Preventing extreme polarization of political attitudes

Robert Axelrod[a,1,2], Joshua J. Daymude[b,1], and Stephanie Forrest[b,c,d,1]

[a]School of Public Policy, University of Michigan, Ann Arbor, MI 48109; [b]Biodesign Center for Biocomputing, Security and Society, Arizona State University, Tempe, AZ 85281; [c]School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ 85281; and [d]Santa Fe Institute, Santa Fe, NM 87501

Edited by Simon Asher Levin, Princeton University, Princeton, NJ, and approved June 10, 2021 (received for review March 11, 2021)

Extreme polarization can undermine democracy by making compromise impossible and transforming politics into a zero-sum game. "Ideological polarization"—the extent to which political views are widely dispersed—is already strong among elites, but less so among the general public [N. McCarty, *Polarization: What Everyone Needs to Know*, 2019, pp. 50–68]. Strong mutual distrust and hostility between Democrats and Republicans in the United States, combined with the elites' already strong ideological polarization, could lead to increasing ideological polarization among the public. The paper addresses two questions: 1) Is there a level of ideological polarization above which polarization feeds upon itself to become a runaway process? 2) If so, what policy interventions could prevent such dangerous positive feedback loops? To explore these questions, we present an agent-based model of ideological polarization that differentiates between the tendency for two actors to interact ("exposure") and how they respond when interactions occur, positing that interaction between similar actors reduces their difference, while interaction between dissimilar actors increases their difference. Our analysis explores the effects on polarization of different levels of tolerance to other views, responsiveness to other views, exposure to dissimilar actors, multiple ideological dimensions, economic self-interest, and external shocks. The results suggest strategies for preventing, or at least slowing, the development of extreme polarization.

political polarization | agent-based models | democracy | opinion change | ideology

Extreme polarization can undermine democracy by making compromise impossible (1, 2) and transforming politics into a zero-sum game, as James Madison observed in Federalist No. 10 (3). When this occurs, even a democratically elected majority may seek to solidify its control over political power by weakening the institutions and norms that ordinarily support the turnover of elites. There are many motivating examples of this dysfunctional political polarization, for example, the rise of Hitler, the American Civil War, the destruction of democracy in Venezuela, the increasing threats to democracy in Hungary, and the growing animosities in American politics over the past two or three decades (4).

At least two kinds of polarization, if carried to extremes, can undermine democracy. "Affective polarization" is already a serious problem in the United States; for example, Americans increasingly dislike and distrust those of the other party, whether Democrat or Republican (1, 5). The other kind, "ideological polarization," is the extent to which political views are widely dispersed. Ideological polarization is already strong among elites but is less pronounced among the general public; see ref. 6 and pp. 50–68 of ref. 7. In the future, ideological polarization among the US public may increase due to the already strong affective polarization, rising social inequality, and the collapse of cross-cutting belief structures into consolidated clusters (8, 9). Therefore, it is important to understand how to prevent the public from reaching dangerous degrees of ideological polarization.

This paper focuses on ideological polarization (henceforth "polarization") among the general public and asks: Is there a level of ideological polarization above which polarization becomes a runaway process? And, if so, what policy interventions can prevent such a dangerous positive feedback loop?

To address these questions, we developed an agent-based model (ABM) of ideological polarization to explore situations in which many actors influence each other in ways that don't lend themselves to equation-based models. In the ABM paradigm, each individual actor is represented explicitly and rules specify the mechanisms for interaction between actors. Simple ABMs, such as ours, are not intended to predict any particular historical event or future outcome. Instead, they can provide insight about important mechanisms and the role they play in determining system trajectories, for example, to highlight the consequences of a few simple assumptions about how people are influenced by each other. By design, ABMs can capture a distribution of possible outcomes, characterizing both typical and rare behavior.

To study ideological polarization, we selected a small set of mechanisms that influence opinion change: attraction to those with similar ideological (i.e., political) positions and repulsion from those with dissimilar positions. In our Attraction–Repulsion Model (ARM), the actors are assumed to follow a few simple rules about giving and receiving influence. The rules are not based on principles of rational calculation, that is, costs and benefits, or the forward-looking strategic analysis typical of game theory. Instead, the actors simply adapt their position in ideological space based on interactions with other actors. Note that, when one actor changes its position, the environment of all of the other actors is affected. Based on idealized simple rules of interaction, we investigate the emergent properties of the system over time. Because the proposed mechanisms can exist alongside other mechanisms, they are complementary with other treatments of polarization.

Significance

Democracies require compromise. But compromise becomes almost impossible when voters are divided into diametrically opposed camps. The danger is that intolerance will grow, democratic norms will be undermined, and winners will be reluctant to let the losers ever regain power. To better understand how polarization can be prevented, or at least slowed, we developed a simple model in which people tend to be exposed to and attracted by views similar to their own, but are repulsed by views that are too dissimilar. The policy implications are described in terms of level of tolerance to other views, responsiveness to other views, exposure to dissimilar views, multiple ideological dimensions, economic self-interest, and external shocks.

Author contributions: R.A., J.J.D., and S.F. designed research, performed research, analyzed data, and wrote the paper; and J.J.D. implemented and ran simulation code.

The authors declare no competing interest.

[1]R.A., J.J.D., and S.F. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: axe@umich.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2102139118/-/DCSupplemental.

Published December 6, 2021.

Most agent-based (10, 11) and statistical physics (12, 13) models of opinion change, including Axelrod's culture model (14), investigate the effects of "homophily" and "assimilation": the tendency to interact with and attract toward others with similar opinions. Our ARM joins a growing body of literature that additionally considers "differentiation" (or "negative influence"): the tendency to amplify difference from others with dissimilar opinions (15–20). Although empirical evidence for differentiation is mixed—for example, negative interactions are apparent in the US Senate (21) and on social media (22) but are not always observed in group discussion and opinion exchange experiments (23, 24)—differentiation can capture the empirical effects of external messaging (16) and public debate on controversial topics (18) that purely assimilative models do not. Unlike most other models with both assimilation and differentiation (16–20), our model assumes that 1) both the likelihood of interaction and the magnitude and direction of opinion change are affected by ideological distance and 2) these effects are uncorrelated. A notable exception is the model of Baldassarri and Bearman (15), but their model includes several additional features such as issue engagement and perception, complicating the task of characterizing the effects of interaction and opinion change.

Our analysis of polarization has several interesting features. First, our model of polarization dynamics is unique in its simplicity, representing opinion change based only on an individual's attraction to or repulsion from others' positions. Its very simplicity allows us to attribute the exact mechanisms that yield particular outcomes. Second, unlike many models of opinion change, the ARM avoids the common assumption that the direction and magnitude of opinion change are correlated, which may not necessarily be true. Third, in the context of models of polarization dynamics that include repulsion as well as attraction, our discoveries include 1) the identification of conditions under which a population first approaches convergence around a moderate position but then reverses direction and becomes highly polarized, that is, conditions under which the center does not hold; 2) the identification of conditions under which a few extremists can actually help prevent polarization; 3) the discovery that even weak attraction to one's own initial position (such as the effect of economic self-interest) can prevent polarization; and 4) insight into the seeming paradox that—contrary to many policy proposals—exposure to dissimilar views can actually exasperate rather than alleviate polarization.

The Attraction–Repulsion Model

The ARM has only two rules. Stated informally, one rule says that an actor tends to interact with those who have similar views. The second rule says that a (pairwise) interaction between similar actors reduces their difference, while an interaction between dissimilar actors increases their difference. The ARM is, in fact, a Markov process where the choice of interacting actors is stochastic but interactions' effects are deterministic. Conditioned on a given sequence of interacting pairs of actors, the ARM is a deterministic dynamical system. While some analytic results from Markov theory are potentially available, a simulation approach is preferred here.

In particular, we assume a population of N actors whose ideological positions are distributed in D-dimensional space, where each dimension corresponds to an ideological, or political, issue. Each actor has a position between 0.0 and 1.0 on any given dimension; our results focus primarily on one-dimensional (1D) space.* Many ABMs assume that actors' positions are initially distributed uniformly or randomly; we, instead, assume the normal (Gaussian) distribution which is more realistic for modeling human political views (6).† Initial positions are normally distributed with a mean of 0.5 and SD of $\sigma = 0.2$, unless stated otherwise. The two rules are defined as follows.

