PNAS LETTER

The complexity of polarization

Scott de Marchi[a,1]

Axelrod et al. (ADF; ref. 1) have a laudable goal: to create a simple model that reveals the causal mechanisms underlying polarization in order to better understand and "prevent extreme polarization." All else equal, we should prefer simple models where comparative statics are straightforward. Policymakers desire monotonic effects, and models that produce highly conditional or nonlinear effects will get short shrift.

Models can, however, be too simple and fail to uniquely identify or test the causal mechanisms that would inform policy interventions. There are four reasons why the model presented by ADF (1) does not uniquely identify the mechanisms responsible for polarization.

First, ADF believe the model depicted in Fig. 1 produces emergent/unexpected results. Yet, in this model the "distance" between the assumptions and results is quite small and behavior depends on global parameters. When tolerance is low the population splits into two modes at the edges of the ideological space. When tolerance is high, agents "stick" in the center. Similarly, low exposure or responsiveness slows the rate of change in the model. Is the finding that more tolerance leads to less polarization a case where local rules lead to "surprising" macrolevel results, or does this repeat conventional wisdom (2)?

Second, the model seems incomplete. ADF initialize their model with a centrist population but also state that the United States suffers from "growing animosities" akin to the Civil War. The logic of their model indicates that intolerance leads to a "runaway process" with "little hope of avoiding" extreme polarization. If their model is correct, what has kept the (currently centrist) US population from the expected polarized outcome? Is it that intolerance is less than some critical value? Or is the model missing dynamics promoting moderation?

Third, the best evidence from scholars working in this area indicates that the behavioral rules chosen by ADF are incorrect. Affective polarization, not policy polarization, is crucial (3). ADF argue that one could simply change the labels of the policy space to "parties" and keep the same dynamics—but is it plausible that the cognitive mechanisms underlying policy preferences and emotional attachment to parties (or political leaders) are identical? There is also substantial research on when interactions between opposed ideologies yield positive versus negative outcomes; none of this work indicates that such encounters are uniformly "repulsive," as ADF assume (4).

Finally, it would be difficult to falsify ADF's model. By selecting parameter values for T, E, and R, one can achieve end states ranging from polarization to centrism. Given the weight of the assumptions and lack of empirical tests, should we trust their recommendation that our democracy consider limiting citizens' "exposure to dissimilar views"? Their conclusion is especially problematic given existing evidence that finds that limiting exposure increases rather than decreases polarization (5).

This raises more general questions. Can, as the authors argue, the simple algorithm in Fig. 1 explain such disparate phenomena as polarization, controversy over school desegregation, and the "rise of Hitler"? Or do overly simple models and the lack of hypothesis testing allow researchers to analogize too broadly (6)?

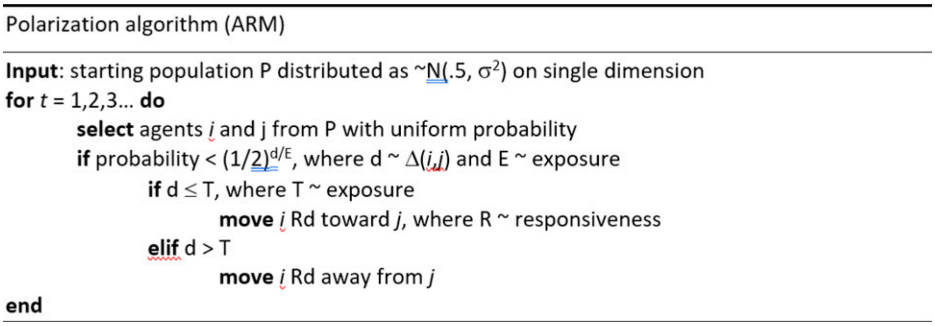
Author affiliations: [a]Department of Political Science, Duke University, Durham, NC 27708

Author contributions: S.d.M. designed research, performed research, and wrote the paper.

The author declares no competing interest.

[1] Email: demarchi@duke.edu.

Published April 21, 2022.



Polarization algorithm (ARM)

```
Input: starting population P distributed as ~N(.5, $\sigma^2$) on single dimension
for $t$ = 1,2,3... do
    select agents $i$ and j from P with uniform probability
    if probability < $(1/2)^{d/E}$, where d ~ $\Delta(i,j)$ and E ~ exposure
        if d $\leq$ T, where T ~ exposure
            move $i$ Rd toward $j$, where R ~ responsiveness
        elif d > T
            move $i$ Rd away from $j$
end
```

Fig. 1. Polarization algorithm. ARM, Attraction–Repulsion Model. Reprinted with permission from ref. 1.

1. R. Axelrod, J. J. Daymude, S. Forrest, Preventing extreme polarization of political attitudes. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2102139118 (2021).
2. J. Miller, S. Page, *Complex Adaptive Systems* (Princeton University Press, 2009).
3. S. Iyengar *et al.*, The origins and consequences of affective polarization in the United States. *Annu. Rev. Polit. Sci.* **22**, 129–146 (2019).
4. D. M. Ryfe, Does deliberative democracy work? *Annu. Rev. Polit. Sci.* **8**, 49–71 (2005).
5. N. J. Stroud, Polarization and partisan selective exposure. *J. Commun.* **60**, 556–576 (2010).
6. S. de Marchi, B. Stewart, "Wrestling with complexity in computational social science: Theory, estimation and representation" in *The SAGE Handbook of Research Methods in Political Science and International Relations*, L. Curini, R. Franzese, Eds. (SAGE Publications, 2020), chap. 17.