The Structural Turn in Polarization Research: Why Individual-Level Interventions Do Not Suffice

Abstract

The dominant scholarly understanding of political polarization in the United States has shifted three times in three decades: from a focus on elite ideological distance, to mass affective polarization, and most recently to a structural account in which institutional and economic environments shape both the strength of partisan animus and the political incentives to mobilize it. This essay reconstructs that shift through four bodies of evidence: Kleinfeld's (2023) synthesis of a decade of polarization research; Boxell, Gentzkow, and Shapiro's (2021) cross-national measurement of affective polarization in twelve OECD democracies; Voelkel et al.'s (2023) experimental finding that interventions reducing affective polarization do not reliably reduce antidemocratic attitudes; and Stewart, McCarty, and Bryson's (2020) formal model of polarization as an evolutionarily stable response to economic decline and rising inequality. Read together, these findings make three claims defensible. First, the cross-country variation in affective polarization trends rules out monocausal explanations rooted in technology or inequality and points instead toward features of national political structure. Second, depolarization interventions targeting attitudes can succeed at changing feelings without changing the antidemocratic behaviors those feelings were supposed to cause. Third, polarization once established exhibits hysteresis—it persists after the conditions that produced it have reversed—which means the policy frame must shift from "reduce dislike" to "alter the incentive structure." I close by arguing that this configuration of properties is what Rittel and Webber (1973) identified as a wicked problem, and that the methodological consequences of accepting that classification are non-trivial.

https://doi.org/10.33774/apsa-2026-5qmzx **ORCID:** https://orcid.org/0009-0005-2942-0839 Content not peer-reviewed by APSA. **License:** All Rights Reserved

1. Introduction

Two facts that ought to be related appear not to be. Affective polarization—the gap between how warmly Americans rate their own party versus the opposing party on a feeling thermometer—has risen substantially in the United States over four decades. By Boxell, Gentzkow, and Shapiro's (2021) measurement, the average partisan rated in-party members 27.4 points higher than out-party members in 1978; by 2020 the difference was 56.3, an increase of more than a full standard deviation in the 1978 distribution. Over the same period, however, the United States is not noticeably more affectively polarized than several European democracies whose democratic institutions are not under comparable strain. The implication is that affective polarization, on its own, is doing less explanatory work than two decades of research assumed.

This essay reconstructs how the field arrived at that conclusion, and why the implications for both research and intervention are larger than the field has yet absorbed. The argument proceeds through four bodies of recent evidence and one older theoretical frame. Section 2 follows Kleinfeld's (2023) periodization of the field into three generations of understanding. Section 3 uses Boxell, Gentzkow, and Shapiro's (2021) cross-national data to show why several intuitive explanations fail and which one survives. Section 4 examines the experimental finding, established most directly by Voelkel et al. (2023) and synthesized by Kleinfeld, that interventions reducing affective polarization do not reliably reduce antidemocratic attitudes. Section 5 introduces Stewart, McCarty, and Bryson's (2020) formal model, which provides a mechanism for both the irreversibility of polarization and its emergence under economic stress. Section 6 returns to Rittel and Webber (1973) to argue that the configuration of properties revealed by sections 3 through 5—essential uniqueness, irreversibility, intractability to first-generation problem-solving—is precisely what their original treatment of wicked problems described, and that the methodological consequences of that classification have not been fully drawn out.

A note on scope. I take affective polarization, antidemocratic attitudes, and support for political violence as conceptually distinct, following the operationalization the cited literature has

https://doi.org/10.33774/apsa-2026-5qmzx **ORCID:** https://orcid.org/0009-0005-2942-0839 Content not peer-reviewed by APSA. **License:** All Rights Reserved

converged on. The essay does not extend the empirical record. Its contribution is interpretive: to read these particular findings against one another and to make explicit a structural framing of the polarization problem that the empirical literature increasingly supports but rarely states in those terms.

2. Three Generations of Understanding

Kleinfeld (2023) organizes the polarization literature into three generations distinguished less by their findings than by what they take the object of study to be.

