Does Fear of Retaliation Constrain Support for Democratic Backsliding?*

Daniel B. Markovits†

October 1, 2025

Abstract

To what extent can fear of partisan retaliation deter support for anti-democratic behavior? In contemporary American politics, many voters worry about the opposing party's commitment to democracy. Such concerns could, in theory, promote compromise if voters believe opponents will break democratic rules only when provoked. I investigate these beliefs among the American public, distinguishing them from other drivers of democratic support. In two prediction experiments (N = 7,000), I find that partisans view the opposing party as only modestly more likely to undermine democracy if provoked than if unprovoked. These expectations fall well below both theoretical maximums and benchmark estimates I obtain from a sample of political elites. Further, references to retaliation are rare in open-ended responses. In follow-up experiments (N = 5,500), I show that randomized warnings about the likelihood of retaliation reduce support for violating democratic rules. I argue that this form of strategic reasoning can promote pro-democratic attitudes even when normative commitments are weak and elites fail to champion democratic principles.

*The most recent version of this paper is available here.

†PhD. Candidate, Department of Political Science, Columbia University. The author thanks the Columbia Political Science Department and the Institute for Humane Studies for funding and Taylor Carlson, Anthony Fowler, Don Green, Matt Graham, Shigeo Hirano, Mohamed Hussein, Yphtach Lelkes, Neil Malhotra, Tamar Mitts, Brendan Nyhan and Carlo Prato for comments as well as participants at the Montreal Summer School on Democratic Decline and Resilience (2024), MPSA (2025), the Polarization Research Lab Annual Meeting (2025), the Institute for Humane Studies Junior Fellows convening (2025), and the Columbia American Politics Speaker Series (2025). The original studies in this project were approved by the Columbia Institutional Review Board, protocol numbers AAAV2044, AAAV3443, and AAAV3445.

1 Introduction

American democracy is engulfed today in spiraling violations of democratic norms. In areas as diverse as gerrymandering and judicial nominations, partisans of both sides justify their own transgressions with reference to provocations from the opposing party. When the newly re-elected Donald Trump targeted his political opponents with Department of Justice investigations, he cited his prosecution by the Biden administration. When both parties seek to redraw congressional maps to their benefit, they reference identical past violations from the other party. This form of retaliation is well-established in academic literature (Bateman, 2025; Lupu et al., 2025; Janssen et al., 2025), frequently invoked by partisan actors, and often highlighted by media outlets seeking to caution political allies. The Wall Street Journal warned that conservatives "might not like (Trump's precedent) when President Ocasio-Cortez is in charge" (WSJ Editorial Board, 2025) while the Washington Post warned Democrats enthusiastic about the prosecution of Trump that "legal escalations beget legal escalations" and that such maneuvers "further inflame (Trump) against his opponents" (Willick, 2025).

The prospect of aggression from one party prompting retaliation from the other recalls a simple logic of deterrence which is common to studies of conflict. Across diverse contexts, strategic consideration of how an opponent will respond shapes a first-mover's willingness to seek conflict (Keohane, 1984; Fearon and Laitin, 1996; Weingast, 1997). In game-theoretic or qualitative accounts of conflict over democratic rules, concerns that opponents will retaliate for escalations are a major constraint on supporting otherwise attractive efforts to revise or overturn the rules of the game. This mechanism holds even when idealistic commitment to democracy is weak and there are substantial material interests at stake (North and Weingast, 1989; Weingast, 1997; Helmke et al., 2022). Existing theories suggest that knowledge of the cycle of escalation can help to deter rather than provoke attacks on democratic rules.

When politicians engage in incremental anti-democratic behaviors, voters have ample opportunity to remove these leaders from positions of power, and scholars have long highlighted the mass public as the ultimate check on anti-democratic politicians (Graham and Svolik, 2020). Yet despite a burgeoning literature on public opinion and democracy, there is only limited evidence on the role strategic considerations play in voter support for democratic backsliding, with the narrow though important exception of papers exploring how partisans perceive the opinions of their opponents (Druckman et al., 2023; Dias et al., 2024). This literature suggests that negative information about opponents increases willingness to violate democratic rules. However, we do not know from existing scholarship if or when more complex reasoning is possible.

