and $\alpha_d(t)$ are overall increasing, which leads to increasing polarization as illustrated in Fig. 6. This is consistent with the mechanistic explanation illustrated in Fig. 4*A*. The Republican self-reinforcement level α_r remains well below the threshold for runaway polarization (equal to 1 in our model) until around 1980, corresponding to the beginning of the Reagan period. Around that date, α_r undergoes a dramatic increase, which brings the Republican elite close enough to the runaway polarization threshold to see escalating polarization. In line with empirical data (Fig. 1), a second Republican polarization bump is observed during the Newt Gingrich era. This second bump pushes α_r above the threshold for runaway polarization.

According to the model results of Fig. 6, Democratic elites had two periods of leftward movement: in the early 1960s and again in the early 1990s. The latter accords with empirical data but the former does not. This suggests Democratic elites may have missed an opportunity to capitalize fully on a leftward swing in policy mood. It could also reflect the ideological heterogeneity of the Democratic coalition. A Democratic party that includes both liberals and conservatives would be less likely to fully embrace a liberal policy mood signal than a party that contains only liberals. After the Civil Right Act in 1964, conservatives steadily sorted out of the Democratic party, likely increasing its responsiveness to public opinion.

We can adjust our model to allow for Democratic responsiveness to large liberal policy mood swings to start relatively small and grow over time by letting the sensitivity threshold U_d decrease over time in a linear fashion across the 1960s and 1970s, as shown in Fig. 7*A*. Fig. 7 shows that this modification reduces Democratic polarization in the early 1960s and lets our model better approximate the empirical data (Fig. 8). The smaller leftward movement of the Democrats in the early 1960s distinguishes Fig. 7 *B* and *C* from Fig. 5 *B* and *C*, respectively. In line with empirical data (Fig. 1), a marked Republican polarization begins substantially earlier than Democratic polarization. The Democratic self-reinforcement level α_d remains below the threshold for runaway polarization in Figs. 5*C* and 7*C*, although there is a steep increase recently during the Trump presidency.

We emphasize that no data fitting was performed in obtaining Figs. 5–8. The largely agnostic linear filtering and threshold nonlinearity applied to policy mood data are responsible for the sharply asymmetric course of voter influence over elites. These results are independent of modeling details, for the most part, and robust to parameter variations. In other words, the results are intrinsic to the policy mood data. The results are also agnostic to whether or not the voters are polarized. The resulting asymmetric increase in polarizing positive feedback is both in line with existing self-reinforcement theory of polarization and unique, in that it connects those theories to an asymmetric influence of policy mood over elites.

We evaluate how well the results of Fig. 7 match the DW-NOMINATE scores by plotting in Fig. 8 both the normalized simulated trajectories, $\bar{x}_r$ and $\bar{x}_d$, and the normalized DW-NOMINATE scores (see *SI Appendix*, section S2.F for details on the normalization). The difference curves (purple) in Fig. 8, *Bottom* measure the asymmetry in the polarization: The simulation (solid curve) slightly overpredicts the asymmetry in the DW-NOMINATE scores (dashed curve) after 1996. This suggests that the asymmetry in elite polarization could have been worse, given how much swings in policy mood have favored conservatives over the last half century.

There are four key parameters in the model that can make a significant difference in how well the results match the DW-NOMINATE scores and their asymmetry. In *SI Appendix*, section S2.F, we present a robust analysis of hyp. A where we

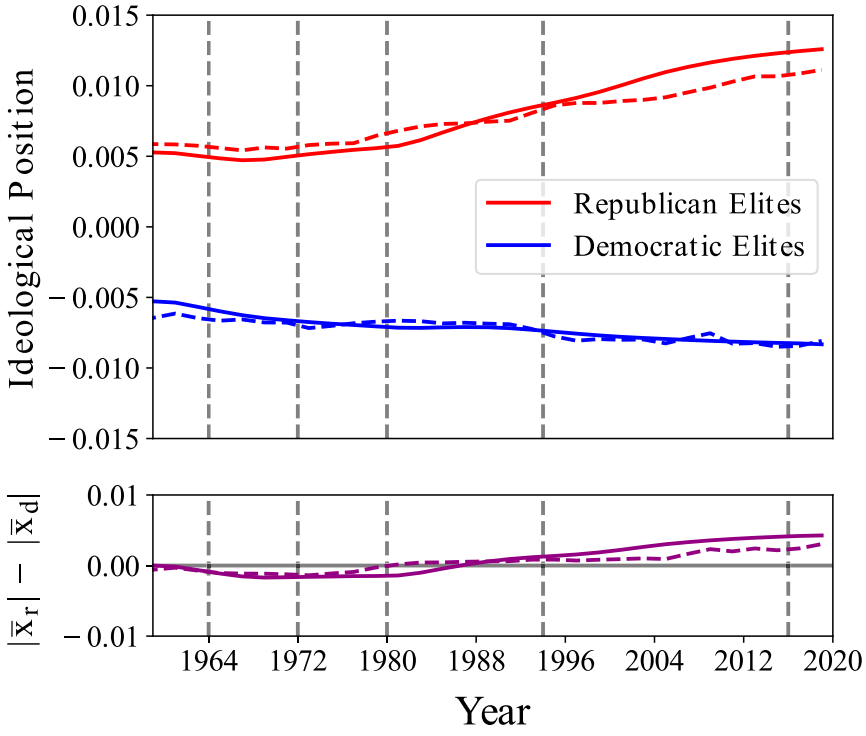


Fig. 8. Comparison of simulation results (solid lines) of Fig. 7 with DW-NOMINATE scores (dashed lines) of Fig. 1, after normalization. *Bottom* panel compares the asymmetry in polarization in the simulation (solid purple line) with that in the DW-NOMINATE scores (dashed purple line).

computed the equivalent of Fig. 8 from 1979 to 2019 over a range of values for each of these four key parameters: U, L, p_0, and k, where $U_r = U_d = U$, $L_r = L_d = L$, $\alpha_r(1979) = \alpha_d(1979) = p_0$, and $k_r = k_d = k$. Each of the 4,000 simulations corresponds to a different combination of parameter values, and for each one we computed the mean-square error (MSE) of normalized trajectories compared to normalized DW-NOMINATE scores. Over the 4,000 simulations, the lowest MSE is 4.37×10^{-7}, achieved with $U = 0.57$ and $L = 0.37$, yielding a difference curve that closely captures the polarization asymmetry in the data. We evaluate how well asymmetry is captured with two measures (*SI Appendix*, Table S2). The first one is the mean-square error in the difference curves (MSEdif). The second one is the "polarization asymmetry index (PAI)," which is the ratio of the simulated difference to the DW-NOMINATE score difference in 2019 (after normalization). Here, MSEdif $= 5.00 \times 10^{-7}$ and PAI $= 0.80$. The PAI indicates that the results slightly underpredict the asymmetry. This is consistent with the relatively large U having more of a moderating effect on x_r than on x_d since the biggest swings during the period analyzed were to the right.

Alternative and Null Hypotheses: Reflexive Partisanship, the Breakdown of Norms, and Additive Response. We consider and reject our alternative and null hypotheses on elite response mechanisms. Considering hyp. B, we let the asymmetric policy mood drive the elite reflexive partisanship levels γ_r,γ_d. However, as predicted by Fig. 4*B*, applying the policy mood inputs $I_r(t)$ and $I_d(t)$ to the reflexive partisanship levels γ_r and γ_d does not lead to asymmetric polarization. This is the result of the mutual response associated with reflexive partisanship, which is independent of the ratio of γ_r to γ_d. This is illustrated in *SI Appendix*, Fig. S5*A*, which shows virtually no asymmetry in the results analogous to Fig. 6, i.e., with policy mood input to γ_r and γ_d given as in **[6]** and **[7]** with $I_r(t)$ and $I_d(t)$ as given by Fig. 5*B* and $\alpha_r = \alpha_d = 0$. Similarly, *SI Appendix*, Fig. S5*B* shows virtually no asymmetry in the results analogous to Fig. 7, i.e., with policy mood input to γ_r and γ_d given as in **[6]** and **[7]** with $I_r(t)$ and $I_d(t)$ as given by Fig. 7*B* and $\alpha_r = \alpha_d = 0$. We plot the equivalent of Fig. 8 in *SI Appendix*, Fig. S6, where we see gross underprediction of asymmetry in polarization.

To make rigorous our rejection of reflexive partisanship as the dominant elite response mechanism, we performed the analogous robust analysis of hyp. B as for hyp. A, running 4,000 simulations over the same combinations of parameters. As reported in *SI Appendix*, Table S2, the lowest MSE is 5.27×10^{-7}, achieved with $U = 0.57$ and $L = 0.50$, which underperforms compared to hyp. A. Further, as the plot in *SI Appendix*, Table S2 shows, even with a best choice of parameters, the simulation under hyp. B can still not capture the asymmetry in polarization in the data (MSEdif $= 6.23 \times 10^{-7}$ and PAI $= 0.57$).

Considering hyp. C, the null hypothesis, we let the asymmetric policy mood drive the elite additive input levels b_r, b_d. However, as predicted by the theory and illustrated in Fig. 4*C*, the simulated trajectories simply track the policy mood. To make rigorous our rejection of additive response as a mechanism that can explain the data, we performed the analogous robust analysis of hyp. C as for hyp. A and hyp. B, running 4,000 simulations over combinations of the same four parameters. As reported in *SI Appendix*, Table S2, the lowest MSE is 9.91×10^{-7}, achieved with $U = 0.30$ and $L = 0.17$, which reflects relatively poor performance overall. Importantly, as seen in the plot in *SI Appendix*, Table S2, even with best parameters, the simulation under hyp. C is not at all representative of the asymmetry in polarization in the data (MSEdif $= 17.39 \times 10^{-7}$ and PAI $= 0.45$).

We additionally hypothesized that perhaps it was not increasing reflexive partisanship but the breakdown of bipartisan norms, which began in the 1970s, that accounts for the rise of asymmetric polarization (31). To test this, we used the same conditions as in Fig. 6 but applied a small negative γ, where $\gamma_r = \gamma_d = -0.1$. As *SI Appendix*, Fig. S7 suggests, even if norms of bipartisanship had endured, it would not have prevented the rise of asymmetric polarization.

We have used estimates of citizen preferences from data (policy mood) as input to our model of the evolution over time of the ideological position of elites. We have compared model output to estimates of ideology from data (DW-NOMINATE scores). Parameters U_r, U_d, L_r, and L_d and variables α_r and α_d are not so easily estimated from data. However, we have shown that for reasonable choices of parameters the model robustly provides output that is consistent with the historical data and, with it, predictions on trends in variables. For a future work, we propose using extended Kalman filtering for our nonlinear model, much as linear Kalman filtering was used in ref. 20. As suggested in ref. 20, extracting the model parameters that best explain historical data constitutes a regression on aggregated macroscopic variables and parameters, such as party self-reinforcement or sensitivity to policy mood swings.