The **first generation**, which dominated through the early 2000s, treated polarization as ideological distance between policy preferences and located its source in Congress. Members of Congress had measurably moved apart on roll-call votes since the mid-1950s in the Senate and the mid-1970s in the House, and that movement was asymmetric—Republican members moved rightward more than Democratic members moved leftward. Three findings of this generation are worth holding in mind for what follows. First, ideological polarization at the congressional level developed in a steady, unpunctuated manner over decades, ruling out event-driven explanations such as the Clinton impeachment or the Tea Party caucus. Second, polarization was equally present in the seven states with a single House representative—districts that cannot be gerrymandered—suggesting that gerrymandering is at best a modest cause. Third, and most consequential for what follows, the supply of candidates rather than the demand of voters appears to be the structural locus: more ideologically extreme politicians have been running for office since the 1980s, and party chairs at the county level prefer extreme candidates over moderates by a ten-to-one margin among Republicans and two-to-one among Democrats.

The **second generation**, taking shape around 2016, redirected attention to mass affective polarization—the dislike between members of opposing parties, measured independently of policy disagreement. Several findings of this generation are by now well established. Americans hold

https://doi.org/10.33774/apsa-2026-5qmzx **ORCID:** https://orcid.org/0009-0005-2942-0839 Content not peer-reviewed by APSA. **License:** All Rights Reserved

systematic misperceptions about the demographic composition of the other party and about the policy beliefs of its members; correcting those misperceptions reduces partisan animosity in lab settings. Affective polarization in the United States grew most quickly among Americans over sixty-five years old, a demographic that uses the internet less but watches television and listens to talk radio far more than younger Americans. The pattern is symmetric across parties—Republicans and Democrats dislike each other to roughly equal degrees. And the rise of affective polarization correlates more closely with the spread of cable news and talk radio than with the diffusion of the internet.

The **third generation** is the one that this essay sits inside. It is characterized by two moves. The first is the observation, developed in section 4 below, that the second generation's intervention research—reduce affective polarization in the lab, expect democratic gains downstream—has not delivered. The second is a reorientation toward the political structures and incentives that allow politicians to instrumentalize affective polarization as a winning strategy, rather than treating polarization as a property of individuals. Jennifer McCoy and Murat Somer's concept of "pernicious polarization"—polarization understood as a process in which a variety of identities collapse into two opposing camps that view each other as immoral, instrumentalized by political leaders—belongs to this generation, as does the broader comparative-politics and civil-war literature on which it draws. Kleinfeld documents that this third-generation framing is more prevalent among scholars who compare political systems and study civil wars than among U.S.-focused democracy scholars, and that it is only beginning to spread to the latter.

The three generations are not refutations of one another. Congress remains very ideologically polarized; mass affective polarization is real and rising; what has changed is the field's understanding of which of these features is doing what work in the larger system. The remainder of this essay is an argument that the third-generation framing has stronger evidentiary support than its current uptake reflects.

https://doi.org/10.33774/apsa-2026-5qmzx **ORCID:** https://orcid.org/0009-0005-2942-0839 Content not peer-reviewed by APSA. **License:** All Rights Reserved

3. Cross-National Variation Constrains the Causal Story

The most useful single piece of evidence for distinguishing among explanations of rising affective polarization comes from comparison across countries. If affective polarization has risen for the same reason everywhere, then candidate causes should rise alongside it everywhere. If it has not, candidate causes that do rise everywhere are unlikely to be the explanation.

Boxell, Gentzkow, and Shapiro (2021) construct comparable measures of affective polarization across twelve OECD democracies over roughly four decades, drawing on 149 surveys. The headline finding is that the trends do not move together. Over the sample period, affective polarization rose substantially in the United States (5.6 points per decade), rose more modestly in five other countries (Switzerland, France, Denmark, Canada, and New Zealand), and *fell* in the remaining six (Japan, Australia, Britain, Norway, Sweden, and Germany). The U.S. exhibited the largest increase among all sample countries; the cross-country trends differ from the U.S. trend significantly except in the case of Switzerland.