This gap matters for two reasons. First, the relevance and magnitude of retaliation expectations may partially explain the status quo of public opinion regarding democracy. Existing research has clearly established that Americans support democratic values in the abstract and condemn political violence in overwhelming majorities (Westwood et al., 2022; Holliday et al., 2024). At the same time, most voters will not punish politicians of their party at the ballot box when they violate democratic rules (Graham and Svolik, 2020). This could reflect an equilibrium in which fear of opponents constrains partisan voters from backing yet more aggressive actions by their own party's politicians; deterrence could be reducing support for anti-democratic behavior compared with a counterfactual where citizens do not fear the reactions of opponents. Second, whatever the baseline may be, changing beliefs about retaliation from opponents could be an important lever for reducing support for anti-democratic behavior compared to the status quo. Notably, this deterrence-based approach offers a clear alternative to existing efforts to reduce anti-democratic attitudes which emphasize the role of norms, common identities, and reassuring messaging about the opposing party (Levendusky, 2023; Voelkel et al., 2024; Weiss et al., 2025).

Scholars have long highlighted the role of the mass public as a check on procedural

escalation and democratic erosion. However, the importance of public opinion on these issues has become more directly relevant as fights over gerrymandering and efforts to reform the U.S. electoral system have increasingly placed the democratic rules of the game directly on the ballot through referenda on topics as diverse as ranked-choice voting and re-drawing state congressional maps.[1]

The relative dearth of research on voters' strategic reasoning creates ambiguity in both these areas. In addition, because there is minimal empirical evidence about whether American political elites actually reason in the manner prescribed by formalized accounts, evidence regarding public opinion is needed to provide micro-foundations for many stylized examples of deterrence. No matter how elites may reason about risks from the other party, sanctioning from voters in primaries or general elections can shape how politicians behave regarding democratic and procedural norms (Bartels and Carnes, 2023; Malzahn and Hall, 2024). Cooperative equilibria at the elite level can be undone by voters who seek escalation.

This paper proceeds through several stages. I begin by outlining how theoretical accounts of deterrence among elites might map onto a form of parsimonious strategic reasoning among the mass public. I describe how this logic differs from existing approaches, including an important strand of research on second-order beliefs (i.e. partisans' beliefs about the preferences of members of the opposing party). I highlight low weights on retaliation from the other party and low perceived probabilities of retaliation as plausible mechanisms that might undermine deterrence and explain what types of assumptions might contribute to these outcomes. Next, I focus on answering my first, descriptive question by demonstrating that Americans believe that democratic violations carry tangible consequences for their own party. I contextualize these findings by comparing them to a pair of benchmarks: (1) The theoretical maximum retaliation prediction and (2) Benchmarks I obtain from numeric predictions embedded in interviews with 52 partisan political elites. I then use text

[1]See California's ballot measure:https://calmatters.org/politics/2025/08/california-redistricting-vote/ and the many initiatives to implement ranked-choice voting via ballot measure:https://fairvote.org/press/fact-sheet-rcv-electionday-2024/

data from open-ended survey responses from partisan samples and comments on partisan YouTube channels to document the inconsistent prominence of retaliation concerns. Next, I show results from a second set of experiments demonstrating that explicitly prompting the conditionality of the opposing party's anti-democratic behavior reduces public support for anti-democratic actions from a respondent's own party.

My results offer cause for both concern and optimism. Before prompting or educational treatments, voters believe that opponents are only modestly likely to retaliate. However, as I show in my final set of experiments, explicit warnings about the risks of retaliation reduce support for partisan efforts to revise the rules of the game.

Throughout, I make three contributions. First, I highlight the theoretical importance of retaliation expectations in American politics. I distinguish these from second-order beliefs, which are partisans' views about opposing voters, and from unconditional predictions that the other party will violate democratic norms. Retaliation predictions are unique because they concern the real-world actions of the opposing party rather than voter preferences, and they reflect a conditional logic that requires partisans to anticipate opponents' behavior in response to their own party's provocations. This distinction clarifies the assumptions that underlie prior accounts of voter support for democratic rule changes. Second, I provide novel empirical evidence on the scale of these expectations across voters and elites. Across survey samples and open-ended responses, retaliation predictions from survey respondents are modest. They are lower than elite benchmarks and far from the theoretical maximum, yet they do not show common failures such as ceiling effects or sharply diminishing returns. Third, I show that randomized warnings about retaliation substantially alter respondents' expectations of opponent behavior and reduce support for potentially provocative changes to democratic rules.