Conclusion

Social scientists have offered a wide range of explanations for the rise of elite polarization in the United States. By focusing on feedback mechanisms, our model integrates the various explanations for polarization within a single framework. Interest group pressure, increases in party discipline, ideological sorting, changes to the media environment, and other suggested hypotheses are different manifestations of reinforcement that drive polarization upward and that feed off one another.

This article underscores why it is so important to understand how political processes reinforce themselves and each other through positive feedback and, therefore, connects polarization to a broader literature on path dependency and self-reinforcement (4, 5, 32). Social processes such as polarization are dynamic—just as so many processes in nature are—and our models must reflect that. In fact, by ruling out hyp. C, we show that explanations that do not account for positive feedback cannot account for historical patterns of polarization. Viewing polarization through this lens reveals critical thresholds or moments when processes become difficult if not impossible to reverse. Our model suggests that this threshold has been crossed by Republicans in Congress and may very soon be breached by Democrats.

Our results also demonstrate that a critical threshold divides a state envisioned by the Responsible Party Model wherein the political parties are distinct, effective, and accountable to the voters and a state that begets unchecked polarization driven by self-reinforcement. Political scientists are, perhaps, more aware than most that democracies die and that polarization can be a leading cause in their demise (1). The authors of the 1950 American Political Science Association report calling for the parties to differentiate themselves could not have imagined where those parties would end up 70 y later. They did emphasize that the parties should not only offer distinct policy platforms but also be "effective" or "able to cope with the great problems of modern government" (ref. 33, p. 17). There may be an optimal level of differentiation beyond which effectiveness is harmed (34).

Our model finds that public opinion has an important role to play in the asymmetric nature of elite polarization. However, rather than mass polarization driving elite polarization (or vice versa), we argue that shifts in aggregate public opinion, i.e., not necessarily mass polarization, can drive asymmetric elite polarization. This finding is all the more remarkable because researchers have not, as far as we know, noted that swings in policy mood are asymmetric, either in magnitude or in duration. These differences, compounded over time, can account for the asymmetry we observe in elite polarization. That said, while we treat policy mood as an input in our analysis, future research could extend our model and use instead DW-NOMINATE scores as inputs to then model policy mood.

While our model suggests that policy mood plays an important role in the asymmetric nature of elite polarization, it also suggests that reflexive partisanship does not. Although scholars have suggested antipathy toward the opposing party might be stronger among Republicans (35), reflexive partisanship cannot account for the magnitude of the asymmetry observed in the system even if this is the case. This is because the polarizing effects of such antipathy outweigh its asymmetric effects.

Finally, perhaps one of the most sobering results of our analysis is that we find that as self-reinforcement increases, parties become increasingly less responsive to policy mood. Once self-reinforcement, driven by policy mood, crosses a critical threshold, polarization dynamics become completely dominated by positive feedback. Hence, while public opinion can drive initial polarization, it may be relatively powerless to stop it. Future research might test the limits of this finding, however. For instance, it might consider how an abrupt and sustained leftward swing in public opinion could affect current levels of polarization.

Additionally, while our model primarily looks backward, and explains the polarization dynamics of the last 70 y, future research can use the model as a testbed to evaluate mechanisms for decreasing polarization. For instance, assuming that policy mood continues to have a conservative bias, would decreasing Republican sensitivity to it overcome the positive feedback effects identified here? Turning to the agent-based version of the model, how many and which legislators need to exogenously decrease their sensitivity to positive feedback to decrease polarization? In other words, can one senator, perhaps centrally located, unilaterally decrease polarization? Agent-based models could also be used to examine the role of party heterogeneity on self-reinforcement and, ultimately, polarization. We hope this model will provide researchers with a simple but powerful tool for exploring ways to mitigate, if not reverse, the polarization dynamics that are threatening the long-term stability of this country.

Data Availability. All study data are included in this article and/or *SI Appendix*.

ACKNOWLEDGMENTS. We are grateful to the anonymous reviewers for many helpful comments. This research was supported by Dirección General de Asuntos del Personal Académico (DGAPA), National Autonomous University of Mexico (UNAM), through the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) research grant IN102420, and by Consejo Nacional de Ciencia y Tecnología (Conacyt) through the research grant CB-A1-S-10610 (to A.F.) and by the NSF Graduate Research Fellowship Program under Grant DGE-2039656 (to A.B.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

1. S. Levitsky, D. Ziblatt, *How Democracies Die* (Broadway Books, 2018).
2. F. E. Lee, *Beyond Ideology: Politics, Principles, and Partisanship in the US Senate* (University of Chicago Press, 2009).
3. M. Fiorina, S. Abrams, *Disconnect: The Breakdown of Representation in American Politics* (University of Oklahoma Press, 2012).
4. P. Pierson, Increasing returns, path dependence, and the study of politics. *Am. Polit. Sci. Rev.* **94**, 251–267 (2000).
5. P. Pierson, E. Schickler, Madison's constitution under stress: A developmental analysis of political polarization. *Annu. Rev. Polit. Sci.* **23**, 37–58 (2020).
6. N. McCarty, *Polarization: What Everyone Needs to Know* (Oxford University Press, 2019).
7. M. Grossmann, D. A. Hopkins, *Asymmetric Politics: Ideological Republicans and Group Interest Democrats* (Oxford University Press, 2016).
8. T. Mann, N. Ornstein, *It's Even Worse Than It Looks: How the American Constitutional System Collided With the New Politics of Extremism* (Basic Books, 2016).
9. F. E. Lee, Patronage, logrolls, and "polarization": Congressional parties of the gilded age, 1876-1896. *Stud. Am. Polit. Dev.* **30**, 116 (2016).
10. N. McCarty, In defense of DW-NOMINATE. *Stud. Am. Polit. Dev.* **30**, 172 (2016).
11. S. Gailmard, J. A. Jenkins, Distributive politics and congressional voting: Public lands reform in the Jacksonian era. *Public Choice* **175**, 259–275 (2018).

12. D. DiSalvo, *Engines of Change: Party Factions in American Politics, 1868-2010* (Oxford University Press, 2012).
13. H. Noel, *Political Ideologies and Political Parties in America* (Cambridge University Press, 2014).
14. T. Skocpol, A. Hertel-Fernandez, The Koch network and Republican Party extremism. *Perspect. Polit.* **14**, 681–699 (2016).
15. A. Abramowitz, *The Disappearing Center: Engaged Citizens, Polarization, and American Democracy* (Yale University Press, 2010).
16. M. Fiorina, S. Abrams, J. Pope, *Culture War: The Myth of a Polarized America* (Longman, 2005).
17. Y. Lelkes, P. M. Sniderman, The ideological asymmetry of the American party system. *Br. J. Polit. Sci.* **46**, 825–844 (2016).
18. S. W. Webster, *American Rage: How Anger Shapes our Politics* (Cambridge University Press, 2020).
19. C. Wlezien, The public as thermostat: Dynamics of preferences for spending. *Am. J. Pol. Sci.* **39**, 981–1000 (1995).
20. R. Erikson, M. MacKuen, J. Stimson, *The Macro Polity* (Cambridge University Press, 2002).
21. J. Stimson, *Public Opinion in America: Moods, Cycles, and Swings* (Routledge, 2018).
22. A. Bizyaeva, A. Franci, N. E. Leonard, Nonlinear opinion dynamics with tunable sensitivity. *arXiv* [Preprint] (2020). https://arxiv.org/abs/2009.04332. (Accessed 6 September 2021).
23. A. Franci, A. Bizyaeva, S. Park, N. E. Leonard, Analysis and control of agreement and disagreement opinion cascades. *Swarm Intell.* **15**, 47–82 (2021).
24. F. P. Santos, Y. Lelkes, S. A. Levin, Link recommendation algorithms and dynamics of polarization in online social networks. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e102141118 (2021).
25. S. Callander, J. C. Carbajal, Cause and effect in political polarization: A dynamic analysis. *J. Polit. Econ.*, in press.
26. J. Stimson, Data from "Public Policy Mood" (2021). https://stimson.web.unc.edu/data/. (Accessed 6 September 2021).
27. S. Soroka, C. Wlezien, *Degrees of Democracy: Politics, Public Opinion, and Policy* (Cambridge University Press, 2010).
28. K. Bawn *et al*., A theory of political parties: Groups, policy demands and nominations in American politics. *Perspect. Polit.* **10**, 571–597 (2012).
29. B. Highton, Issue accountability in US House elections. *Polit. Behav.* **41**, 349–367 (2019).
30. D. Caughey, C. Warshaw, Policy preferences and policy change: Dynamic responsiveness in the American states, 1936–2014. *Am. Polit. Sci. Rev.* **112**, 249–266 (2018).
31. T. Mann, N. Ornstein, *The Broken Branch: How Congress is Failing America and How to Get it Back on Track* (Oxford University Press, 2006).
32. S. Page, Path dependence. *Quart. J. Polit. Sci.* **1**, 87–115 (2006).
33. Toward a more responsible two party system: A report of the Committee on Political Parties. *Am. Polit. Sci. Rev.* **44**, 1–96 (1950).
34. S. A. Binder, *Stalemate: Causes and Consequences of Legislative Gridlock* (Brookings Institution Press, 2004).
35. L. Mason, *Uncivil Agreement: How Politics Became our Identity* (University of Chicago Press, 2018).

PNAS
www.pnas.org

Supplementary Information for

The nonlinear feedback dynamics of asymmetric political polarization

Naomi Ehrich Leonard[a], **Keena Lipsitz**[b], **Anastasia Bizyaeva**[a], **Alessio Franci**[c], **Yphtach Lelkes**[d]

[a] Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544

[b] Department of Political Science, Queens College and The Graduate Center, City University of New York, Flushing, NY 11367

[c] Department of Mathematics, National Autonomous University of Mexico, 04510 Mexico City, Mexico

[d] Annenberg School for Communication and Department of Political Science, University of Pennsylvania, Philadelphia, PA 19104

To whom correspondence should be addressed.
E-mail: naomi@princeton.edu and keena.lipsitz@qc.cuny.edu

This PDF file includes:

Supporting Information Text

S1. Model analysis

A. Derivation from General Model. In this section we derive the model for evolution of the ideological positions of two political party elite populations by specializing a general model of opinion formation recently proposed in (1); see also (2). Suppose that each political party elite population forms positions on n mutually exclusive ideological dimensions. We define the real-valued variable z_{ij} to be party i's position on ideological dimension j, where $i = 1, 2$. In this notation $z_{ij} > 0$ (< 0) corresponds to party i favoring (disfavoring) policy positions that align with ideological dimension j. Additionally $z_{ij} = 0$ corresponds to a neutral stance along ideological dimension j. Mutual exclusivity of the ideological dimensions places a constraint on each party's positions:

$$\sum_{j=1}^{n} z_{ij} = 0 \quad \text{for } i = 1, 2. \tag{1}$$

Let $\mathbf{Z}_i = (z_{i1}, \ldots, z_{in})$ be the vector of ideological positions of party i and define $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$. We can then define the drift in party i's ideological position along dimension j as

$$F_{ij}(\mathbf{Z}) = -d_{ij} z_{ij} + u_i \left(\tanh\left(\alpha_i z_{ij} - \gamma_i z_{kj}\right) + \sum_{\substack{l \neq j \\ l=1}}^{n} \tanh\left(-\beta_{il} z_{il} + \delta_{il} z_{kl}\right) \right) + b_{ij}, \qquad k \neq i, \qquad i, k = 1, 2 \tag{2}$$

where tanh is the hyperbolic tangent function. The drift Eq. (2) contains a number of real-valued parameters, which we interpret in the following way:

1. $d_{ij} > 0$ is the resistance of party i to forming a non-neutral position along the ideological dimension j;
2. $u_i \geq 0$ is the level of attention party i pays to its within-party and cross-party interactions;
3. b_{ij} is the party's intrinsic bias in favor of or against ideological dimension j;
4. $\alpha_i \geq 0$ is the amount of self-reinforcement in party i's ideological positions;
5. γ_i captures the influence of party k on party i along the same ideological dimension;
6. β_{il} and δ_{il} capture the influence of positions along ideological dimension l on party i's position along dimension j.