Interaction Rule. At each step, a random actor selects another actor uniformly at random and they interact with probability $(1/2)^{d/E}$, where d is the distance between them and E is a model parameter capturing actors' exposure to other points of view.

This rule reflects the idea that the probability of interaction decays with the distance between ideological positions, and, in our model, the decay is exponential. The scaling factor is the "halving distance" of exposure E. As an example, if two actors are distance E apart, they have a 50% chance of interacting; if they are $2E$ apart, they have a 25% chance; and if they are $E/2$ apart, they have a $1/\sqrt{2} \approx 71\%$ chance. A large exposure means that an actor is almost as likely to interact with someone with whom they disagree as with someone who has a similar position; that is, the population is largely unsorted.

Attraction–Repulsion Rule. All actors have identical tolerance, T, that determines whether the result of an interaction is attractive or repulsive. If an actor tolerates the position of the other, the actor moves a fraction R (responsiveness) of the distance toward the other. Otherwise, the actor moves the same distance away from the other, subject to the boundaries.‡

A motivating example for the attraction–repulsion (AR) rule is the Bail et al. (22) study showing that exposure to opposing views on social media can lead to repulsion. As an example of AR, if an actor is at 0.4 and $T = 0.15$, the actor will move closer to anyone between 0.25 and 0.55. If $R = 0.25$ and the actor interacts with another at 0.5, the actor will move a quarter of the way from its position at 0.4 to the other's position at 0.5, resulting in a new position of 0.425. On the other hand, if the other actor is outside of the tolerance range, the actor will be repulsed. For example, if the other actor were at 0.1, the distance between them would be $0.4 - 0.1 = 0.3$, which is greater than the tolerance of 0.15. In that case, the actor would move a fraction R of the distance between them away from the other, resulting in a position of $0.4 + (0.25)(0.3) = 0.475$.

Unless the actor is already near an extreme, an interesting effect of this rule is that repulsion moves actors farther than attraction does. The distance moved is always an R fraction of the distance between the interacting actors which is greater than T for repulsion and less than T for attraction. Yet actors within distance T are also more likely to interact because they are closer. Repulsion is less likely than attraction, but it leads to greater movement when it does occur.

Polarization Metric and Default Parameters. We operationalize the degree of polarization as the variance of political positions in a population. For example, suppose there is one ideological dimension with actors distributed according to a normal (Gaussian) distribution with SD $\sigma = 0.2$. Since variance is σ^2, the polarization is $\sigma^2 = 0.04$. Each ideological dimension ranges from 0.0 to 1.0, so the maximum polarization for the 1D setting is 0.25 when half the population are extremists at zero and the other half are at one. The minimum polarization is zero, which occurs if the population converges to a single point.

*We assume that ideological positions are bounded to prevent repulsion from causing unbounded moves and because ideological positions are typically measured in surveys by questions with a limited integer range, for example, on a five- or seven-point scale.

†Populations initialized according to estimates of 1D latent ideology extracted from 2020 survey data (25, 26) yield similar behavior (see *SI Appendix*, Fig. S1 and *Empirical Initial Distribution*).

‡Instead of deterministic attraction within a tolerance distance and repulsion beyond it, one could model the probability of repulsion as a smooth stochastic function that increases with actors' pairwise distance to obtain similar results (see *SI Appendix*, *Stochastic Attraction–Repulsion*).

Table 1. Default parameter values for the ARM

Description	Default value
Number of actors	$N = 100$
Number of ideological dimensions	$D = 1$
Mean and SD of actors' initial normal distribution along an ideological dimension	(0.5, 0.2)
Exposure, the degree to which actors interact with dissimilar points of view expressed as the halving distance	$E = 0.1$
Tolerance, the distance within which interactions are attractive and beyond which interactions are repulsive	$T = 0.25$
Responsiveness, the fractional distance an actor's ideological position moves as a result of an interaction	$R = 0.25$

Table 1 gives the default parameter values used for our experiments. There is no natural calibration of time in the model, so we adopt the unit of actor activations, called steps. A helpful way to think about time is to suppose that, on average, each actor has one activation per day. Of course, only some of the activations result in interaction and movement, with similar actors more likely than distant actors to interact. With 100 actors, 36,500 steps of the simulation represents 1 y, and 1 million steps represents about 27 y.

Results

The findings of our simulation experiments are presented in six categories: tolerance, responsiveness, exposure, multiple ideological dimensions, economic self-interest, and external shock. See *SI Appendix* for details on the model implementation.

Tolerance. We begin by exploring the effects of tolerance (T), that is, the level of ideological difference that actors find attractive rather than repulsive. Fig. 1 shows the development of polarization over time for different values of T. As expected, when actors have low T (≤ 0.15), many of their interactions are outside their tolerance, leading to repulsion which increases their distance. This produces extreme polarization, with roughly half the population at one extreme and the other half at the other. It is also unsurprising that, when T is sufficiently high (≥ 0.55), interactions are usually attractive and the population converges around a single position.

The interesting cases occur at intermediate levels; two representative examples are $T = 0.25$ and $T = 0.35$. As Fig. 1 shows, $T = 0.25$ produces polarization that increases slowly at first and then rapidly takes hold and goes to the extreme. In contrast, $T = 0.35$ produces a nearly constant, low level of polarization that remains stable for the entire simulation. To understand how the dynamics of these two runs play out so differently, Fig. 2 shows snapshots of the population's ideological positions at key points in the run.

Fig. 2 *A* and *E* shows the identical initial condition for both runs. Both the $T = 0.25$ and $T = 0.35$ runs quickly form a majority of moderates near the ideological center flanked by smaller extreme groups at the far left and far right (Fig. 2 *B* and *F*). These three groups are mutually intolerant and repulse one another. In the $T = 0.25$ run, repulsion gradually erodes the moderate majority (Fig. 2*C*) until all actors have joined an extremist group (Fig. 2*D*), while, in the $T = 0.35$ run, the moderate majority remains in dynamic equilibrium among positions $[0.45, 0.55]$ for the entire simulation (Fig. 2 *F–H* and Movie S1).

This small change in tolerance yields such divergent long-term outcomes because of a subtle and unanticipated effect that we label "repulsive extremism." A left extremist at position 0.0 repulses any actor to the right of position T, pushing it farther to the right; likewise, a right extremist at position 1.0 repulses any actor to the left of position $1 - T$, pushing it farther to the left. Moderates that interact repeatedly with the same extreme group may be repulsed out of the center to the opposite extreme, ultimately dissolving the majority (Fig. 2 *B–D*). However, if the repulsive effect leaves the moderates within the tolerance range of the central majority, their mutual attraction, combined with repulsion from the opposing extremists, will reinforce and shift the majority as it attracts and reabsorbs those that were repulsed. The emergent effect is one in which the majority reaches and remains in what appears to be a dynamic equilibrium of centrist positions (Fig. 2 *F–H* and Movie S1), maintaining a diversity of opinion over long time spans. Without repulsive extremism, the population would, instead, converge to a single ideological position from which it would never move.

The determining factor in the long-term effect of repulsive extremism, then, is frequency of interaction between extremists and moderates. This relies critically on the relative size of the extreme groups compared to the moderates, since pairs of actors are chosen uniformly at random to interact. Tolerance plays a key role in determining these sizes: Larger T leads to larger attraction which kick-starts the initial concentration of actors near the center into a strong central majority, while the few remaining actors are repulsed and form small extremist groups. When the extremist groups are small enough relative to the moderate majority, as in the $T = 0.35$ run, the center holds. With larger extremist groups such as those in the $T = 0.25$ run, there is higher probability of moderates having repeated interactions with extremists, eventually dissolving the moderate majority.

> With low tolerance, even interactions between similar actors can result in repulsion, leaving little hope of avoiding runaway polarization with all actors holding extreme positions. Sufficiently high tolerance, on the other hand, leads to consensus at a single ideological position. At intermediate levels, a moderate majority can form that is flanked by repulsive extremists.