Throughout, I argue that voters can exercise a simple form of strategic intuition regarding their party's efforts to revise the rules of the game, though this logic is subject to

important scope conditions. Partisans sometimes refrain from supporting otherwise attractive democratic transgressions when they understand that the downstream consequences of such behaviors can include retaliation directed at their party.

2 Why Do Voters Support Democracy?

This project addresses a disconnect between how existing literature conceptualizes elite versus mass support for democratic norms.[2] Formal models of elite decision-making frequently invoke the "shadow of the future," suggesting that the threat of retaliation deters elites from breaking democratic rules, as in older political economy accounts where the possibility of revolution or coup prompts compromise or policy change (Acemoglu, 2001; Acemoglu and Robinson, 2005). More recent formalized accounts of democratic erosion make similar arguments, with the threat of the opposing party responding aggressively serving as a common feature of many models of iterated competition over democratic rules (Miller, 2021; Helmke et al., 2022; Grillo et al., 2023)

This logic extends beyond conflict over democratic rules. For example, Fearon and Laitin (1996) describe an "in-group policing" equilibrium, where members of a group fear collective punishment if one member violates a norm. In international relations, scholars examine how fear of retaliation constrains aggression (Keohane, 1984), though other work highlights how anarchy may incentivize preemptive strikes (Jervis, 1978). Perhaps most directly relevant, Pauly (2024) highlights the "assurance dilemma," emphasizing that threats must be paired with reassurances that cooperation will not be punished in order for deterrence to prove effective.

Studies of public opinion, by contrast, often treat preferences for democratic values as exogenously given. Graham and Svolik (2020), introducing a framework echoed by others

[2]Throughout, I refer to democratic norms to describe the current rules of the game. Some violations of norms, like changing the allocation of electoral votes in a given state, may not by themselves contradict definitions of democracy, but constitute part of broader efforts to rig or distort electoral outcomes.

(Carey et al., 2022; Grillo and Prato, 2023; Frederiksen, 2024), argue that voters weigh a fixed democratic commitment against policy preferences. Other approaches argue that differing definitions of democracy (Wunsch et al., 2022; Davis et al., 2022), motivated reasoning (Krishnarajan, 2023) and efforts by elites to confuse the nature of democratic transgressions (Clayton et al., 2021; Nalepa et al., 2024) make accountability difficult. These findings align with arguments about voter (in)capacity (Achen and Bartels, 2016; Lucas et al., 2024), suggesting that the electorate often fails to recognize or understand the implications of anti-democratic behavior and cannot act as a reliable democratic safeguard.

Recently, scholars have investigated how pessimism about opponents drives support for anti-democratic behavior. These papers show that voters overestimate the share of opposing partisans who support partisan violence and correcting these misperceptions bolsters democratic attitudes (Mernyk et al., 2022; Braley et al., 2023; Freitag et al., 2025).[3] However, these papers assume voters view their opponents' support for democracy as exogenous and unconditional. Voter beliefs about (1) real-world outcomes, as opposed to second-order beliefs, and (2) conditional probabilities of misbehavior remain underexplored in this literature.[4]. Similar work shows that randomized exposure to *provocations* from opponents reduces support for democracy (Lupu et al., 2025; Janssen et al., 2025). Again however, it is unclear what forward-looking mechanisms are implicated in voter decision making.

The divergence between complex accounts of elite behavior and simple accounts of voter reasoning about democracy turns on a testable assumption: do voters hold conditional expectations about the actions of the opposing party, or do they view its commitment to democracy as a fixed parameter? The public matters because, even if elites are deterred by the threat of retaliation [5], voters who ignore deterrence could oust pro-democratic politicians

[3]Other work raises doubts about the durability and robustness of these treatment effects (Dias et al., 2024)

[4]A parallel strand of work credits social norms among in-groups (Valentim, 2024; Dahlum et al., 2024) for constraining norm violations, consistent with the idea that democratic values vary in their pro-sociality depending on context, unlike standard normative behaviors such as voting (Gerber et al., 2008). This reflects inter-personal strategic behavior among co-partisans rather than inter-party strategic interaction.

[5]Notably, there is no quantitative evidence on how American elites consider the trade-offs inherent in

in primaries, undermining elite-led bargains to defend democracy.[6] By investigating beliefs about how opponents react to provocation I explore a novel mechanism rooted in partisan self-interest. The most substantively proximate efforts have a handful of efforts that explore whether fear of violence or partisan embarrassment reduce support for anti-democratic behavior Voelkel et al. (2024); Connors et al. (2025).