This pattern is what statisticians and methodologists call a useful constraint: it eliminates explanations that would predict uniform trends across the sample. Boxell, Gentzkow, and Shapiro relate the trends in affective polarization to trends in a series of candidate explanatory variables and report the rank correlations. The results are decisive enough to be worth listing.

- **Inequality (Gini coefficient):** rank correlation with the trend in affective polarization is *negative* and not statistically significant. Inequality rose in most sample countries; affective polarization did not.
- **Trade openness (trade share of GDP):** rank correlation is essentially zero and not significant. Trade openness rose in nearly all sample countries.
- **Internet penetration:** rank correlation is *negative* and not significant. Internet penetration rose in every sample country with available data.
- **Share getting news online:** rank correlation is *negative* and not significant.

https://doi.org/10.33774/apsa-2026-5qmzx **ORCID:** https://orcid.org/0009-0005-2942-0839 Content not peer-reviewed by APSA. **License:** All Rights Reserved

- **24-hour private television news, both penetration and count:** rank correlation is positive but not significant.
- **Foreign-born share:** rank correlation is negative and not significant.
- **Non-white share of the population:** rank correlation is positive, with a p-value of 0.052.
- **Ethnic fractionalization and ethnic polarization:** rank correlations are positive but not significant.
- **Partisan–ideological sorting:** rank correlation is positive but not significant.
- **Elite polarization:** rank correlation is 0.782, with a p-value of 0.011.

The single variable whose trend tracks the trend in affective polarization across the twelve democracies, by a comfortable margin, is *elite polarization*—the ideological distance between political parties themselves, rated by experts. This is striking on two counts. First, several of the most-discussed explanations of American polarization—digital media, social media, rising inequality, immigration—turn out not to track affective polarization once we look beyond the United States. Second, the variable that does track it is itself a structural feature of national politics, not a feature of individuals or a feature of the media environment in the consumer-technology sense.

Two clarifications are in order. Boxell, Gentzkow, and Shapiro are careful to note that cross-country differences in survey format, political systems, and other factors make these comparisons challenging, and that the cross-country data make it difficult to reach firm causal conclusions about any single variable. They explicitly note that they have not measured all candidate explanations—social sorting in particular is hard to harmonize across countries. The argument here is not that elite polarization causes affective polarization in a unidirectional sense; it is the more modest claim that the cross-national variation rules out a class of explanations and pushes the remaining candidates toward features of national political structure. That is sufficient for what follows.

The implication for the second-generation research program is uncomfortable. If affective polarization in the United States is rising primarily because of features of the U.S. political system

https://doi.org/10.33774/apsa-2026-5qmzx **ORCID:** https://orcid.org/0009-0005-2942-0839 Content not peer-reviewed by APSA. **License:** All Rights Reserved

rather than because of features of human psychology under digital-age conditions, then interventions that target the latter without altering the former should be expected to underperform. That is exactly what the third-generation experimental literature finds.

4. The Intervention–Democracy Disconnect

The strongest single piece of evidence motivating the third-generation framing is also the most counterintuitive. Interventions that reduce affective polarization do not reliably reduce the antidemocratic attitudes and behaviors that affective polarization was, in the second-generation theoretical framework, presumed to cause.

Voelkel et al. (2023), publishing in *Nature Human Behaviour*, test three previously established depolarization interventions—correcting misperceptions about out-partisans, priming inter-partisan friendships, and observing warm cross-partisan interactions between political leaders—across experimental samples totaling 8,385 participants. The interventions reliably reduce affective polarization, replicating prior findings. But the authors do not find compelling evidence that the same interventions reduce support for undemocratic candidates, support for partisan violence, or the prioritization of partisan ends over democratic means. The conclusion drawn in the abstract is precise and worth quoting in full: the findings "call into question the previously assumed causal effect of affective polarization on anti-democratic attitudes."