The evolution over time of the ideological position of party i along ideological dimension j is then summarized by the ordinary differential equation

$$\tau_z \frac{dz_{ij}}{dt} = F_{ij}(\mathbf{Z}) - \frac{1}{n} \sum_{p=1}^{n} F_{ip}(\mathbf{Z}) \tag{3}$$

where subtracting the average drift in ideological position in Eq. (3) models the mutual exclusivity of the ideological dimensions.

For this paper we further specialize this model to two ideological dimensions, conservative and liberal. With this simplification, each party's ideology is captured by a single variable which we define as

$$x_r := z_{11} = -z_{12} \tag{4}$$

for the Republican party elites and accordingly,

$$x_d := z_{21} = -z_{22} \tag{5}$$

for the Democratic party elites. Additionally, we assume $\beta_{il} = \delta_{il} = 0$ and normalize $d_{ij} = 1$, $u_i = 1$ for all $i, j, l = 1, 2$. Finally, we relabel $\tau_z = \tau_x$, $\alpha_1 = \alpha_r$, $\alpha_2 = \alpha_d$, $\gamma_1 = \gamma_r$, $\gamma_2 = \gamma_d$, and define

$$b_r := \frac{1}{2}(b_{11} - b_{12}), \qquad b_d := \frac{1}{2}(b_{21} - b_{22}). \tag{6}$$

With these assumptions imposed, we arrive at the model equations [1]-[2] from the main paper:

$$\tau_x \frac{dx_r}{dt} = \tanh\left(\alpha_r x_r - \gamma_r x_d\right) - x_r + b_r, \tag{7}$$

$$\tau_x \frac{dx_d}{dt} = \tanh\left(\alpha_d x_d - \gamma_d x_r\right) - x_d + b_d. \tag{8}$$

Although we have arrived at this model by treating the two parties as two distinct entities, the general modeling framework proposed in (1) can also be utilized to model evolution of ideological positions of many interacting party members. For such an agent-based model, each node would represent an individual policymaker rather than the party as a whole. Clustering results, e.g. (1, Theorem III.5), suggest that with proper parametrization of the inter-agent interactions, an agent-based model can behave in a manner that is formally equivalent to the two-node model Eq. (7),Eq. (8). This means that the average opinions of

the nodes comprising each of the respective parties would behave the same as x_r, x_d in the two-party model. In the main paper we perform analysis and numerical experiments with the two-party model. At the end of this supplement we use the general agent-based model to represent 50 Republican elites and 50 Democratic elites and illustrate their dynamics in simulation. We show that even with (small) parametric differences among the 50 Republican elites and among 50 Democratic elites, the average behavior of each population agrees with the behavior of the two-population model.

B. Bifurcation analysis of model with constant parameters. In this section we establish using bifurcation analysis the existence of a critical value in one or more of the parameters of the model Eq. (7),Eq. (8). Consider the model with $b_r = b_d = 0$. The Jacobian matrix of this system evaluated at the origin $(x_r, x_d) = (0, 0)$ is

$$J = \frac{1}{\tau_x}\begin{pmatrix} -1+\alpha_r & -\gamma_r \\ -\gamma_d & -1+\alpha_d \end{pmatrix}. \quad [9]$$

The eigenvalues of Eq. (9) are

$$\lambda_{1,2} = -1 + \frac{1}{2}(\alpha_r + \alpha_d) \pm \frac{1}{2}\sqrt{(\alpha_r - \alpha_d)^2 + 4\gamma_r\gamma_d} \quad [10]$$

and one of the two eigenvalues is zero whenever

$$(-1+\alpha_r)(-1+\alpha_d) = \gamma_r\gamma_d. \quad [11]$$

The origin of the nonlinear system Eq. (7), Eq. (8) is stable whenever $\mathrm{Re}(\lambda_{1,2}) < 0$, which is true whenever one of the following conditions is met:

1. $\alpha_r + \alpha_d < 2$ and $\gamma_r\gamma_d < -\frac{1}{4}(\alpha_r - \alpha_d)^2$;
2. $\alpha_r + \alpha_d < 2(1 + \alpha_r\alpha_d - \gamma_r\gamma_d)$ and $\gamma_r\gamma_d \geq -\frac{1}{4}(\alpha_r - \alpha_d)^2$.

Whenever Eq. (11) is satisfied, the Jacobian Eq. (9) is singular. As one or more of the parameters $\alpha_r, \alpha_d, \gamma_r, \gamma_d$ is varied near this singularity, the origin can lose stability and new branches of steady-state solutions can emerge in a nonlinear phenomenon called a steady-state bifurcation. These new solutions will appear along the kernel of J at the singularity. Next we illustrate how this bifurcation analysis specializes to two scenarios examined in the main paper and predicts the emergence of polarized outcomes. Although we only formally handle these two cases, their results (namely, the appearance of a pitchfork bifurcation) apply more generally in Eq. (7),Eq. (8) with heterogeneous parameters.

B.1. Case I: $\gamma_r = \gamma_d = 0$**.** Without party interactions, the evolution in time of each party's ideological position is decoupled from the other and summarized by a one-dimensional equation of the form

$$\tau_x \frac{dx_i}{dt} = -x_i + \tanh(\alpha_i x_i) + b_i. \quad [12]$$

By (1, Proposition IV.1) when $b_i = 0$, Eq. (12) exhibits a supercritical pitchfork bifurcation at $\alpha_i = 1$. For $\alpha_i < 1$ the neutral state ($x_i = 0$) is stable, and for $\alpha_i > 1$ it is unstable. Two non-neutral stable branches of steady-state solutions appear for $\alpha_i > 1$, one corresponding to a right-leaning position, and the other to a left-leaning position - see Figure S1(A). These ideological positions rapidly become polarized as α_i increases in value.

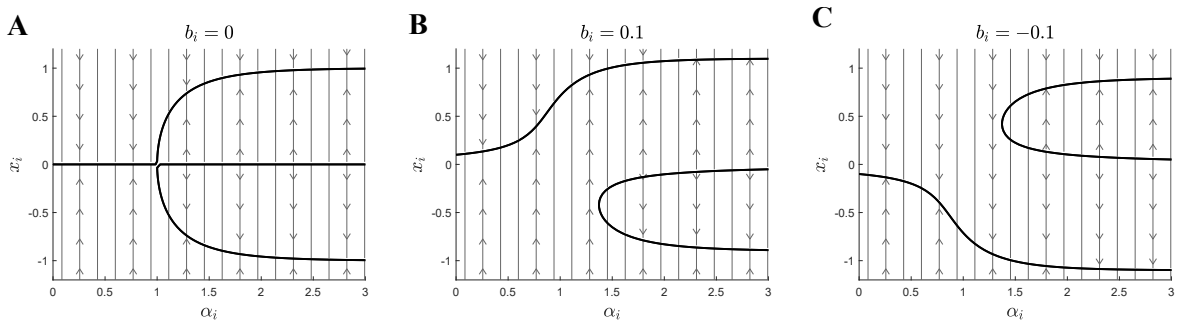


Fig. S1. Bifurcation diagrams for Eq. (12) with (A) no bias, (B) positive bias, and (C) negative bias. Black lines plot steady-state solutions (nullclines) and gray arrows are streamlines showing direction of the flow.

When $b_i \neq 0$, *unfolding theory* (3, Chapter III) predicts that the general shape of the bifurcation diagram resembles the unbiased case pictured in Figure S1(A), but, near the singularity, the equilibrium favored by the additive bias b_i is selected - see Figure S1(B),(C). In the context of political polarization, this means that the ideological position of each party can become strongly polarized in the direction of a small bias, as long as the party's self-reinforcement is sufficiently strong. The degree of polarization increases monotonically with magnitude of α_i, becoming steepest when α_i approaches the critical value of 1. Thus, when α_r and α_d have different values, the party with the greater self-reinforcement is more polarized in its ideological position. This difference in the degree of polarization is particularly strong when one of the α_i coefficients is below its critical value of 1, and the second one is above 1.

B.2. Case II: $\alpha_r = \alpha_d = 0$**.** Let $b_r = b_d = 0$. In this case, the Jacobian in Eq. (9) simplifies to

$$J = \frac{1}{\tau_x}\begin{pmatrix} -1 & -\gamma_r \\ -\gamma_d & -1 \end{pmatrix} \quad [13]$$

and is singular whenever

$$\gamma_r \gamma_d = 1. \quad [14]$$

This corresponds to two potential scenarios:

1. $\gamma_r > 0$ and $\gamma_d > 0$ (reflexive partisanship): At the singularity the kernel of J is

$$\text{span}\{(\sqrt{\gamma_r}, -\sqrt{\gamma_d})\} \quad [15]$$

and therefore new steady-state solution branches appear along a space tangent to

$$x_d = -\sqrt{\frac{\gamma_d}{\gamma_r}} x_r. \quad [16]$$

Qualitatively these solutions correspond to ideological positions of the two parties diverging, with one party taking on a left-leaning stance and the second taking on a right-leaning stance. Additionally when $\gamma_d \neq \gamma_r$, the party with a stronger degree of reflexive partisanship takes on a stronger ideological position. Restricting Eq. (7),Eq. (8) to the kernel of J, the equilibria are fully described by the one-dimensional equation

$$\frac{dx_r}{dt} = -x_r + \tanh(\sqrt{\gamma_r \gamma_d} x_r) \quad [17]$$

coupled with Eq. (16), which is the same equation as Eq. (12) with $\sqrt{\gamma_r \gamma_d}$ acting as a bifurcation parameter. Figure S1(A) illustrates the structure of the equilibria of this equation, if the variable along the horizontal axis is replaced with $\sqrt{\gamma_r \gamma_d}$. The bifurcation point corresponds to $\sqrt{\gamma_r \gamma_d} = 1$, and the two parties' ideological positions become polarized, satisfying Eq. (17) for $\sqrt{\gamma_r \gamma_d} > 1$.