My theoretical perspective is similar to work that explores whether citizens can understand the second-order effects of policies. For example, (Bowen et al., 2023) shows that updating about the response of rival states reduces support for nuclear use. More generally, Dal Bó et al. (2018) shows that participants understate the second-order, equilibrium effects of the rules of laboratory games. Work on strategic behavior in electoral contexts often involves similar second-order reasoning, as in the selection of a candidate in the first round of an election who is most able to win a general election (Corbett et al., 2022; Cohen, 2025) or participating in an opposing primary for similar reasons (Markovits and Cohen, 2025; Cohen and Markovits, 2025). More stylized accounts of this style of reasoning in lab games show that the shadow of the future affects behavior (Dal Bó et al., 2021), though subjects may struggle to comprehend formalized games (Koppel et al., 2025).

Given mixed evidence of voter capacity for strategic intuitions, what would a realistic model of citizen strategic reasoning about procedural rules look like? Democratic contestation typically involves complex mechanisms such as assumptions about electoral victories (which may prevent retaliation) or judicial limits on transgressions. As an alternative, I investigate a theoretically simple set of beliefs around retaliation, which I define as how much more likely voters view the opposing party to violate norms if provoked compared to unprovoked.

negotiating democratic rules; a gap I begin to address later. Qualitative accounts also suggest partisan elites in backsliding democracies often fail to anticipate how their behaviors provoke opponents (Levitsky and Ziblatt, 2018; Gamboa, 2022), suggesting even sophisticated actors have difficulty estimating true effects

[6] Accounts of individual cooperation in psychology make similar claims, often focusing on how repeated interactions can induce cooperation even among self-interested subjects (Van Lange et al., 2011)

I define a simple model of public opinion and deterrence where voters consider their payoffs over two periods. A partisan of Party A considers his utility if Party A violates ($D_1 = 1$) or upholds a democratic or procedural norm ($D_1 = 0$). The partisan has beliefs about the actions of the opposing party such that the opposing party violates a democratic norm absent provocation with probability $P_{unprovoked}$. When a democratic norm is violated by the partisan's party, the opposing party violates with an additional probability $P_{retaliate}$. Meanwhile, the partisan places some weight ω on the cost to his or her interests caused by the opposing party's violations. For parsimony, ω incorporates a time discount factor. Notably, this equation isolates the strategic cost faced by a voter and does not consider features like their normative commitment to democracy.

$$U_i(D_1 \in [0,1]) = -\omega_i(P_{\text{unprovoked}} + P_{retaliate}(D_1)) \tag{1}$$

This simple theoretical framework helps identify several potential sources of deterrence failure. First, partisans may believe that the opposing party will always violate democratic norms, such that ($P_{\text{unprovoked}} = 1, ; P_{\text{retaliate}} = 0$); this belief could derive from assumptions that opponents have already been provoked in prior periods. Second, partisans may believe that the opposing party sometimes violates norms but does not do so in response to the actions of co-partisans ($P_{\text{unprovoked}} \neq 1, ; P_{\text{retaliate}} = 0$). Substantively, this belief could stem from expectations about preemption (that is, the idea that violating a democratic norm might mechanically prevent opponents from retaliating) or from the belief that the opposing party's commitment to democracy is exogenous to whether or not it is first provoked. Finally, partisans may acknowledge the possibility of retaliation but place little weight on its costs.

3 Testing Beliefs about Retaliation

I begin by exploring baseline beliefs about retaliation by estimating $P_{retaliate}$. To do so, I conduct a pair of survey experiments and then analyze open-ended text data from a range of sources, including comments from YouTube videos discussing democratic violations. In this section, my goal is to investigate sources of current preferences about democratic norms and the extent to which retaliation fears may be limiting support for democratic backsliding compared with a counterfactual of no such concerns. I investigate the narrow question of status quo retaliation expectations: the extent to which partisans believe that the opposing party will violate democratic or procedural norms in response to their own party's provocations.

3.1 Method

My first two experiments share a common format and many similar design features. Across both studies, respondents estimated the probability that a party would violate democratic norms across scenarios with randomized attributes (five attributes in Experiment 1, three in Experiment 2). Both are analyzed as conjoint experiments with a single profile per task. In each study, subjects judged the likelihood that Party B would break norms in response to randomized behaviors from Party A.