Responsiveness. Actors can be more or less responsive to interactions, depending on how far they move when attracted

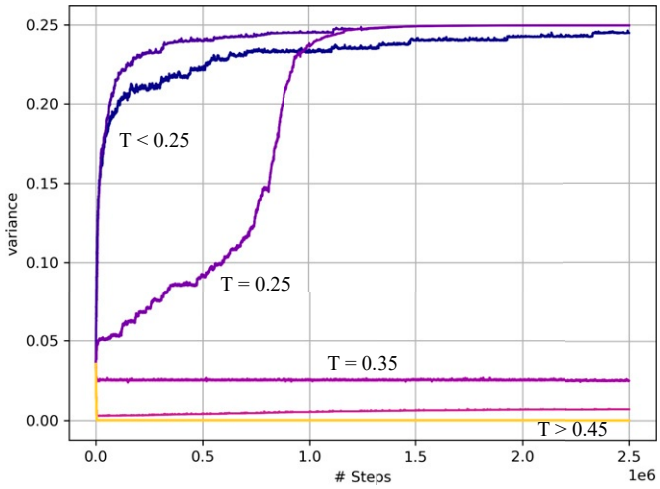

Fig. 1. The effects of tolerance (*T*). Polarization of the population's ideological positions over time when varying tolerance over the range $T = 0.05, 0.15, \ldots, 0.95$ (dark blue to yellow). Low tolerance ($T \leq 0.25$) leads to extreme polarization, intermediate tolerance ($0.35 \leq T \leq 0.45$) leads to small but nonzero polarization, and high tolerance ($T \geq 0.55$) leads to convergence.

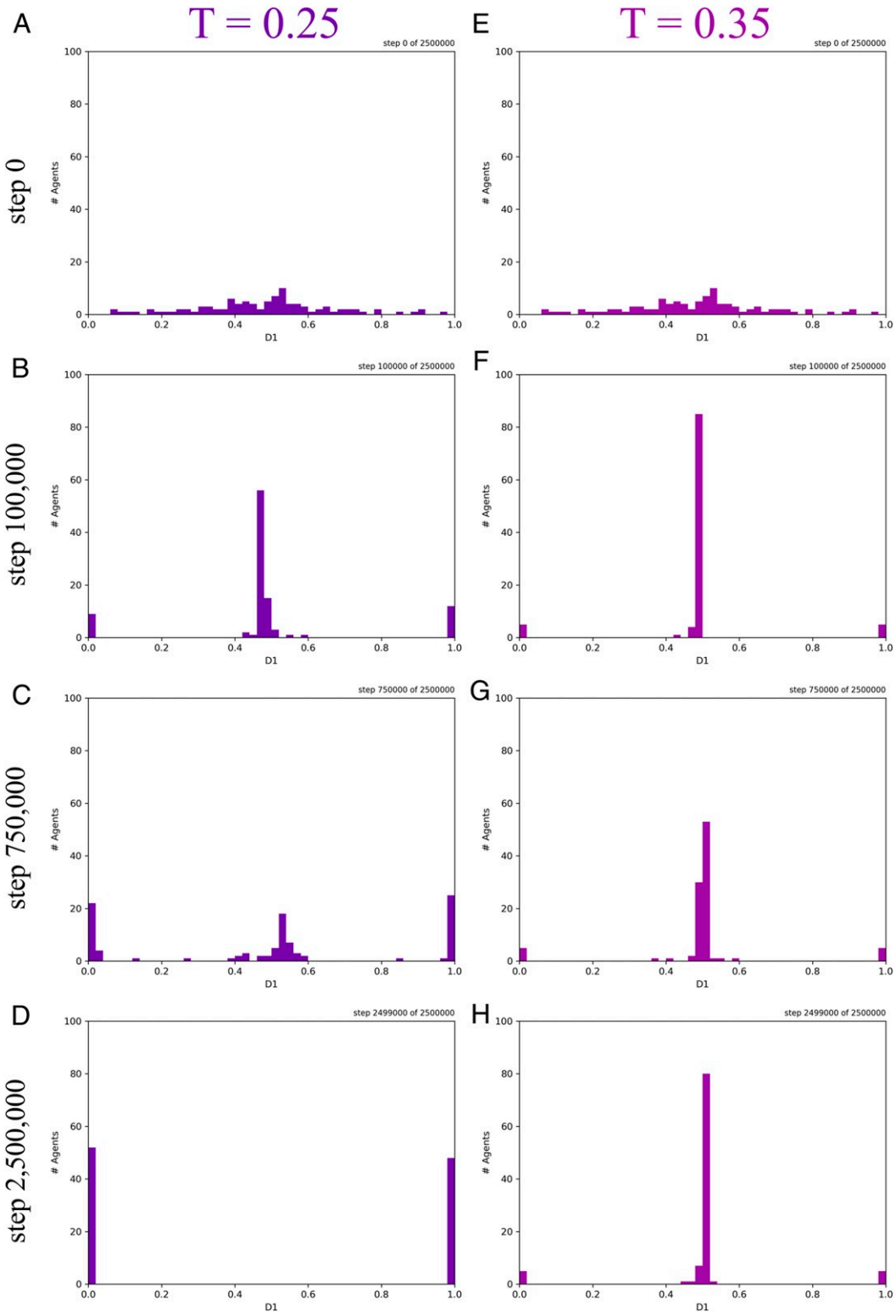

Fig. 2. With intermediate tolerance, can the center hold? Snapshots of the population's one-dimensional ideological positions (D1) over time, shown as histograms for the $T = 0.25$ and $T = 0.35$ runs shown in Fig. 1. (*A*) Initially, the 100 actors are normally distributed with mean 0.5 and SD $\sigma = 0.2$. (*B*) At step 100,000, the $T = 0.25$ run forms a moderate majority of ∼80 actors flanked by extreme groups at the far left and far right of ∼10 actors each. (*C*) The extreme groups grow steadily as the moderate majority dissolves. (*D*) After 2,500,000 steps or—using our estimation of one interaction per actor per day—about 70 y, all actors have converged to the extremes in equal proportions. (*E*–*H*) The $T = 0.35$ run forms and maintains a larger moderate majority (∼90 actors) that remains stable over all 2,500,000 steps. See Movie S1 for animations.

or repulsed by another. Recall that the AR rule says that, if an actor tolerates another's position, the actor moves a fraction R of the distance between them toward the other, and otherwise moves the same distance away from the other. Thus, larger values of R represent increased responsiveness.

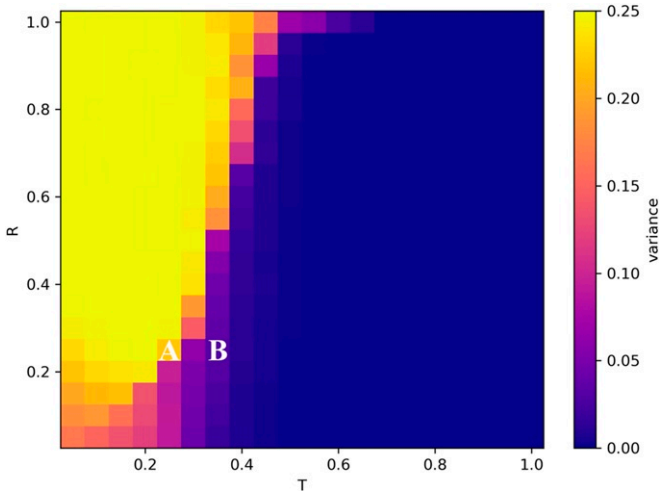

Fig. 3. The effects of responsiveness (*R*) as a function of tolerance (*T*). Average polarization of the population's ideological positions after 1,000,000 steps, averaged over 20 iterations for each (*T*, *R*) pair. *T* and *R* are both varied over the range 0.05, 0.10, . . . , 1.0. There is a phase change from extreme polarization (yellow) with low *T* to convergence (dark blue) with high *T*. The phase change is largely independent of *R*. A and B indicate the $T = 0.25$ and $T = 0.35$ cases shown in Fig. 2 on the boundary of the phase change.