Addition of small nonzero biases b_r, b_d to the model Eq. (7),Eq. (8) will have a two-fold effect: 1) qualitatively changing the structure of the equilibria near the singular point, as pictured in Figure S1(B)),(C), and perturbing the solution vector (x_r, x_d) slightly away from the manifold defined by Eq. (16). When $b_r > 0$ and $b_d < 0$, the branch of equilibria that is selected, for γ_r, γ_d near the singular point, corresponds to $x_r > 0$ and $x_d < 0$. Overall, this analysis means the two parties can develop polarized and asymmetric ideological positions in the direction of their respective small biases. Whether or not the polarization occurs depends on the product of γ_r and γ_d whereas the degree of asymmetry in the ideological positions is determined by their ratio. A much more significant level of difference between γ_r and γ_d is necessary in order to capture a similar level of asymmetry to the $\gamma_r = \gamma_d = 0$ case with α_d slightly below its critical value of 1 and α_r slightly above it.

2. $\gamma_r < 0$ and $\gamma_d < 0$ (bipartisan cooperation): at the singularity the kernel of J is

$$\text{span}\{(\sqrt{\gamma_r}, \sqrt{\gamma_d})\}. \quad [18]$$

Following the same analysis as was carried out in the positive γ_i case, we find that the model undergoes a pitchfork bifurcation at $\sqrt{\gamma_r \gamma_d} = 1$, and whenever $\sqrt{\gamma_r \gamma_d} > 1$, equilibrium solutions appear near the manifold defined by

$$x_d = \sqrt{\frac{\gamma_d}{\gamma_r}} x_r. \quad [19]$$

This analysis predicts that when both parties exhibit bipartisan cooperation, they can overcome the differences in their intrinsic biases b_r, b_d and develop an ideological position that leans in the same direction. In order for this to happen, one (or both) parties must put in sufficient effort to cooperate. This means that it is possible for a party which is putting in a lot of effort to be more cooperative with the other party to entirely switch its ideological leaning.

C. Dynamic parameters. Analysis performed in Section B assumes that parameters of the model Eq. (7),Eq. (8) are static. More generally, equilibria of the static-parameter model inform the behavior of the state trajectories when the parameters become dynamic. For example when there is a slow drift in the bifurcation parameter, geometric singular perturbation theory (4) predicts that trajectories of the system state will remain close to a normally hyperbolic manifold defined by the nullclines of the static-parameter problem. Therefore we can use the analysis in Section B to gain intuition about the dynamic-parameter simulation studies in this paper. In particular we can deduce from this analysis the degree of asymmetry in the trajectories of the ideological positions of the two parties, as well as the parameter values that define the critical point beyond which ideological positions will rapidly polarize.

S2. Numerical experiments

All simulations of the model are run using Python. In each of the simulations in the main text and in this section where policy mood input is used, we first run a short simulation, with all parameters held constant, that starts before the initial time t_0. This allows the dynamics to settle to an equilibrium at the initial time t_0 and prevents initial inadvertent transients. The initial values $x_r(1959) = 0.3$, $x_d(1959) = -0.3$ are chosen to resemble those in the DW-NOMINATE scores.

A. Sensitivity study: asymmetry in γ_r, γ_d growth rate. Figure 4(B) of the main paper shows how increasing reflexive partisanship levels, γ_r and γ_d, yields polarization. However, that polarization exhibits much less asymmetry between the elite ideological positions, when γ_r increases at twice the rate as γ_d, as compared to the asymmetry between elite ideological positions in the case α_r increases at twice the rate as α_d, as shown in Figure 4(A). Here we examine even greater differences between γ_r and γ_d.

As in Figure 4(B), let $x_r(1959) = 0.3$, $x_d(1959) = -0.3$, $\tau_x = 1$ year, $b_r = 0.05$, $b_d = -0.05$, and $\alpha_r = \alpha_d = 0$. Let γ_r and γ_d increase over time (in years) at rate r_r and r_d, respectively:

$$\gamma_r(t) = r_r(t - 1959) + 0.3, \quad \gamma_d(t) = r_d(t - 1959) + 0.3.$$

In Figure 4(B), $r_d = 0.01$ and $r_r = 0.02$. Figure S2(A) shows that there is not much more asymmetry between elite ideological positions, even when $r_d = 0.01$ and $r_r = 0.04$, i.e., when γ_r increase four times as quickly as γ_d. In Figures S2(B) and (C), $r_d = 0$, i.e., γ_d is kept constant at $\gamma_d = 0.3$, and $r_r = 0.04$ and $r_r = 0.06$, respectively. These extreme cases are sufficient to yield asymmetry between elite ideological positions.

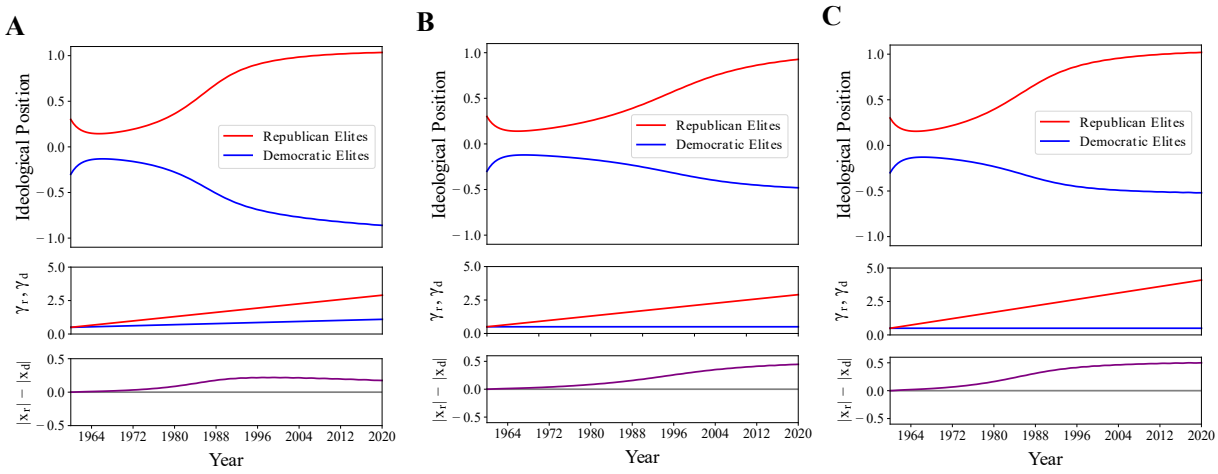


Fig. S2. A. $r_r = 0.04$, $r_d = 0.01$; B. $r_r = 0.04$, $r_d = 0$; C. $r_r = 0.06$, $r_d = 0$.

B. Sensitivity study: time scale τ_x. In all of the time simulations presented in the main text, we let the time scale associated with the evolution of elite ideological position be $\tau_x = 4$ years. In Figure S3, we run the same simulation that produced Figure 6 in the text, but with 1-year, 2-year, 6-year, and 8-year time scales. The runs with 2-year and 6-year time scales perform in a similar fashion to the 4-year scale. We use the 4-year scale, however, because it accords with the length of a presidential term in office and researchers have noted that policy mood inflection points often coincide with party regime change.

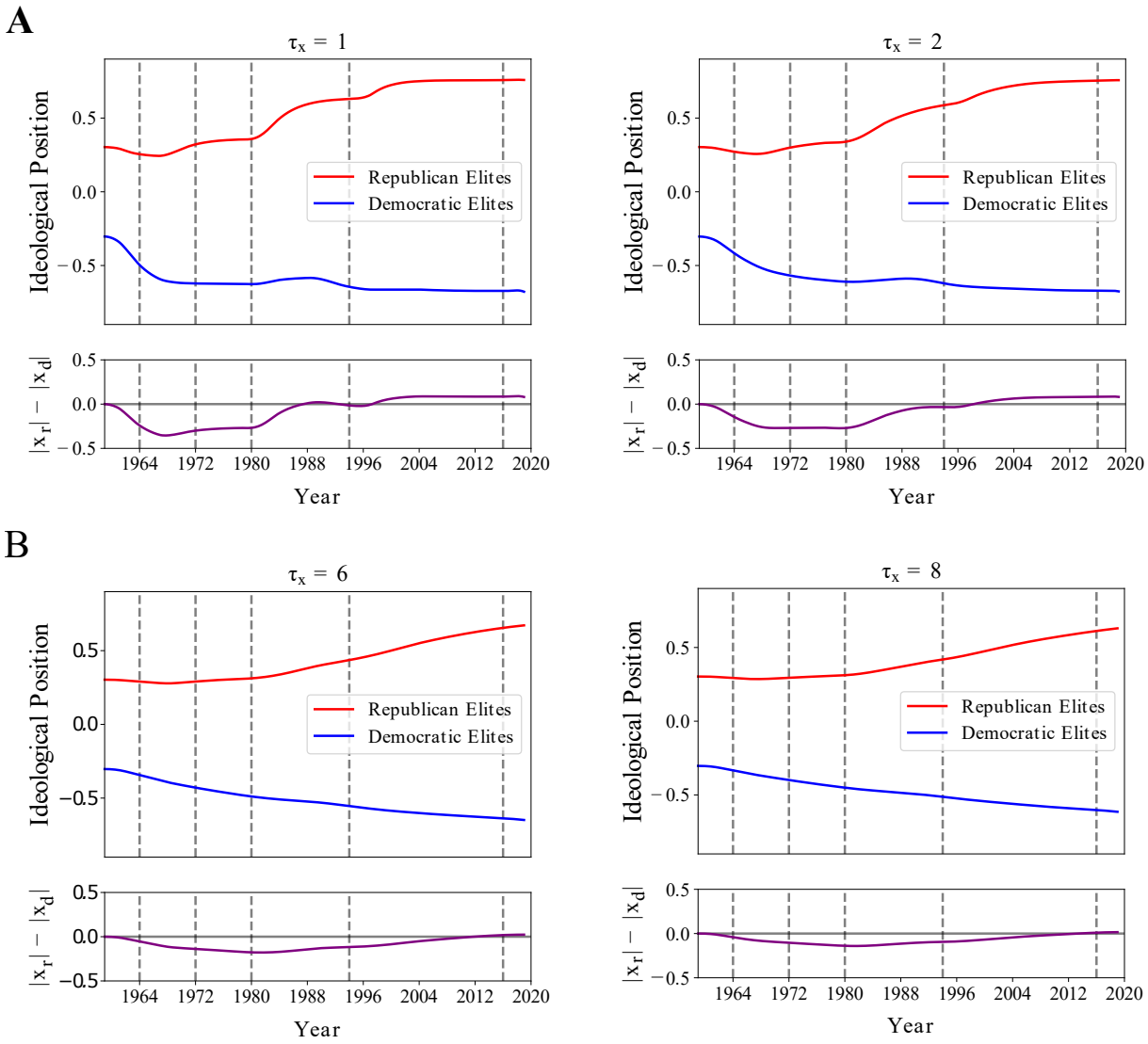


Fig. S3. Same simulation as Figure 6 in main text, with faster (A) and slower (B) time scales.