The implication is striking when laid alongside the broader experimental literature that Kleinfeld (2023) synthesizes. The Strengthening Democracy Challenge tested twenty-five short interventions, selected from 252 crowdsourced proposals, on more than 30,000 online participants. Several interventions reduced affective polarization. The same interventions did not reduce antidemocratic attitudes or justifications for political violence. One intervention—based on creating a shared identity as an "exhausted majority"—actually *increased* support for antidemocratic practices. A video depicting social instability and violence in collapsing

https://doi.org/10.33774/apsa-2026-5qmzx **ORCID:** https://orcid.org/0009-0005-2942-0839 Content not peer-reviewed by APSA. **License:** All Rights Reserved

democracies reduced affective polarization but *increased* conservative Republican support for political violence. The pattern is consistent across studies: affective polarization is a real and measurable construct, and it can be moved in the lab, but it is not the variable doing the explanatory work the second-generation framework assigned to it.

Kleinfeld synthesizes the implication as follows. The only interventions that appeared, in the Strengthening Democracy Challenge, to reduce both affective polarization and antidemocratic attitudes were those that addressed a specific belief: the perception that the other party intends to break democratic norms. A video of Utah's gubernatorial candidates committing to honor the 2020 election results, for example, reduced participants' support for their own side taking antidemocratic action. The distinguishing feature of these interventions is not that they reduced animus toward members of the other party. It is that they targeted the variable that actually predicts antidemocratic behavior: the perception of threat to democracy from the other side.

This pattern admits a coherent interpretation, which Kleinfeld articulates and which the comparative-democracy literature has been suggesting for some time. Affective polarization in itself does not directly cause antidemocratic action; what causes antidemocratic action is the belief that the other side is about to break democratic norms, which makes preemptive norm-breaking on one's own side appear justified. The two variables are correlated—the more one dislikes the other party, the easier it is to believe they are bent on democratic subversion—but the causal chain runs through belief about threat, not through dislike itself.

If this is the right picture, then the second-generation research program's central operational assumption—that reducing affective polarization is a road to democratic stabilization—is incorrect, and the experimental literature has begun to verify that it is incorrect.

https://doi.org/10.33774/apsa-2026-5qmzx **ORCID:** https://orcid.org/0009-0005-2942-0839 Content not peer-reviewed by APSA. **License:** All Rights Reserved

5. A Formal Mechanism: Hysteresis, Social Dilemma, and the Limits of Reversal

The cross-national constraint developed in section 3 and the experimental finding developed in section 4 leave a question open. Why should polarization, once established, be so resistant to interventions that target it directly? Stewart, McCarty, and Bryson (2020), publishing in *Science Advances*, provide a formal model that supplies one mechanism.

The setup is a stylized cultural-evolution model in which individuals in a population face repeated choices between in-group and out-group interactions. Out-group interactions have higher expected reward but also higher risk; in-group interactions have lower reward but lower risk. Individual fitness depends both on accumulated interaction benefits and on the underlying economic environment, with a non-linear (sigmoidal) component capturing the idea that falling below some threshold—poverty, insolvency—produces a sharp loss of utility. Individuals copy the strategies of more successful neighbors with a probability that depends on the fitness gap.

The model produces three results that matter for the present argument.

First, under conditions of economic decline or rising inequality, a risk-averse, in-group-favoring strategy becomes evolutionarily stable even when out-group interactions retain higher expected benefit. The mechanism is straightforward: when the threshold below which utility collapses comes into play, the variance reduction from in-group interaction matters more than the expected-value advantage of out-group interaction. The authors describe this as a transition in the optimal behavioral strategy from out-group toward in-group interactions—that is, toward greater polarization.

Second, and this is the result that is most consequential for intervention design, the model is bistable in the social version where out-group interaction success depends on the willingness of the other party to engage. Once a population has shifted to the high-polarization state, the low-polarization state is hard to recover via gradual change even when economic conditions improve. The authors describe this directly: “a population that becomes polarized can remain trapped in that

https://doi.org/10.33774/apsa-2026-5qmzx **ORCID:** https://orcid.org/0009-0005-2942-0839 Content not peer-reviewed by APSA. **License:** All Rights Reserved

state, even after a reversal of the conditions which generated the risk aversion and polarization in the first place." Polarization, in this model, exhibits hysteresis.