Fig. 3 shows that the phase change from complete polarization to convergence to a single ideological position, shown in Fig. 1, is largely invariant with respect to R. Although actors move much more slowly when R is low (e.g., $R \leq 0.2$), it does not change the outcome. Intermediate polarization occurs on the boundary of the two regimes ($0.25 \leq T \leq 0.45$), as shown in Fig. 2. Interventions focused on responsiveness are thus unlikely to mitigate polarization because outcomes are largely determined by T except when R is very small.

Exposure. For a variety of reasons, people tend to be exposed less frequently to opinions that are different from their own than they are to similar opinions. The strength of this tendency is called the population's "exposure." Low exposure means the tendency is strong, while high exposure means that actors listen to distant and similar opinions almost equally. Empirically, exposure is an important factor in both affective and ideological polarization (1, 5). In the ARM, the interaction rule captures exposure by stating that the probability of an actor interacting with another is halved as the distance between them is doubled, scaled by the halving distance E. Put another way, population exposure increases with E.

Fig. 4 shows the effect of different levels of exposure for different levels of tolerance. For all but the lowest exposures ($E \geq 0.1$), tolerance dominates exposure in determining the population's outcome: Just as in Fig. 3, low tolerance ($T \leq 0.25$) leads to extreme polarization, while sufficiently high tolerance ($T \geq 0.5$) yields total convergence.

At intermediate levels of tolerance ($0.3 \leq T \leq 0.4$), polarization increases with exposure. To explore this further, we fix $T = 0.3$ and investigate the population's polarization over a longer period for varying degrees of exposure (Fig. 5). No runs produce convergence, but we observe two distinct types of polarization. When $E \geq 0.15$, actors often interact with and are repulsed by those who have dissimilar opinions, quickly leading to extreme polarization. Maximum population variance is not always achieved because the extreme groups may have asymmetric sizes (see, e.g., Movie S2, where the $E = 0.25$ run produces 30% left extremists and 70% right extremists, for a variance of $\sigma^2 \approx 0.2$). On the other hand, when actors rarely interact with anyone with different opinions ($E \leq 0.1$), the population can maintain a stable moderate majority flanked by extremist groups for many time steps (Movie S2).

> When a population is stubbornly intolerant, the majority of interactions are repulsive. Increasing exposure only makes these interactions more likely, leading to polarization with rapid adoption of extreme positions. Low exposure, however, preserves clusters of like-minded individuals by greatly decreasing the probability of repulsive "cross-cultural" interactions, inhibiting polarization.

Multiple Ideological Dimensions. The previous investigations can be extended to a setting with multiple ideological dimensions. For simplicity, we consider a 2D ideological space modeled as the unit square in which each actor has an ideological position ranging from 0.0 to 1.0 on each dimension. The AR rule parameterized by tolerance T and responsiveness R remains the same, but with pairwise distances between actors computed in 2D Euclidean space. Note that, in two dimensions, the maximum distance between two ideological positions is $\sqrt{2} \approx 1.414$ and the maximum variance is $\sigma^2 = 0.5$.

> Tolerance and responsiveness have similar effects in two ideological dimensions (*SI Appendix*, Fig. S4) as in one (Fig. 3), yielding a phase change from extreme polarization with low tolerance to consensus with sufficiently high tolerance.

With two ideological dimensions, we can investigate populations that have different exposures per dimension. We generalize the interaction rule to consider two exposures, E_1 and E_2. Similar to the 1D case, with high E_1 and E_2, actors often interact with and are repulsed by others beyond their tolerance, causing extreme polarization (*SI Appendix*, Fig. S5). However, populations with low exposure even on just one dimension avoid extreme polarization for most degrees of exposure on the other dimension.

To better understand how low exposure on just one dimension can help mitigate polarization, we consider the situation where actors with disparate views on the first ideological dimension are less likely to interact ($E_1 = 0.1$) while varying

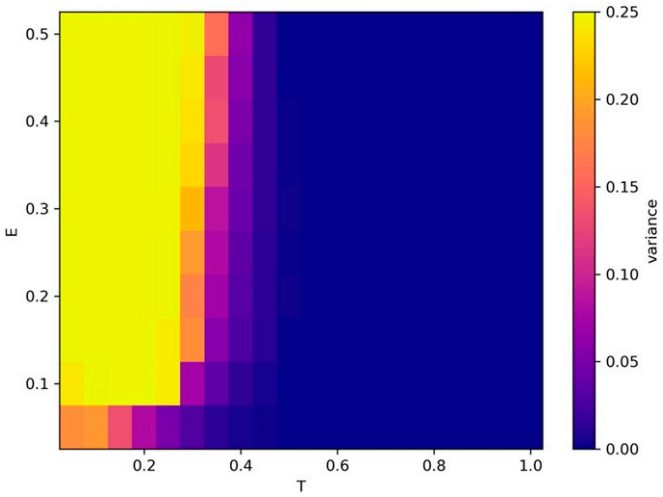

Fig. 4. The effects of exposure (*E*) as a function of tolerance (*T*). Average polarization of the population's ideological positions after 2,000,000 steps, averaged over 20 iterations for each (*T*, *E*) pair. Tolerance is varied over $T = 0.05, 0.1, \ldots, 1.0$, and exposure is varied over $E = 0.05, 0.1, \ldots, 0.5$.

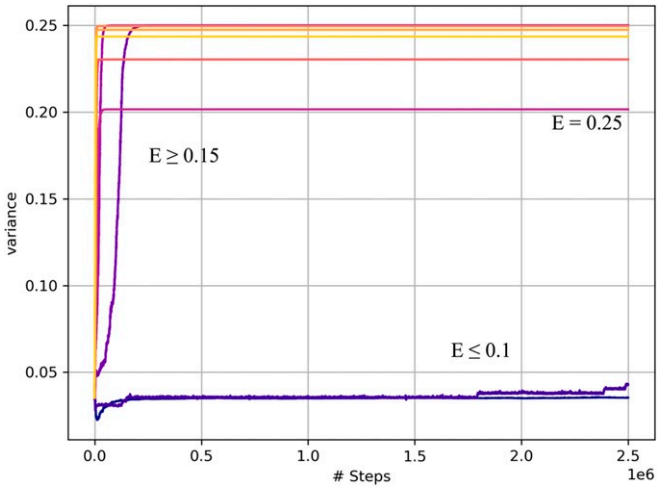

Fig. 5. The effects of exposure (E) for intermediate tolerance (T). Polarization of the population's ideological positions over time when $T = 0.3$ is fixed and exposure is varied over the range $E = 0.05, 0.1, \ldots, 0.5$ (dark blue to yellow). $E \leq 0.1$ leads to a stable moderate majority flanked by repulsive extremists, while $E \geq 0.15$ leads to rapid polarization.

E_2. Fig. 6 shows two distinct population behaviors. When $E_2 = 0.05$, intermediate polarization occurs slowly. This effect is a generalization of repulsive extremists to two dimensions: Mutually repulsive groups emerge and move to positions at the boundaries and corners of ideological space that are the most stable, interacting only occasionally with other groups (Fig. 6, *Bottom Inset*).

The second behavior shown in Fig. 6 is large but oscillating polarization occurring when $0.2 \leq E_2 \leq 0.45$. As an example, Fig. 6, *Top Inset* shows the final configuration of the $E_2 = 0.4$ run. The population has polarized completely along the second dimension, with all actors on either the top or bottom edge, due to high exposure on that dimension enabling many repulsive interactions. Surprisingly, the same is not true of the first dimension. After 2,500,000 steps, only the top edge has polarized into opposite corners; actors on the bottom edge are distributed bimodally. The left- and right-leaning bottom actors repulse each other as in the 1D setting, but, unlike in one dimension, they maintain a diversity of ideological positions. Whenever a bottom actor gets close to the bottom-right corner, for example, the first dimension's low exposure causes it to interact frequently with the top-right corner of actors. These two right corners are mutually intolerant and repulsive, causing the bottom actor to be pushed back out of the bottom-right corner. This subtle and surprising effect shows that frequent interactions between actors close to one another on the low-exposure dimension but potentially very far from each other on the high-exposure dimension can cause oscillations on the low-exposure dimension that remain stable for long time periods.