C. Asymmetric thresholds: Republicans less responsive to public mood swings to the left. In Figure 6 in the main text, we assume Republican and Democratic elites have the same thresholds for responding to swings in PM. Yet, we know elites can have a biased perception of public opinion. While elites of both parties tend to overestimate support for conservative policies, Republicans are particularly prone to making this mistake (5, 6). Research also suggests that the Republican Party is more ideological than the Democratic Party (7) and that Republican members of Congress are more tethered to the national party than their Democratic counterparts (8). This suggests Republicans may be less responsive to leftward swings in policy mood than Democrats to rightward swings. Thus, in Figure S4, we use $U_r = U_d = L_d = 0.45$ as in Figure 6 but set $L_r = 0.55$, increasing the threshold L_r above which the Republican elites will respond to the moderating effect of liberal swings. The simulation results suggest Republicans would have crossed the critical threshold for "run away" polarization much earlier (closer to 1980 than 1990) and, more importantly, that the asymmetric nature of elite polarization overall would have been much worse. Thus, concerns about biased perceptions of public opinion may be over-blown.

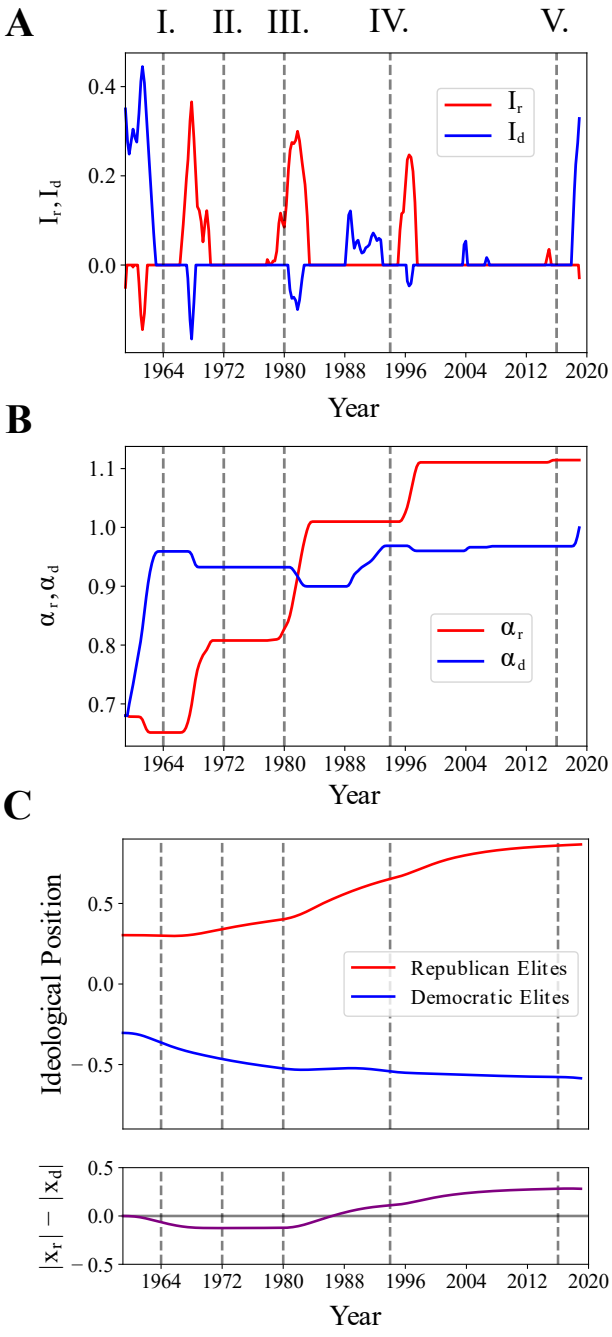


Fig. S4. Asymmetric thresholds in response to policy mood. Model parameters: $U_r = U_d = L_d = 0.45$, $L_r = 0.55$, $I_0 = 0.1$, $\alpha_r(t_0) = \alpha_d(t_0) = 0.68$, $k_\alpha = 0.25$, $x_r(t_0) = 0.3$, $x_d(t_0) = -0.3$, $b_r = 0.1$, $b_d = -0.1$, $\tau_x = 4$, and $t_0 = 1959$.

D. Policy mood drives γ_r, γ_d. Here we test the hypothesis that reflexive polarization levels γ_r and γ_d are driven by policy mood, and not self-reinforcement levels α_r and α_d. All conditions in Figure S5(A) and (B) are the same as in Figures 6 and 7(D), respectively, except with the roles of the γ_r and γ_d swapped with α_r and α_d, respectively. These simulations serve to rule out this hypothesis since there is virtually no asymmetry exhibited in the polarization.

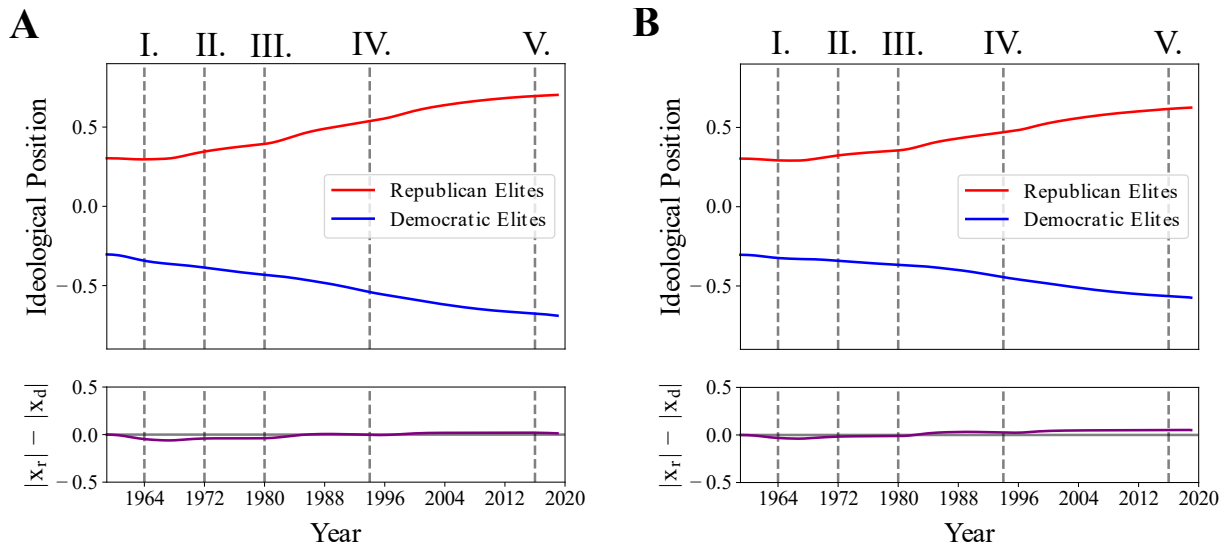


Fig. S5. A. Policy mood drives γ_r, γ_d with symmetric thresholds as in main text Figs. 5 and 6, $\alpha_r = \alpha_d = 0$; B. Policy mood drives γ_r, γ_d with time-varying U_d as in main text Figure 7, $\alpha_r = \alpha_d = 0$.

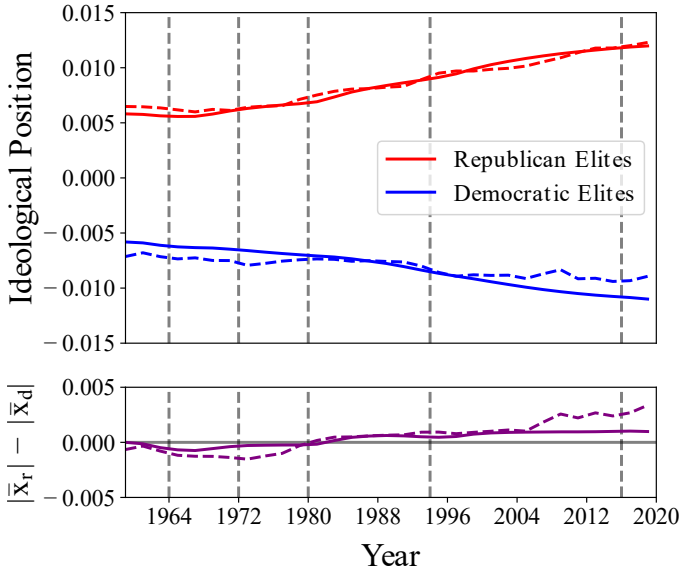


Fig. S6. A. Same as Figure S5(B), with the areas between the curves normalized to 1, and similarly normalized DW-NOMINATE scores (dashed lines) superimposed. The bottom panel shows that the simulated difference in ideological position (solid purple) grossly underpredicts the asymmetry in polarization in the DW-NOMINATE data (dashed purple). See Section S4 for definition of the normalization.

E. Bipartisan cooperation. Here we test whether the break down of bipartisan norms, which began in the 1970s, can account for the rise of asymmetric polarization. We do this by using the same conditions as in Figure 7 but by also applying a small negative γ where $\gamma_r = \gamma_d = -0.1$. As Figure S7 suggests, even if norms of bipartisanship had endured, it would not have prevented the rise of asymmetric polarization.

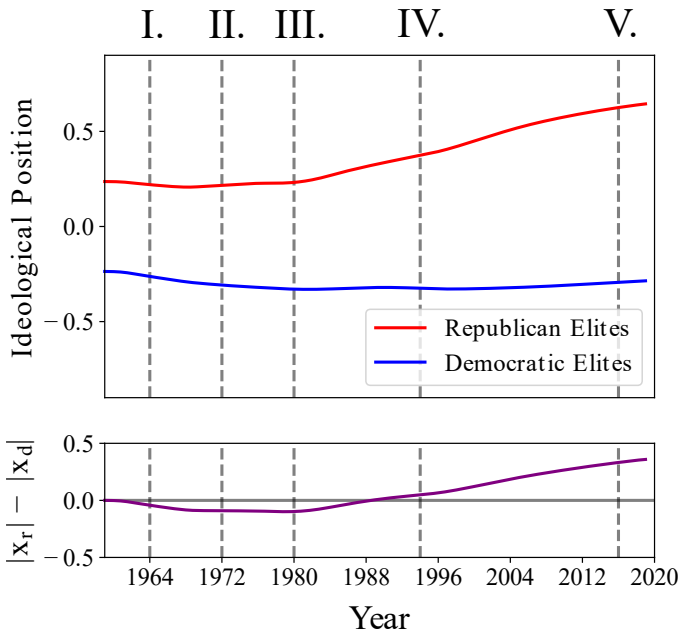


Fig. S7. Same conditions as main text Figure 7 except $\gamma_r = \gamma_d = -0.1$, $x_r(t_0) = 0.24$, $x_d(t_0) = -0.24$.