Third, polarization in the bistable regime "takes on the character of a social dilemma, in which there would be a collective benefit if everyone switched to a low polarization strategy, but each individual has incentives to maintain a high polarization strategy." The conclusion the authors draw is methodologically consequential: "Recovering low polarization would therefore require coordination, possibly introduced in response to an exogenous shock altering salience of the polarized identities relative to the identity of the full population as an aggregate."

The empirical companion to the model is more modest but in the right direction. State-level Gini coefficients and state-level affective polarization, pooled across the 2008, 2012, and 2016 presidential cycles, exhibit a positive and statistically significant correlation ($p<0.01$) in the U.S. data. This is consistent with the model's prediction—rising inequality and economic stress can produce a transition toward risk-averse in-group strategies that manifest as affective polarization—but the authors are clear that the empirical evidence is suggestive rather than dispositive, and that the model's primary contribution is theoretical.

Three implications follow for the broader argument of this essay.

The first is reconciliatory: the model "unifies what are often seen as conflicting explanations for political polarization: identity threat versus economic anxiety." In the framework, economic stress and identity-threat dynamics are not competing accounts but coupled mechanisms—economic stress lowers the threshold below which in-group preference becomes individually rational, and identity-based grouping is the dimension along which the in-group/out-group line is drawn.

The second is structural: the bistability result implies that an intervention which marginally reduces affective polarization in a population stuck in the high-polarization basin of attraction will not, by itself, tip the system back. The basin is stable to small perturbations. This is consistent with the experimental finding from section 4: interventions that move affective polarization a measurable amount in the lab do not move the larger democratic equilibrium.

https://doi.org/10.33774/apsa-2026-5qmzx **ORCID:** https://orcid.org/0009-0005-2942-0839 Content not peer-reviewed by APSA. **License:** All Rights Reserved

The third concerns the kind of intervention that the model implies could work. The authors point toward exogenous shocks that re-establish the salience of a larger, encompassing identity—their examples include wartime altruism and post-disaster cohesion. The relevant point for present purposes is not the specifics of those examples but that the implied intervention class is structural rather than psychological. The unit of change is the salience of collective identity at the population scale, not the temperature of dislike at the individual scale.

This is precisely where the comparative-democracy literature has been pointing as well. Kleinfeld (2023) notes that majoritarian, winner-take-all political systems appear to favor affective polarization by creating a binary of political actors, and that this is one reason a class of structural reforms—proportional representation, nonpartisan primaries, ranked-choice voting—has gained renewed attention in the third-generation literature. The empirical evidence on these reforms is uneven, and I do not want to overstate it. But the direction is consistent: where the second-generation literature looked for intervention at the level of mass attitudes, the third-generation literature increasingly looks for intervention at the level of political-structural incentives.

6. Polarization as a Wicked Problem

The previous three sections have a shared structure that may not be immediately apparent. Each documents a feature of the polarization problem that systematically frustrates the application of standard problem-solving methods: causal heterogeneity across cases (section 3), decoupling between treatment and outcome (section 4), and bistability with hysteresis (section 5). The configuration is recognizable. It is what Rittel and Webber (1973), in a paper now half a century old, classified as a *wicked problem*.

Rittel and Webber's argument is methodological rather than substantive. They distinguished a class of social-policy problems—including, in their examples, crime, education, housing, and urban planning—from the "tame" problems that the scientific and engineering paradigms of the

https://doi.org/10.33774/apsa-2026-5qmzx **ORCID:** https://orcid.org/0009-0005-2942-0839 Content not peer-reviewed by APSA. **License:** All Rights Reserved

early twentieth century had been built to solve. The distinction is not that wicked problems are harder. It is that they have ten properties that make the apparatus of definition–analysis–solution inapplicable to them in principle.

Three of those properties bear directly on the present argument.

Their first property: a wicked problem has no definitive formulation, because the information one needs to understand it depends on the solution one is considering. “To find the problem is thus the same thing as finding the solution”; the problem cannot be defined until the solution is found. This is precisely what the three-generation history in section 2 illustrates. The first generation’s definition of polarization as ideological distance and the second generation’s definition as affective dislike are not competing answers to the same question. They are different problems entailed by different intervention frames. The third generation’s reframing as a structural-incentive problem is, in Rittel and Webber’s terms, a new problem-formulation made visible by the inadequacy of the previous solution-formulations.