Although asymmetric exposures yield interesting new behaviors, the key takeaway for higher dimensions is analogous to what was observed in one dimension. Low exposure can limit interactions between mutually intolerant groups, enabling the population to avoid a flurry of repulsive interactions that can dissolve moderate clusters. Extreme polarization is avoided as long as these moderate clusters persist, as shown in Fig. 6.

Economic Self-Interest. We next consider a 1D variant of the ARM in which each actor has a preferred ideological position based on economic self-interest.[§] For example, an actor with low income might have an economic self-interest in a position to the left of the median. Although income, wealth, and many other attributes are highly skewed, it is not unreasonable to assume that positions follow a normal distribution (see *SI Appendix*, Fig. S1 and *Empirical Initial Distribution* for a discussion of empirical ideological data). Therefore, we assume that an actor's initial position represents its preferred position. We model the effect of economic self-interest by assuming that, with fixed probability P, an actor is attracted to its preferred position rather than interacting with another actor. Thus, P can be thought of as the strength of one's own self-interest in comparison to the effects of other actors in the population.

Fig. 7 shows polarization over time for P ranging from 0 to 10%. The top line is the familiar case when self-interest has no effect ($P = 0\%$) and the population polarizes to the extreme. But even a very small increase in P avoids extreme polarization: With $P = 1\%$, the population's polarization is roughly halved. As P increases ($4\% \leq P \leq 10\%$), the population rarely strays from the low level of polarization present in the initial normal distribution.

Without self-interest ($P = 0\%$), the population polarizes in roughly equal proportions to each extreme (Fig. 7, *Top Right Inset*). We use this run as a baseline to understand the relative impact of increasing self-interest. At very low levels ($P = 1\%$; Fig. 7, *Middle Right Inset*), self-interest does not prevent the formation of two mutually repulsive clusters, but it still has a moderating influence. Because actors are attracted to their more moderate preferred positions, the runs produce bimodal distributions with peaks that oscillate in the $[0.1, 0.2]$ and $[0.8, 0.9]$ ranges, respectively, avoiding extreme polarization. A 10-fold increase in self-interest ($P = 10\%$; Fig. 7, *Bottom Right Inset*) maintains the center and avoids polarization. The combination of self-interest and attracting interactions produces a tighter concentration of actors near the center than the initial/preferred normal distribution (Fig. 7, *Left Inset*).

> Even a small amount of self-interest, which biases actors toward their initial positions, can dramatically reduce polarization (Fig. 7 and *SI Appendix*, Fig. S6).

External Shock. Our final experiments consider external shocks that exogenously shift actors' ideological positions so they become less polarized as they unify around a common problem. Four examples of such shocks are as follows: 1) climate change awareness, where, as more people experience the reality of increasingly frequent and damaging natural disasters, they may become more willing for the government to take costly action to mitigate its effects; 2) the COVID-19 pandemic, where, as infection rates and death counts soar, people may become more willing to support actions such as mask mandates, shutdowns, and emergency funding for testing, contact tracing, and vaccine deployment; 3) economic recession, where, when economic downturns such as the Great Depression, the 2008 recession, and the 2020 economic crisis occur, people become more willing to accept large budget deficits to stimulate the economy; and 4) war, where, when a country declares war or an existing war escalates, people become more willing to pay higher taxes to support the military.

In the ARM, an external shock has a strength Δ and a time (i.e., step) at which it occurs. At the shock's specified step, all actors' ideological positions are shifted to the right by the shock strength Δ, subject to the usual constraint that no actor can

[§]We describe the preference as due to economic self-interest, but it could represent many other reasons for an actor's innate preference for a certain position.

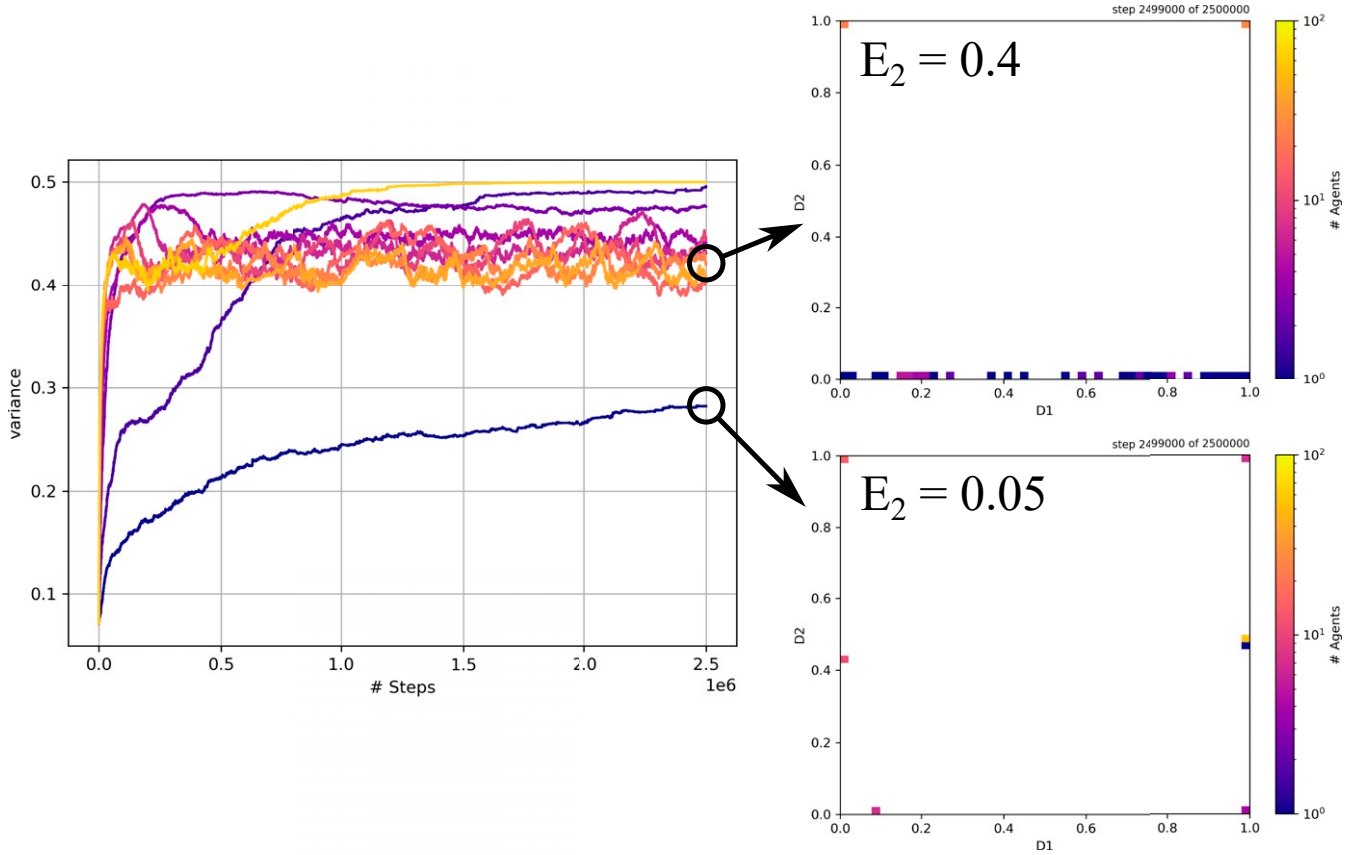

Fig. 6. Avoiding maximum polarization with low exposure (E). Polarization of the population's ideological positions over time when exposure on the first ideological dimension is fixed at $E_1 = 0.1$ and exposure on the second ideological dimension is varied over $E_2 = 0.05, 0.1, \ldots, 0.5$ (dark blue to yellow). (*Insets*) Final configurations of the population after 2,500,000 steps for the $E_2 = 0.4$ (*Top*) and $E_2 = 0.05$ (*Bottom*) runs as a 2D histogram whose colors indicate concentrations of actors on a log-scale. See Movie S3 for animations.

have a position greater than 1.0. We use the default parameters (Table 1) to investigate the effects of external shocks of varying strengths introduced at step 500,000. Fig. 8*A* shows three distinct outcomes stemming from the same underlying effect: Either the shock strength Δ is large enough to shift ideological groups that were previously repulsive within one another's range of tolerance, enabling them to attract and eventually merge, or the distinct ideological groups remain mutually intolerant and repulse one another.