F. Parameter sweep: finding parameter values that minimize mean square error between model results and data for Hypotheses A, B, and C. To rigorously analyze and compare the simulated trajectories to the DW-NOMINATE score data, we introduce the following normalization:

$$\bar{x}_r = \frac{x_r}{\hat{x}}, \quad \bar{x}_d = \frac{x_d}{\hat{x}}, \tag{20}$$

where the normalization factor $\hat{x}$ is the area between the curves $x_r(t)$ and $x_d(t)$, computed using the trapezoidal rule over the time period of interest, which is from 1979 to 2019 in this section. We then run the model over the range of values defined in Table S1 of four parameters over this time period. We begin the simulations in 1979 because the DW-NOMINATE scores are close in magnitude in 1979, which makes more natural a comparison between the data and the simulated results that start from symmetric initial conditions.

In this section we consider the three separate hypotheses on elite response mechanism:

Hypothesis A. Policy mood drives party self-reinforcement levels α_r and α_d, with $\gamma_r = \gamma_d = 0$ and $b_r = -b_d = 0.1$;

Hypothesis B. Policy mood drives reflexive partisanship levels γ_r and γ_d with $\alpha_r = \alpha_d = 0$ and $b_r = -b_d = 0.1$;

Hypothesis C. Policy mood drives additive inputs b_r and b_d, with $\alpha_r = \alpha_d = \gamma_r = \gamma_d = 0$.

For each of the three hypotheses we simulate the model dynamics over a range of values for each of four parameters: U, L, p_0, and k, where $U_r = U_d = U$, $L_r = L_d = L$, $k_r = k_d = k$, $\alpha_r(1979) = \alpha_d(1979) = p_0$ for Hyp. A, $\gamma_r(1979) = \gamma_d(1979) = p_0$ for Hyp. B , and $b_r(1979) = b_d(1979) = p_0$ for Hyp. C. The range of values simulated are described in Table S1 and the results of the simulation are presented in Table S2.

The first measure of comparison we consider is the mean square error (MSE) between the normalized simulated trajectories ($\bar{x}_r$ and $\bar{x}_d$) and the normalized DW-NOMINATE scores ($\bar{x}_r^{DW}$ and $\bar{x}_d^{DW}$), defined as

$$\text{MSE} = \frac{1}{N}\sum_{n=1}^{N}\left((\bar{x}_r(t_n) - \bar{x}_r^{DW}(t_n))^2 + (\bar{x}_d(t_n) - \bar{x}_d^{DW}(t_n))^2\right), \tag{21}$$

where N is the number of time instances at which the comparison is made and t_n is the set of indexed time instances, where $n = 1, \ldots, N$. The MSE measures how well the normalized trajectories of the simulation agree with the normalized trajectories of the data, such that the smaller the MSE the better the agreement. In Table S2 we present the results for each hypothesis corresponding to the simulation yielding the lowest MSE over all combinations of parameters as listed in Table S1. The results include the corresponding parameter values and two measures of the asymmetry in the simulated polarization. The first is

	Hyp. A (α)	Hyp. B (γ)	Hyp. C (b)
U_{min}	0.1	0.1	0.1
U_{max}	0.7	0.7	0.7
No. values U	10	10	10
L_{min}	0.1	0.1	0.1
L_{max}	0.7	0.7	0.7
No. values L	10	10	10
k_{min}	0.1	0.1	0.1
k_{max}	0.5	0.5	0.5
No. values k	8	8	8
$p_{0,min}$	0.6	0.6	0.1
$p_{0,max}$	0.8	0.8	0.8
No. values p_0	5	5	5

Table S1. Minimum and maximum values and number of values used for each of the four parameters: U, L, k, p_0 in the set of analyses performed for Hypotheses A, B, and C. A total of 4000 simulations were run for each hypothesis, where each simulation used a different combination of parameter values. We present in Table S2 the results of the simulation with the best results, defined as the lowest mean square error (MSE) of modeled ideological positions with respect to the DW-NOMINATE data.

the polarization asymmetry index (PAI), which we define as the ratio of difference between magnitudes of the normalized trajectories to difference between magnitudes of the normalized DW-NOMINATE scores at the end of the simulation:

$$\text{PAI} = \left(\frac{|\bar{x}_r| - |\bar{x}_d|}{|\bar{x}_r^{DW}| - |\bar{x}_d^{DW}|} \right) (2019). \quad [22]$$

If PAI > 1, the simulation overpredicts the asymmetry polarization in 2019 and if PAI < 1, the simulation underpredicts it. The second measure is the mean square error (MSEdif) between the differences in magnitude of the ideological positions of the parties, defined as

$$\text{MSEdif} = \frac{1}{N} \sum_{n=1}^{N} \left(\left(|\bar{x}_r(t_n)| - |\bar{x}_d(t_n)| \right) - \left(|\bar{x}_r^{DW}(t_n)| - |\bar{x}_d^{DW}(t_n)| \right) \right)^2. \quad [23]$$

The MSEdif measures how well the simulated asymmetry in polarization resembles the asymmetry in polarization in the data over time, such that the lower the MSEdif the better the resemblance.

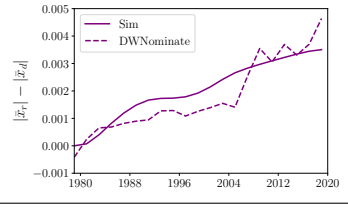
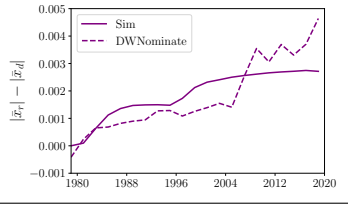
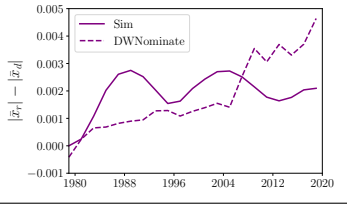
	Hyp. A (α)	Hyp. B (γ)	Hyp. C (b)
MSE ($\times 10^{-7}$)	**4.37**	5.27	9.91
U	0.57	0.57	0.30
L	0.37	0.50	0.17
p_0	0.65	0.60	0.80
k	0.10	0.44	0.10
PAI	**0.80**	0.57	0.45
MSEdif ($\times 10^{-7}$)	**5.00**	6.23	17.39
$\lvert\bar{x}_r\rvert - \lvert\bar{x}_d\rvert$			

Table S2. Parameters and results from the simulation with lowest MSE over the complete set of 4000 simulations, with parameters ranging as described in Table S1, for each of the three hypotheses. $N = 21$ for MSE and MSEdif calculations. Hyp. A performs best over all measures (see bolded values). In particular, we note how well Hyp. A captures the asymmetry in the polarization in the data as illustrated in the plot and in the PAI and MSEdif values. The best simulation for Hyp. B does not do as well with respect to the MSE; however, what is most striking is that even this best run for Hyp. B still underpredicts the asymmetry in polarization in the data. The best simulation for Hyp. C underperforms with respect to MSE and with respect to the asymmetry in polarization. As can be seen in the plot for Hyp. C, $|\bar{x}_r| - |\bar{x}_d|$ tracks the asymmetry in the PM, as predicted by the theory, and does not resemble the asymmetry in the DW-NOMINATE scores. The parameters U, L, and p_0 are quite similar for Hyp. A and Hyp. B. The value $k = 0.1$ for Hyp. A implies relatively slow α_r and α_d dynamics and with it flexibility to match the data. The value $k = 0.44$ for Hyp. B implies relatively fast γ_r and γ_d dynamics and inflexibility to match the data (and notably the asymmetry) since these dynamics saturate early.

S3. Filtering of Policy Mood Data

PM data (Figure S8(A)) was filtered through a first-order high-pass filter with transfer function

$$H_{HP}(s) = \frac{s}{\tau_{HP}s + 1}$$

and a first-order low-pass filter with transfer function

$$H_{LP}(s) = \frac{1}{\tau_{LP}s + 1}.$$

Filters with polynomial transfer functions are realizable as simple differential equations and thus they provide simple interpretable models of generic dynamical responses to inputs.

The high-pass filter models the sensitivity of elite response only to variations of PM, i.e., elites are not sensitive to low-frequency components (the zero-frequency average, in particular) of PM variations. The filter time constant τ_{HP} roughly determines the threshold frequency below which PM variations are filtered-out. In our model, $\tau_{HP} = 1.0$ year.

The low-pass filter models memory of elite response to PM variations. Such a first-order filter "forgets" about the input past history exponentially with time-constant τ_{LP} (expressed in years in our model). Input events more recent than τ_{LP} have relatively large influence on the filter response. Input events more remote than τ_{LP} have relatively small influence on the filter response. In our model, $\tau_{LP} = 10.0$ years.

The filtered PM data is shown in Figure S8(B) and also in Figure 5(A) in the main paper.

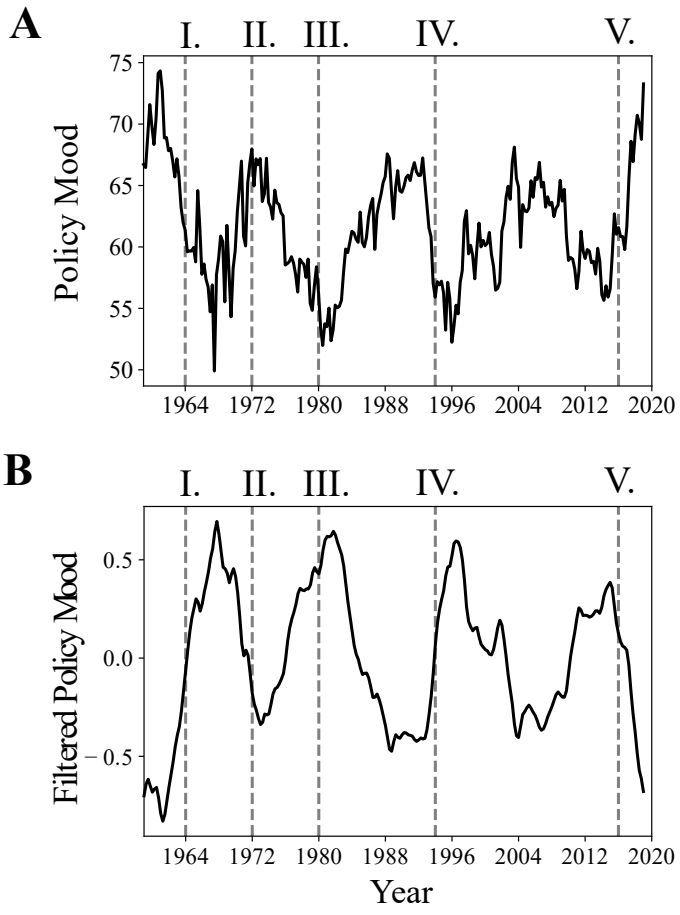


Fig. S8. A. Policy mood data. B. Filtered policy mood data, multiplied by -1 to match the left-right sign convention used throughout this work.