Their fifth property: every implemented solution to a wicked problem is “a one-shot operation,” consequential, and effectively irreversible. The Stewart, McCarty, and Bryson (2020) result that polarization, once established, exhibits hysteresis is a formal restatement of this property in a specific domain. The implication is the same in both treatments: corrective interventions cannot be evaluated cheaply by trial and error, because every trial leaves traces that cannot be undone, and the trial itself alters the conditions under which the next trial will be conducted.

Their eighth property: every wicked problem can be considered a symptom of another problem, and the level at which the problem is settled is not given by logic but by judgment. The cross-national finding in section 3—that affective polarization is correlated most strongly with elite polarization across the OECD—is a working example. Elite polarization is itself produced by structural features of party systems, candidate recruitment, primary rules, campaign finance, and media markets. Settling the polarization problem at the level of mass affect treats a symptom; settling it at the level of political-structural incentives treats one layer deeper, but is itself a symptom of decisions made about constitutional design and election law.

https://doi.org/10.33774/apsa-2026-5qmzx **ORCID:** https://orcid.org/0009-0005-2942-0839 Content not peer-reviewed by APSA. **License:** All Rights Reserved

Rittel and Webber argue, prescriptively, that the systems-approach of the "first generation"—the linear pipeline of understand, gather, analyze, synthesize, solve—is inadequate for wicked problems. They sketch an alternative, the "second generation" approach, in which planning is treated as an argumentative process where the image of the problem and of the solution emerges gradually among the participants, as a product of incessant judgment subjected to critical argument. This is a less satisfying methodological proposal than its first-generation alternative. It does not promise the linear convergence to optimal solutions that the systems-approach promises. But it has the modest virtue of being a method that can actually operate on problems with the properties Rittel and Webber describe.

Two consequences of accepting this classification matter for the polarization literature.

The first is epistemic humility about the units of evidence that the field has been generating. Self-reported surveys based on short online interventions, trust games, and other lab tests—the workhorses of the second-generation depolarization literature—are precisely the kind of evidence the wicked-problems frame predicts will under-deliver. Not because the experiments are badly conducted but because the construct they measure ("change in affective polarization at the individual level immediately after an intervention") is not on the causal pathway the field cares about ("change in the political-structural conditions that make antidemocratic strategies pay"). Kleinfeld (2023) draws this conclusion directly: claims of success or failure from such interventions should be taken with a grain of salt, and the field has been drawing "overly broad, even hyperbolic, conclusions in both positive and negative directions from short-term lab experiments."

The second consequence is harder. Rittel and Webber's tenth property—the planner has no right to be wrong—imposes a different standard of accountability than the scientific one. In the scientific case, a refuted hypothesis is a contribution. In the planning case, a wrong intervention is a harm done to specific people. The third-generation literature is beginning to grapple with this. Kleinfeld observes that some interventions intended to mobilize voters by emphasizing threats to democracy may be *enhancing* the very perceptions of threat that have been shown experimentally to *increase*

https://doi.org/10.33774/apsa-2026-5qmzx **ORCID:** https://orcid.org/0009-0005-2942-0839 Content not peer-reviewed by APSA. **License:** All Rights Reserved

support for antidemocratic action on one's own side. This is the wicked-problem signature: well-intended interventions producing the harms they were intended to prevent, because the system they operate on has dynamics that the intervention frame did not anticipate.

7. Conclusion: Where the Argument Leaves the Reader

Let me state what this essay has and has not done.

It has not produced new empirical evidence. The four pieces of empirical work it draws on—Kleinfeld's synthesis, Boxell-Gentzkow-Shapiro's cross-national measurement, Voelkel et al.'s experimental study, and Stewart-McCarty-Bryson's formal model—were produced between 2020 and 2023 and have been read together in this essay through the methodological lens of Rittel and Webber. The contribution, if there is one, is interpretive: to make explicit a structural framing of the polarization problem that the empirical literature increasingly supports.