In more detail, recall that runs with default parameters quickly form a moderate majority flanked by small groups of repulsive extremists (Fig. 8*B*, *Left*). Weak shocks ($\Delta \leq 0.15$) shift the left extremists and moderates to the right by a small amount, but the right extremists are already at the maximum (Fig. 8*B*, $\Delta = 0.10$). Thus all three groups remain mutually repulsive, the moderate majority is eventually dissolved, and the population converges to extreme polarization as if the shock never happened. At the other end of the spectrum, very strong shocks ($\Delta \geq 0.75$) make all actors mutually attractive, leading to consensus (e.g., Fig. 8*B*, $\Delta = 0.8$). Intermediate shock strengths ($0.2 \leq \Delta \leq 0.7$) weaken intolerance between the moderates and right extremists but are not strong enough to do so for the moderates and left extremists (Fig. 8*B*, $\Delta = 0.40$). The left extremists repulse the now-merged moderates and right extremists, resulting in ~20 actors at the extreme left and the remaining ~80 at the extreme right.[¶]

We next consider the effect of introducing shocks at different times. Fig. 9 shows that the three phases of weak, intermediate, and strong shocks observed in Fig. 8 occur regardless of when the shock is applied. Medium-strength shocks lead to intermediate polarization, with earlier shocks resulting in smaller variance. However, no medium-strength shock avoids polarization entirely once groups of extremists have already formed. In particular, if the left extremists remain repulsive with all other actors (which are necessarily farther to the right) after the shock, they will be repulsed back to the left extreme. The timing of this type of shock only affects how large the consolidation of moderate and right-leaning actors is, which affects the final level of polarization.

> An external shock, which moves all actors in the same direction, only reduces polarization if it occurs so early that extremist groups have not yet formed or is strong enough to move all actors within a single tolerance distance.

Model Extensions

The ARM can facilitate many natural extensions to study other aspects of polarization, including asymmetry, elites, affective polarization, politics, geography, and other interventions.

Asymmetry. To explore some of the asymmetries in current American politics (4), one could use asymmetric initial distributions or vary T across actors.

[¶]Since all runs in Fig. 8 use the same initial distribution and random seed, all intermediate shock strengths lead to the same proportion of left and right extremists.

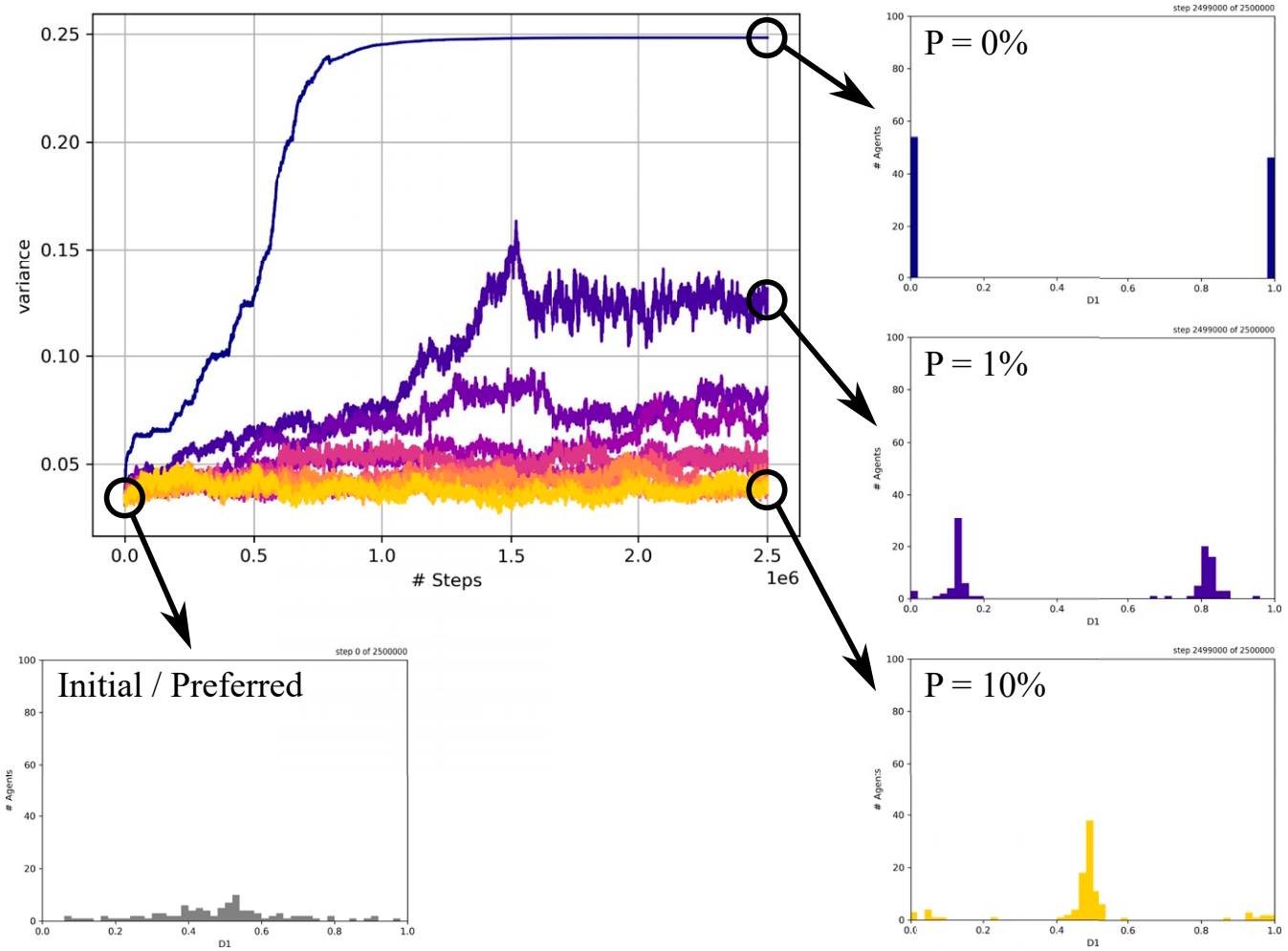

Fig. 7. The effects of economic self-interest (*P*). Polarization of the population's ideological positions over time with varying levels of economic self-interest, $P = 0\%, 1\%, \ldots, 10\%$ (dark blue to yellow) that an actor will be attracted to its preferred (initial) position. (*Left Inset*) The initial normal distribution of actors' ideological positions, which also represent their preferred positions when acting in self-interest. (*Right Insets*) Final configurations of the population after 2,500,000 steps for $P = 0\%$, 1%, and 10%. See Movie S4 for animations.

Elites. Elites with large social influence can be represented as fixed ideological positions that others occasionally attract to as in economic self-interest, or as actors who are more likely to be selected by others for interaction.

Affective Polarization. To study affective polarization, one could assign each actor as Democrat, Republican, or Independent, with initial positions normally distributed around their party's mean, for example, 0.4, 0.5, and 0.7, respectively. The degree of affective polarization could be represented as the probability that a member of one party will tolerate a member of the same party and be repulsed by a member of the other party, regardless of distance.

Politics. Starting with this model of party affiliations, the primary campaigns of the two parties could be represented as selection among members of one's own party, stopping after a given number of steps. At that point, the candidate for each party would be assigned the median position of that party's members. The general election campaign would then treat the two candidates as elites, as described above. A simple way to represent the general election would have each actor vote for its nearest candidate after a fixed number of steps (the campaign), with the candidate with the most votes declared the winner.

Geography. To represent a world in which geography limits interactions, assign each actor to a location on a 2D grid. Then have an actor select another only from their immediate neighborhood, as in Axelrod's culture model (14).

Other Interventions. Beyond incentives for self-interest or externally applied shocks, one could consider other interventions such as educational campaigns that nudge the population or interventions that affect actors differentially.