S4. Agent-based model

Consider a group of N agents split into non-overlapping Republican and Democratic elite groups.* We associate to each agent a unique index between 1 and N. Let $\mathcal{R} \subset \{1, \ldots, N\}$ be the set of indices of the agents in the Republican group and $\mathcal{D} \subset \{1, \ldots, N\}$ be the set of indices of the agents in the Democratic group. Then $\mathcal{R} \cap \mathcal{D} = \emptyset$ and $\mathcal{R} \cup \mathcal{D} = \{1, \ldots, N\}$. For simplicity, we assume that these index sets do not change over time.

*Please contact AF (afranci@ciencias.unam.mx) for the Julia code used to run the agent-based simulations.

Let $x_{i_r}(t)$ (resp. $x_{i_d}(t)$) represent the scalar ideological position of agent i_r (resp. i_d) in the Republican (resp. Democratic) elite. The *ideological self-reinforcement level of agent i* is modeled by $\alpha_i(t) \geq 0$. The *level of Republican self-reinforcement of agent i_r with respect to the ideological position of agent j_r* is modeled by $\gamma^{rr}_{i_r j_r}(t) \geq 0$, $i_r, j_r \in \mathcal{R}$, $i_r \neq j_r$. The *level of Democratic self-reinforcement of agent i_d with respect to the ideological position of agent j_d* is modeled by $\gamma^{dd}_{i_d j_d}(t) \geq 0$, $i_d, j_d \in \mathcal{D}$, $i_d \neq j_d$. The *level of Republican reflexive partisanship of agent i_r with respect to the ideological position of agent j_d* is modeled by $\gamma^{rd}_{i_r j_d}(t) \geq 0$, $i_r \in \mathcal{R}$, $j_d \in \mathcal{D}$. The *level of Democratic reflexive partisanship of agent i_d with respect to the ideological position of agent j_r* is modeled by $\gamma^{dr}_{i_d j_r}(t) \geq 0$, $i_d \in \mathcal{D}$, $j_r \in \mathcal{R}$. Each agent i possesses an *ideological bias* $b_i(t)$, which is conservative for Republican agents, i.e., $b_{i_r}(t) \geq 0$ for $i_r \in \mathcal{R}$, and liberal for Democratic agents, i.e., $b_{i_d}(t) \leq 0$ for $i_d \in \mathcal{D}$.

The agent-based model equations are

$$\tau_x \frac{dx_{i_r}}{dt} = S\left(\alpha_{i_r} x_{i_r} + \sum_{\substack{j_r \in \mathcal{R} \\ j_r \neq i_r}} \gamma^{rr}_{i_r j_r} x_{j_r} - \sum_{j_d \in \mathcal{D}} \gamma^{rd}_{i_r j_d} x_{j_d}\right) - x_{i_r} + b_{i_r}, \quad i_r \in \mathcal{R}\,, \tag{24a}$$

$$\tau_x \frac{dx_{i_d}}{dt} = S\left(\alpha_{i_d} x_{i_d} + \sum_{\substack{j_d \in \mathcal{D} \\ j_d \neq i_d}} \gamma^{dd}_{i_d j_r} x_{j_d} - \sum_{j_r \in \mathcal{R}} \gamma^{dr}_{i_d j_r} x_{j_r}\right) - x_{i_d} + b_{i_d}. \quad i_d \in \mathcal{D}\,. \tag{24b}$$

Note that using (1, Theorem III.5) this agent-based model can be shown to be formally equivalent to the two-party model Eq. (7),Eq. (8) where each node represents a within-group average opinion. In all simulations $N = 100$, with $\mathcal{R} = \{1, \ldots, 50\}$ and $\mathcal{D} = \{51, \ldots, 100\}$.

In all simulations we let parameter values for individuals vary from the average of the individual's group by a deviation drawn from a Normal distribution. First, we examine the *symmetric parameter case* in which the average parameter magnitudes for the Republicans are the same as for the Democrats. Second, we examine the *asymmetric parameter case* in which average parameter magnitudes for the Republicans are not the same as for the Democrats. In both cases α_i, ideological self-reinforcement for agent i, varies with respect to the same average across Republicans and Democrats.

A. Agent-based simulations with symmetric parameters between Republicans and Democrats. In the symmetric setting, we let the average magnitude of Republican self-reinforcement level, Republican reflexive partisanship level, and Republican bias, be equal to the average magnitude of Democratic self-reinforcement level, Democratic reflexive partisanship level, and Democratic bias, respectively. More precisely, in the simulations, we let

- α_i, $i = 1, \ldots, N$, are kept constant and drawn from a Normal distribution with mean $\bar{\alpha}$ and variance $\Delta\alpha$;
- for $i_r, j_r \in \mathcal{R}$, $i_r \neq j_r$, $\gamma^{rr}_{i_r,j_r}(t) = \bar{\gamma}_{self}(t) + \Delta\gamma^{rr}_{i_r,j_r}$, where $\Delta\gamma^{rr}_{i_r,j_r}$ is drawn from a Normal distribution with zero mean and variance $\Delta\gamma_{self}$;
- for $i_d, j_d \in \mathcal{D}$, $i_d \neq j_d$, $\gamma^{dd}_{i_d,j_d}(t) = \bar{\gamma}_{self}(t) + \Delta\gamma^{dd}_{i_d,j_d}$, where $\Delta\gamma^{dd}_{i_d,j_d}$ is drawn from a Normal distribution with zero mean and variance $\Delta\gamma_{self}$;
- for $i_r \in \mathcal{R}$, $j_d \in \mathcal{D}$, $\gamma^{rd}_{i_r,j_d}(t) = \bar{\gamma}_{reflex}(t) + \Delta\gamma^{rd}_{i_r,j_d}$, where $\Delta\gamma^{rd}_{i_r,j_d}$ is drawn from a Normal distribution with zero mean and variance $\Delta\gamma_{reflex}$;
- for $i_d \in \mathcal{D}$, $j_r \in \mathcal{R}$, $\gamma^{dr}_{i_d,j_r}(t) = \bar{\gamma}_{reflex}(t) + \Delta\gamma^{dr}_{i_r,j_r}$, where $\Delta\gamma^{dr}_{i_d,j_r}$ is drawn from a Normal distribution with zero mean and variance $\Delta\gamma_{reflex}$;
- b_{i_r}, $i_r \in \mathcal{R}$, are kept constant and drawn from a Normal distribution with mean $\bar{b}$ and variance $\Delta\bar{b}$;
- b_{i_d}, $i_d \in \mathcal{D}$, are kept constant and drawn from a Normal distribution with mean $-\bar{b}$ and variance $\Delta\bar{b}$.

A.1. Symmetric polarization by symmetric increase in ideological self-reinforcement. Figure S9. Simulation parameters:

- $\bar{\alpha} = 0.05$, $\Delta\alpha = 0.0025$;
- $\bar{\gamma}_{self}(t) = 0.1/50 + (1.1 - 0.1)/50 \cdot t/T$, $\Delta\gamma_{self} = 0.05$;
- $\bar{\gamma}_{reflex}(t) = 0.0$, $\Delta\gamma_{reflex} = 0.0$;
- $\bar{b} = 0.8$, $\Delta\bar{b} = 0.8$;

A.2. Symmetric polarization by symmetric increase in reflexive partisanship. Figure S10. Simulation parameters:

- $\bar{\alpha} = 0.05$, $\Delta\alpha = 0.0025$;
- $\bar{\gamma}_{reflex}(t) = -0.1/50 - (1.1 - 0.1)/50 \cdot t/T$, $\Delta\gamma_{reflex} = 0.05$;
- $\bar{\gamma}_{self}(t) = 0.0$, $\Delta\gamma_{self} = 0.0$;
- $\bar{b} = 0.8$, $\Delta\bar{b} = 0.8$;

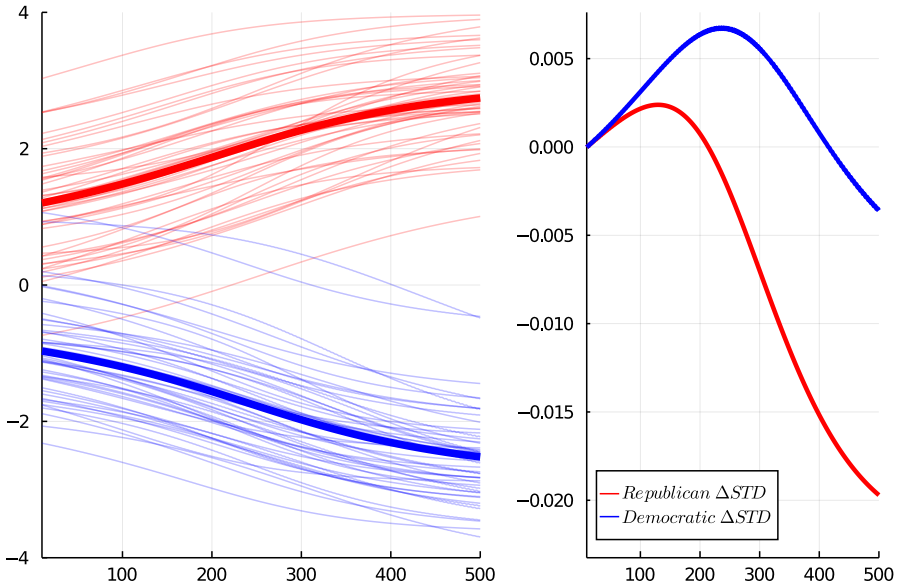


Fig. S9. Left. Thin lines are the evolution of ideological position of each Republican elite (red) and each Democratic elite (blue) as a function of time. Bold lines are the average Republican elite ideological position (red) and average Democratic elite ideological position (blue) as a function of time. Right: Evolution over time of the standard deviation of the ideological positions of Republican elites (red) and Democratic elites (blue) as compared to the standard deviations at the initial time.

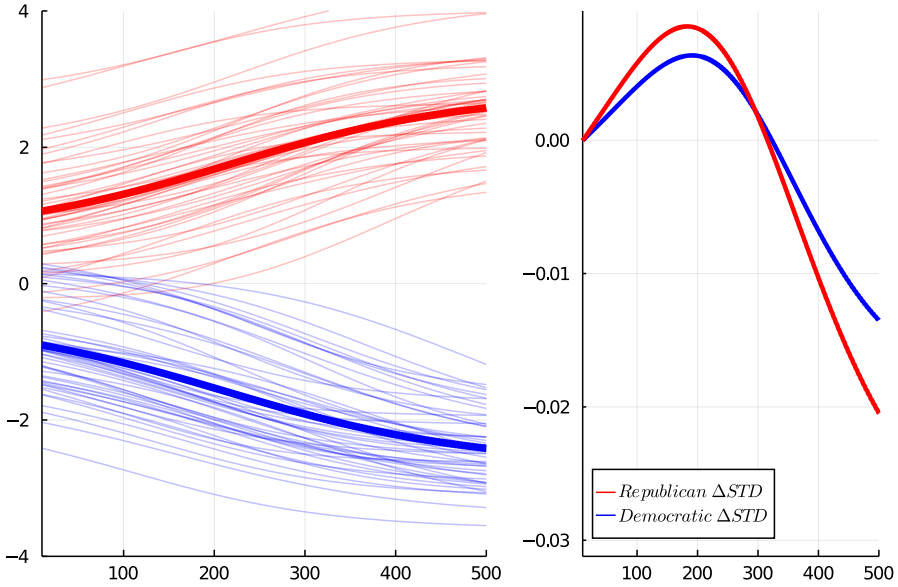


Fig. S10. Left. Thin lines are the evolution of ideological position of each Republican elite (red) and each Democratic elite (blue) as a function of time. Bold lines are the average Republican elite ideological position (red) and average Democratic elite ideological position (blue) as a function of time. Right: Evolution over time of the standard deviation of the ideological positions of Republican elites (red) and Democratic elites (blue) as compared to the standard deviations at the initial time.