It has argued that the cross-national pattern of affective polarization trends is inconsistent with explanations based on inequality, trade openness, or digital media, and is most strongly correlated with elite polarization, a structural feature of national politics. It has argued that experimental interventions that reduce affective polarization do not reliably reduce antidemocratic attitudes, and that this decoupling is precisely what one would expect if the second-generation theoretical framework had misidentified the causal variable. It has argued that a formal model of polarization as an evolutionarily stable response to economic stress produces bistability and hysteresis, which together explain why marginal interventions in the high-polarization basin do not return the system to the low-polarization basin, and why coordinated, structural interventions are the kind of intervention the dynamics imply will work. And it has argued that the configuration of properties revealed by these results matches Rittel and Webber's original treatment of wicked problems closely enough that the methodological apparatus they developed for that class of problems is the appropriate apparatus for this one.

https://doi.org/10.33774/apsa-2026-5qmzx **ORCID:** https://orcid.org/0009-0005-2942-0839 Content not peer-reviewed by APSA. **License:** All Rights Reserved

What this essay has not argued, and could not argue from the sources it relies on, is that any particular structural reform—proportional representation, nonpartisan primaries, ranked-choice voting, reform of campaign finance, restoration of trusted local media—will work. The case for those reforms is being made elsewhere, and the empirical evidence for any specific one of them is uneven. The case the present essay makes is the prior case: that the level at which solutions should be sought has shifted, and that the standards of evidence appropriate to that level are different from the standards the second-generation depolarization literature has been generating.

One implication may be worth stating directly. The polarization problem has been, in popular and professional discourse, framed as a problem about how Americans feel about one another. The literature reviewed here suggests it is not primarily that. It is a problem about the political-structural conditions under which polarized feeling can be instrumentalized to win elections, about the institutional incentives that make extremist candidate-recruitment a rational strategy, about the bistable dynamics that make the polarized equilibrium hard to escape, and about the cross-national variation that makes a domestic-structural explanation more credible than a universal-psychological one. If that reframing is right, then both the diagnostic and prescriptive vocabularies of the field need revision, and the methodological standards for interventions need to come up to the level of the wicked-problem class the underlying object actually belongs to.

A final caveat. The empirical literature is moving fast. The claims advanced here are calibrated to the state of evidence in the sources cited and may need revision as the field continues to develop. The methodological frame—that polarization is a wicked problem in Rittel and Webber's specific sense—is, I think, more durable. It would survive most plausible developments in the empirical record because it is a claim about the *kind* of problem polarization is, not a claim about its specific empirical content.

https://doi.org/10.33774/apsa-2026-5qmzx **ORCID:** https://orcid.org/0009-0005-2942-0839 Content not peer-reviewed by APSA. **License:** All Rights Reserved

References

Boxell, Levi, Matthew Gentzkow, and Jesse M. Shapiro. 2021. “Cross-Country Trends in Affective Polarization.” NBER Working Paper 26669 (August 2021 revision). Subsequently published in *Review of Economics and Statistics* 106 (2024): 557–565.

Kleinfeld, Rachel. 2023. *Polarization, Democracy, and Political Violence in the United States: What the Research Says*. Washington, DC: Carnegie Endowment for International Peace.

Rittel, Horst W. J., and Melvin M. Webber. 1973. “Dilemmas in a General Theory of Planning.” *Policy Sciences* 4 (2): 155–169.

Stewart, Alexander J., Nolan McCarty, and Joanna J. Bryson. 2020. “Polarization Under Rising Inequality and Economic Decline.” *Science Advances* 6 (50): eabd4201.

Voelkel, Jan G., James Chu, Michael N. Stagnaro, Joseph S. Mernyk, Chrystal Redekopp, Sophia L. Pink, James N. Druckman, David G. Rand, and Robb Willer. 2023. “Interventions Reducing Affective Polarization Do Not Necessarily Improve Anti-Democratic Attitudes.” *Nature Human Behaviour* 7 (1): 55–64.

https://doi.org/10.33774/apsa-2026-5qmzx **ORCID:** https://orcid.org/0009-0005-2942-0839 Content not peer-reviewed by APSA. **License:** All Rights Reserved