Discussion

Despite its simplicity, the ARM sheds light on the original questions. First is the question of when ideological polarization becomes a runaway process leading to extremism. Intolerance is the key component of runaway polarization observed in our experiments (Figs. 1–3), especially when enhanced by high exposure that enables frequent repulsive interactions between dissimilar individuals (Fig. 4). These results also generalize to additional ideological dimensions; for example, runaway polarization occurs with low tolerance when both dimensions have high exposure (*SI Appendix*, Figs. S1 and S2).

Our second question concerns policy: What interventions can prevent extreme polarization? Sufficiently high tolerance can

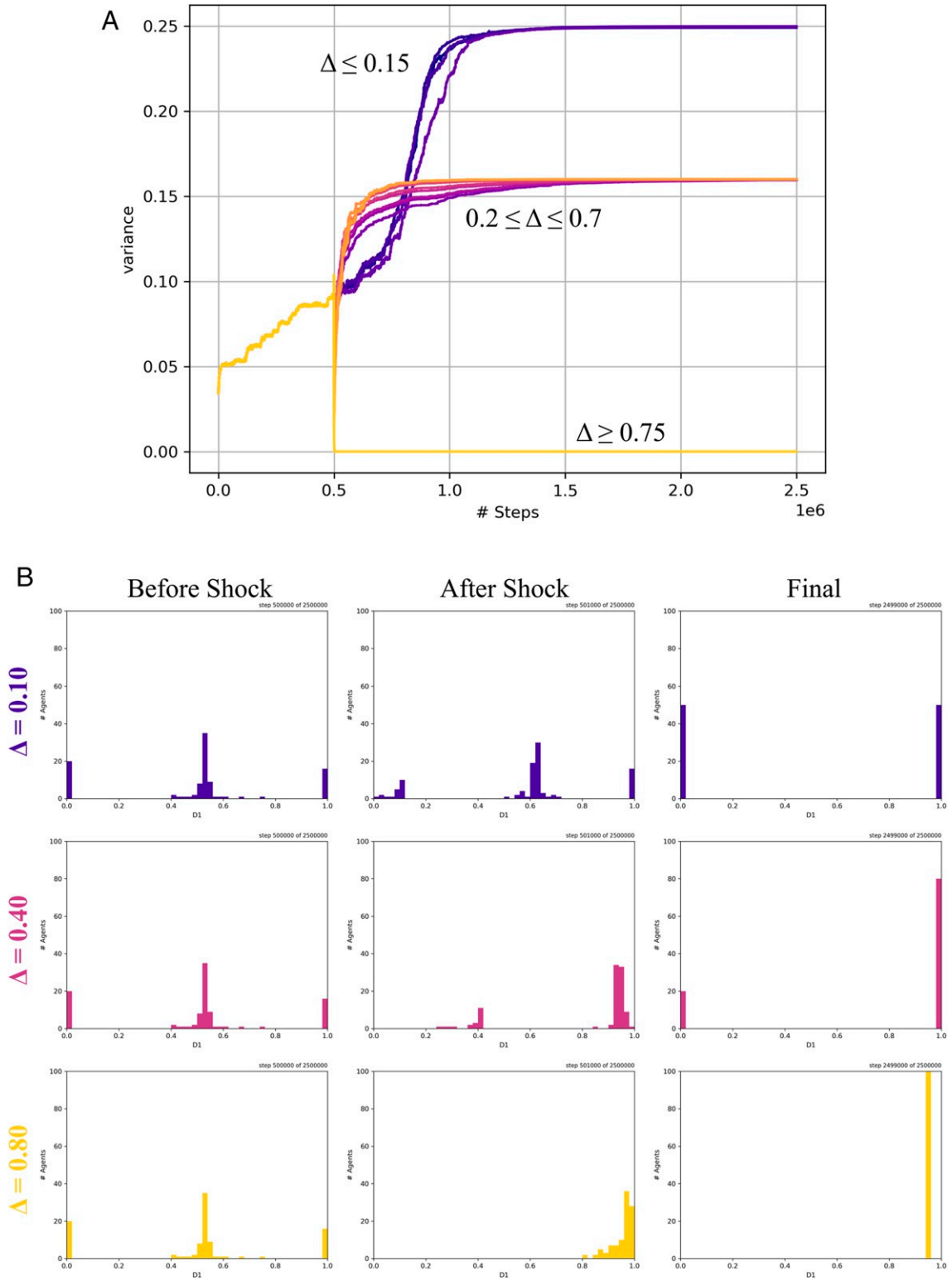

Fig. 8. The effects of external shock strength (Δ) on repulsive extremists. (*A*) Polarization of the population's ideological positions over time with external shocks of varying strengths $\Delta = 0.0, 0.05, \ldots, 0.8$ (dark blue to yellow) introduced at step 500,000. (*B*) Snapshots of the population's ideological positions as histograms for the $\Delta = 0.1$, 0.4, and 0.8 runs just before the shock (step 500,000; *Left*), shortly after the shock (step 501,000; *Middle*), and in the final configuration (step 2,500,000; *Right*). See Movie S5 for animations.

prevent or dramatically slow polarization (Figs. 1 and 2). For example, when $T = 0.25$, the population polarizes but only very slowly; the center forms but does not hold. A small increase in tolerance can prevent runaway polarization, but the critical value depends on other parameters such as responsiveness (Fig. 3) and exposure (Fig. 4).

Interventions focused on responsiveness are unlikely to mitigate polarization, because outcomes are largely determined by

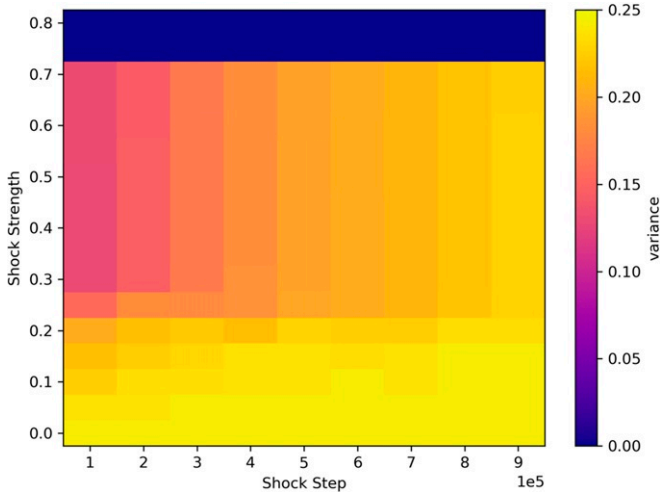

Fig. 9. The effects of external shock, by time and strength. Average polarization of the population's ideological positions after 2,000,000 steps, averaged over 20 iterations for each (Δ, step) pair. Shocks vary in strength over $\Delta = 0.0, 0.05, \ldots, 0.8$ and are introduced at steps 100,000, 200,000, . . ., 900,000.

tolerance effects unless responsiveness is very small (Fig. 3). Strictly limiting exposure to dissimilar views, however, is an effective mechanism for avoiding rapid polarization (Figs. 4 and 5). This may, at first, appear contrary to practical experience: Encouraging interactions among those with different views might be expected to decrease polarization by fostering increased tolerance. However, the ARM treats tolerance and exposure as independent features of a population. If a population is stubbornly intolerant and on a trajectory to runaway polarization, low exposure decreases the probability of interactions between mutually intolerant groups. This, in turn, preserves ideological diversity by inhibiting repulsion.

These results help resolve the controversy about whether contact between different groups tends to increase or decrease their hostility (27, 28). For example, school desegregation brought people of different races together, expecting that exposure to others would reduce hostility. However, experience in places like Boston in the 1970s (29) show the opposite because the differences were too great and may have exacerbated preexisting hostility.

When there are two ideological dimensions, extreme polarization can be avoided if the population has low exposure on either issue (*SI Appendix*, Fig. S5). In other scenarios, however, we observe a "durable correlation of multiple dimensions." This so-called collapse of dimensions can be a threat to democracy, as observed by James Madison in Federalist No. 10 (3).

Attraction to one's preferred ideological position is surprisingly effective at preventing polarization, even if this preference is very small compared to the influence of others. Under these conditions, the population tends to moderate (Fig. 7 and *SI Appendix*, Fig. S6). This is perhaps the most promising result of the model, because it suggests a direction for policy intervention by which a polarizing dynamic could be moderated.