B. Agent-based simulations with asymmetric parameters between Republicans and Democrats. In the asymmetric setting, we let the average magnitude of Republican self-reinforcement level and Republican reflexive partisanship level be different from the average magnitude of Democratic self-reinforcement level and Democratic reflexive partisanship level, respectively. We let the average magnitude of the Republican bias and the Democratic bias be the same. More precisely, in the simulations, we let

- α_i, $i = 1, \ldots, N$, are kept constant and drawn from a Normal distribution with mean $\bar{\alpha}$ and variance $\Delta\alpha$;
- for $i_r, j_r \in \mathcal{R}$, $i_r \neq j_r$, $\gamma^{rr}_{i_r,j_r}(t) = \bar{\gamma}^{r}_{self}(t) + \Delta\gamma^{rr}_{i_r,j_r}$, where $\Delta\gamma^{rr}_{i_r,j_r}$ is drawn from a Normal distribution with zero mean and variance $\Delta\gamma^{r}_{self}$;
- for $i_d, j_d \in \mathcal{D}$, $i_d \neq j_d$, $\gamma^{dd}_{i_d,j_d}(t) = \bar{\gamma}^{d}_{self}(t) + \Delta\gamma^{dd}_{i_d,j_d}$, where $\Delta\gamma^{dd}_{i_d,j_d}$ is drawn from a Normal distribution with zero mean and variance $\Delta\gamma^{d}_{self}$;
- for $i_r \in \mathcal{R}$, $j_d \in \mathcal{D}$, $\gamma^{rd}_{i_r,j_d}(t) = \bar{\gamma}^{r}_{reflex}(t) + \Delta\gamma^{rd}_{i_r,j_d}$, where $\Delta\gamma^{rd}_{i_r,j_d}$ is drawn from a Normal distribution with zero mean

and variance $\Delta\gamma^r_{reflex}$;

- for $i_d \in \mathcal{D}$, $j_r \in \mathcal{R}$, $\gamma^{dr}_{i_d,j_r}(t) = \bar{\gamma}^d_{reflex}(t) + \Delta\gamma^{dr}_{i_r,j_r}$, where $\Delta\gamma^{dr}_{i_d,j_r}$ is drawn from a Normal distribution with zero mean and variance $\Delta\gamma^d_{reflex}$;
- b_{i_r}, $i_r \in \mathcal{R}$, are kept constant and drawn from a Normal distribution with mean $\bar{b}$ and variance $\Delta\bar{b}$;
- b_{i_d}, $i_d \in \mathcal{D}$, are kept constant and drawn from a Normal distribution with mean $-\bar{b}$ and variance $\Delta\bar{b}$.

B.1. Asymmetric polarization by asymmetric increase in ideological self-reinforcement. Figure S11. Simulation parameters:

- $\bar{\alpha} = 0.05$, $\Delta\alpha = 0.0025$;
- $\bar{\gamma}^r_{self}(t) = 0.1/50 + (1.1 - 0.1)/50 \cdot t/T$, $\Delta\gamma^r_{self} = 0.05$;
- $\bar{\gamma}^d_{self}(t) = 0.1/50 + (0.7 - 0.1)/50 \cdot t/T$, $\Delta\gamma^d_{self} = 0.05$;
- $\bar{\gamma}^r_{reflex}(t) = 0.0$, $\Delta\gamma^r_{reflex} = 0.0$;
- $\bar{\gamma}^d_{reflex}(t) = 0.0$, $\Delta\gamma^d_{reflex} = 0.0$;
- $\bar{b} = 0.8$, $\Delta\bar{b} = 0.8$;

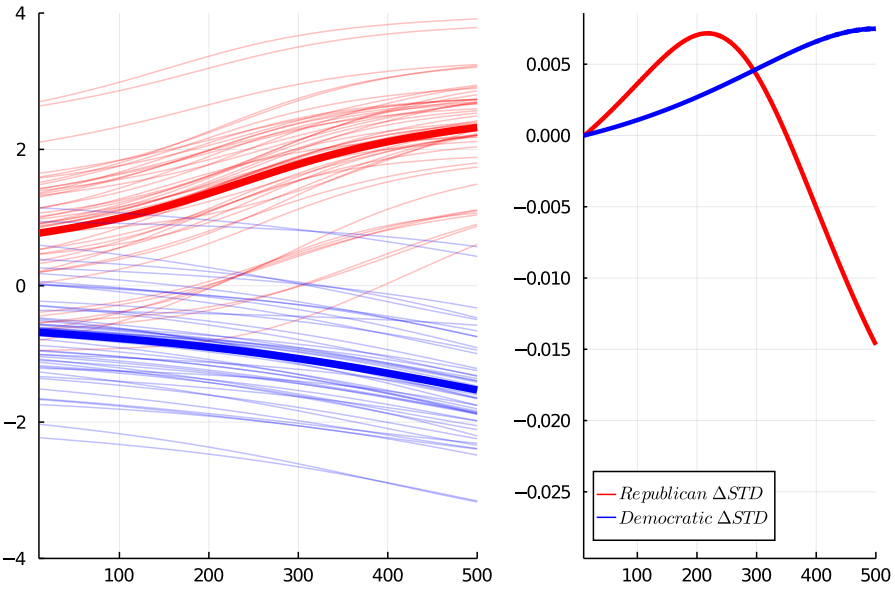


Fig. S11. Left. Thin lines are the evolution of ideological position of each Republican elite (red) and each Democratic elite (blue) as a function of time. Bold lines are the average Republican elite ideological position (red) and average Democratic elite ideological position (blue) as a function of time. Right: Evolution over time of the standard deviation of the ideological positions of Republican elites (red) and Democratic elites (blue) as compared to the standard deviations at the initial time.

B.2. Symmetric polarization by asymmetric increase in reflexive partisanship. Figure S12. Simulation parameters:

- $\bar{\alpha} = 0.05$, $\Delta\alpha = 0.0025$;
- $\bar{\gamma}^r_{reflex}(t) = 0.1/50 + (1.2 - 0.1)/50 \cdot t/T$, $\Delta\gamma^r_{reflex} = 0.05$;
- $\bar{\gamma}^d_{reflex}(t) = 0.1/50 + (0.7 - 0.1)/50 \cdot t/T$, $\Delta\gamma^d_{reflex} = 0.05$;
- $\bar{\gamma}^r_{self}(t) = 0.0$, $\Delta\gamma^r_{self} = 0.0$;
- $\bar{\gamma}^d_{self}(t) = 0.0$, $\Delta\gamma^d_{self} = 0.0$;
- $\bar{b} = 0.8$, $\Delta\bar{b} = 0.8$;

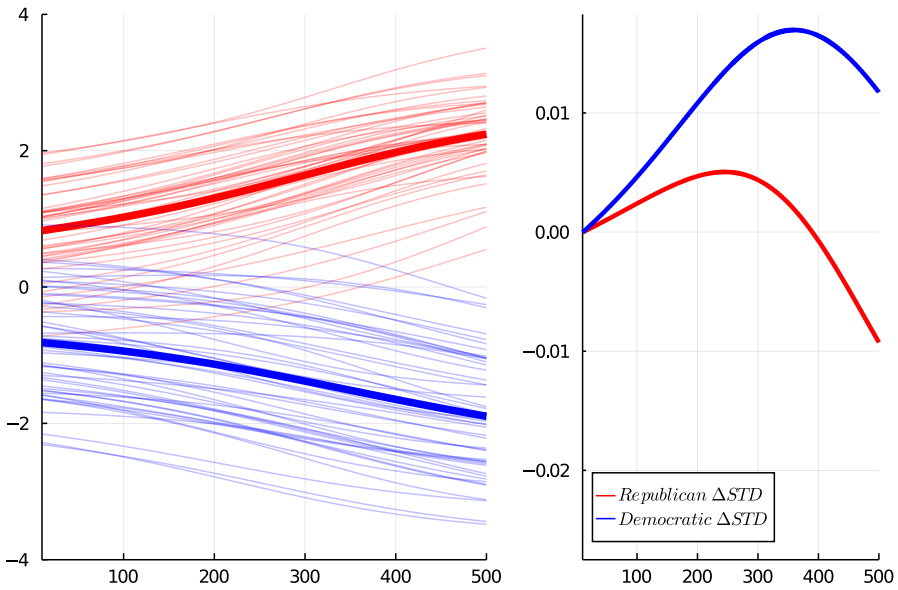


Fig. S12. Left. Thin lines are the evolution of ideological position of each Republican elite (red) and each Democratic elite (blue) as a function of time. Bold lines are the average Republican elite ideological position (red) and average Democratic elite ideological position (blue) as a function of time. Right: Evolution over time of the standard deviation of the ideological positions of Republican elites (red) and Democratic elites (blue) as compared to the standard deviations at the initial time.

References

1. A Bizyaeva, A Franci, NE Leonard, Nonlinear opinion dynamics with tunable sensitivity. *arXiv:2009.04332 [math.OC]* (2020).
2. A Franci, A Bizyaeva, S Park, NE Leonard, Analysis and control of agreement and disagreement opinion cascades. *Swarm Intell.* **15**, 47–82 (2021).
3. M Golubitsky, DG Schaeffer, *Singularities and Groups in Bifurcation Theory*, Applied Mathematical Sciences. (Springer-Verlag, New York, NY) Vol. 51, (1985).
4. N Fenichel, Geometric singular perturbation theory for ordinary differential equations. *J. Differ. Equations* **31**, 53–98 (1979).
5. DE Broockman, C Skovron, Bias in perceptions of public opinion among American political elites. *Am. Polit. Sci. Rev.* **112**, 542–563 (2018).
6. B Highton, Issue accountability in U.S. House elections. *Polit. Behav.* **41**, 349–367 (2019).
7. M Grossmann, DA Hopkins, *Asymmetric Politics: Ideological Republicans and Group Interest Democrats*. (Oxford University Press), (2016).
8. TL Brunell, B Grofman, S Merrill, Components of party polarization in the U.S. House of Representatives. *J. Theor. Polit.* **28**, 598–624 (2016).