One might imagine that an external event that moves everyone in one direction could be another mechanism for controlling polarization, but our model shows that such a shock has to be surprisingly large to succeed (Fig. 8). If the population is on a polarizing trajectory, the sooner the shock occurs, the more likely it is to have an effect (Fig. 9). If the population is already somewhat polarized, a weak shock does not stop the process, although a strong shock can. Even a medium shock (e.g., to the right) only changes the relative sizes of extreme groups, but, in the end, all actors become extremists, with moderates and right extremists combined and polarized from the left extremists (Fig. 8*B*, $\Delta = 0.40$). In future work, it will be interesting to compare these unidirectional shocks to those that produce centrist, "rally around the flag" effects.

Without intervention, the ARM typically exhibits either extreme polarization or complete convergence (Figs. 3 and 4). Although we have focused on preventing extreme polarization, we note that some diversity of opinions is bound to exist in any open society and is likely necessary for holding elites accountable and sustaining healthy democracy. In the ARM, we observe some scenarios where a few repulsive extremists help reinforce the center, allowing a moderate majority to remain stable for long periods of time while continuing to react against extremists (Fig. 2 *E–H*). With too many extremists, however, the cluster of moderates ultimately dissolves (Fig. 2 *A–D*). Ideal levels of ideological diversity in a society thus may be sensitive to factors such as interaction patterns and group structures.

Conclusion

With just two simple rules, the ARM yields complex dynamics that provide policy-relevant insights into mechanisms for preventing extreme polarization. While repulsion is often omitted from models of political polarization, we find that, in some circumstances, repulsion from those of whom we are intolerant can reinforce a moderate majority.

One advantage of an ABM such as the ARM is its ability to generate distributions of possible outcomes on different runs of the model, such as rare "black swan" events which are difficult to obtain with equation-based models. For example, while typical runs with $T = 0.35$ yield low polarization (Figs. 2 *E–H* and 3*B*), rare initial distributions and interaction patterns can lead to polarization 4 times as large (Movie S6). While such events are rare by definition, they are important because they can have large and lasting consequences when they do occur. An interesting question is how to determine whether particular historical events, such as the rise of Hitler, are black swans or expected outcomes of general dynamical processes. Because democracies require compromise, which is almost impossible when electorates are deeply divided, understanding the forces that promote or inhibit political polarization is crucial for sustaining democracies when they are challenged—whether by black swans or predictable trends. Models like those in this Special Feature on the dynamics of polarization can help, especially if they are both simple enough to understand where the results come from and subtle enough to provide relevant new insights.

Materials and Methods

Details of the ARM and its simulation implementation are given in *SI Appendix*. The corresponding source code and execution instructions are available online at https://github.com/jdaymude/AttractionRepulsionModel.

Data Availability. Source code and execution instructions are available on GitHub (https://github.com/jdaymude/AttractionRepulsionModel). All study data are included in the article and *SI Appendix*.

ACKNOWLEDGMENTS. We thank D. Axelrod, V. Axelrod, S. Croker, D. Kinder, C. Tausanovitch, C. Warshaw, B. Edwards, and the participants in this PNAS Special Feature, especially the organizers H. Milner and S. Levin. R.A. thanks the University of Michigan for research support. J.J.D. is supported in part by the NSF under Award CCF-1733680, the Army Research Office under MURI Award W911NF-19-1-0233, and the Arizona State University Biodesign Institute. S.F. is supported in part by the NSF under Award CCF-1908633, the Defense Advanced Research Projects Agency under Awards FA8750-19C-0003 and N6600120C402, the Air Force Research Laboratory under Award FA8750-19-1-0501, and the Santa Fe Institute.

1. L. Mason, A cross-cutting calm: How social sorting drives affective polarization. *Publ. Opin. Q.* **80**, 351–377 (2016).
2. J. W. Patty, E. M. Penn, Are moderates better representatives than extremists? A theory of indirect representation. *Am. Pol. Sci. Rev.* **113**, 743–761 (2019).
3. J. Madison, "Federalist No. 10" in *The Federalist Papers*, C. Rossiter, Ed. (New American Library, New York, 1961), pp. 77–84 (1787).
4. S. Levitsky, D. Ziblatt, *How Democracies Die* (Crown, 2018).
5. S. Iyengar, Y. Lelkes, M. Levendusky, N. Malhotra, S. J. Westwood, The origins and consequences of affective polarization in the United States. *Annu. Rev. Polit. Sci.* **22**, 129–146 (2019).
6. V. C. Yang, D. M. Abrams, G. Kernell, A. E. Motter, Why are U.S. parties so polarized? A "satisficing" dynamical model. *SIAM Rev.* **62**, 646–657 (2020).
7. N. McCarty, *Polarization: What Everyone Needs to Know* (Oxford University Press, 2019).
8. D. DellaPosta, Pluralistic collapse: The "oil spill" model of mass opinion polarization. *Am. Socio. Rev.* **85**, 507–536 (2020).
9. R. D. Putnam, S. R. Garrett, *The Upswing: How America Came Together a Century Ago and How We Can Do It Again* (Simon & Schuster, 2020).
10. A. Flache *et al.*, Models of social influence: Towards the next frontiers. *J. Artif. Soc. Soc. Simulat.* **20**, 2 (2017).
11. A. Flache, Between monoculture and cultural polarization: Agent-based models of the interplay of social influence and cultural diversity. *J. Archaeol. Method Theor.* **25**, 996–1023 (2018).
12. C. Castellano, S. Fortunato, V. Loreto, Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 591–646 (2009).
13. S. Redner, Reality-inspired voter models: A mini-review. *Compt. Rendus Phys.* **20**, 275–292 (2019).
14. R. Axelrod, The dissemination of culture: A model with local convergence and global polarization. *J. Conflict Resolut.* **41**, 203–226 (1997).
15. D. Baldassarri, P. Bearman, Dynamics of political polarization. *Am. Socio. Rev.* **72**, 784–811 (2007).
16. T. V. Martins, M. Pineda, R. Toral, Mass media and repulsive interactions in continuous-opinion dynamics. *Europhys. Lett.* **91**, 48003 (2010).
17. A. Flache, M. W. Macy, Small worlds and cultural polarization. *J. Math. Sociol.* **35**, 146–176 (2011).
18. S. M. Krause, F. Weyhausen-Brinkmann, S. Bornholdt, Repulsion in controversial debate drives public opinion into fifty-fifty stalemate. *Phys. Rev.* **100**, 042307 (2019).
19. 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**, e2102141118 (2021).
20. M. W. Macy, M. Ma, D. R. Tabin, J. Gao, B. K. Szymanski, Polarization and tipping points. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2102144118 (2021).
21. C. C. Liu, S. B. Srivastava, Pulling closer and moving apart: Interaction, identity, and influence in the U.S. Senate, 1973 to 2009. *Am. Socio. Rev.* **80**, 192–217 (2015).
22. C. A. Bail *et al.*, Exposure to opposing views on social media can increase political polarization. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 9216–9221 (2018).
23. M. Mäs, A. Flache, Differentiation without distancing. Explaining bi-polarization of opinions without negative influence. *PloS One* **8**, e74516 (2013).
24. K. Takács, A. Flache, M. Mäs, Discrepancy and disliking do not induce negative opinion shifts. *PloS One* **11**, e0157948 (2016).
25. B. Schaffner, S. Ansolabehere, S. Luks, Cooperative election study common content, 2020. Harvard Dataverse. https://doi.org/10.7910/DVN/E9N6PH. Accessed 14 April 2021.
26. C. Tausanovitch, C. Warshaw, Measuring constituent policy preferences in Congress, state legislatures, and cities. *J. Polit.* **75**, 330–342 (2013).
27. G. W. Allport, *The Nature of Prejudice* (Addison-Wesley, 1954).
28. T. F. Pettigrew, L. R. Tropp, A meta-analytic test of intergroup contact theory. *J. Pers. Soc. Psychol.* **90**, 751–783 (2006).
29. R. P. Formisano, *Boston Against Busing: Race, Class, and Ethnicity in the 1960s and 1970s* (University of North Carolina Press, 2004).