The first experiment, fielded on Cloud Research Connect in August 2024, included 3,626 respondents (1,346 Democrats and 2,280 Republicans, with partisan leaners grouped with partisans). Republicans were oversampled because the study was embedded in a survey on conservative voters' perceptions of the presidential campaign.[7] Participants were asked to estimate the likelihood that the opposing party would (1) have state attorneys general prosecute opponents without evidence, or (2) engage in partisan violence. Each respondent evaluated five hypothetical scenarios describing actions of their own party. The scenarios included five randomized attributes (two policy proposals and three norm-related behaviors),

[7]Pure independents (n=300) completed parallel tasks but are excluded from the main analysis; their predictions were broadly similar to those of partisans.

summarized later in Table 1.

The second experiment, conducted in November 2024, explored retaliation predictions in a more realistic context by randomizing revelations of real violations of democratic norms in the lead-up to the 2024 elections. Respondents made incentivized predictions about Donald Trump's vote share in three swing states (North Carolina, Georgia, and Wisconsin) and then answered questions about Democratic behavior in those states. The sample included 1,450 Democrats, 1,330 Republicans, and 420 pure independents, recruited through Cloud Research Connect. Each state-level vignette presented randomized information about campaign spending, Republican threats against election officials, or Republican-led restrictions on polling places and mail voting. Control conditions omitted mentions of election administration.[8] After each vignette, respondents predicted how Democrats in the state would behave in the year following the election—specifically, whether they would change electoral rules or threaten officials.[9] Baseline randomization levels are summarized in Table 1 and full experimental materials for both studies are provided in the *Experimental Materials* section of the appendix.

Violation	Baseline	Study
Polling Places Closed	Equal Poll Access	Study 1
Partisan Violence	Peaceful Election	Study 1
Politicized Arrests	Fair Justice	Study 1
Extreme Social (Abortion/Immigration)	Moderate Policy	Study 1
Extreme Econ (Tax/Social Security)	Moderate Policy	Study 1
Harris Spending Advantage	Heavily Contested	Study 2
Republicans Closed Polls	No Mention	Study 2
Republicans Threatened Officials	No Mention	Study 2

Table 1: Summary of Violations, Baselines, and Study Assignment

[8]This design choice was intended to increase external validity because media coverage rarely addressed the smooth and fair functioning of electoral institutions.

[9]This experiment included incentivized predictions of Trump's vote share in each state, which are the subject of a companion paper. These prediction results demonstrate that voters considered information about democratic violations credible, as it led to meaningful updating of their predicted election results

Hypotheses

Across the pair of experiments, I preregistered a number of specific hypotheses which are included in the appendix. For parsimony, I summarize my main hypotheses as follows: For Experiment 1, I hypothesized that voters will perceive greater odds their opponents will violate norms in scenarios where they have been provoked (H1A), that they will perceive greater odds their opponents will violate norms in scenarios where they have been provoked compared to scenarios absent provocation and also when they run on extreme compared to moderate social policies (H1B). Predicted retaliation will be smaller for subjects with higher meta-perceptions (H1C) and predicted direct retaliation will be greater than indirect (H1D).[10] For the second experiment, I hypothesized subjects would again predict retaliation (H2A) and that there would be spillovers between state-level vignettes such that subjects exposed to early real-world examples of provocation would predict greater retaliation in later states (H2B).

3.2 Estimation

I estimate average marginal component effects (AMCEs)[11] for all models, using linear regression with predictions about the second-moving party as the dependent variable. All models control for respondents' party, age, race and education level, with standard errors clustered at the respondent level to account for the multiple observations from each respondent (Hainmueller and Hopkins, 2015). Each experiment has two main outcomes corresponding to the two anti-democratic behaviors in each design. The main estimand of interest described in Equation 2 is the perceived retaliatory risk: that is, the probability Party B, the responding party, violates norms when it is provoked compared to when it is not provoked. Unlike in many conjoint designs, these profiles do not feature unrealistic combinations, such that the

[10] In this case, direct retaliation refers to an in-kind response, for example, meeting arrests with arrests. In contrast, indirect retaliation means responding to a democratic violation through a substantively distinct mechanism

[11] Marginal means are reported in the Appendix as an alternate specification

variation in policy is between plausible platforms for each party, and both upholding and violating norms is realistic.[12].

$$\Pr(B \text{ violates} \mid A \text{ violates}) \; - \; \Pr(B \text{ violates} \mid A \text{ upholds}) \tag{2}$$

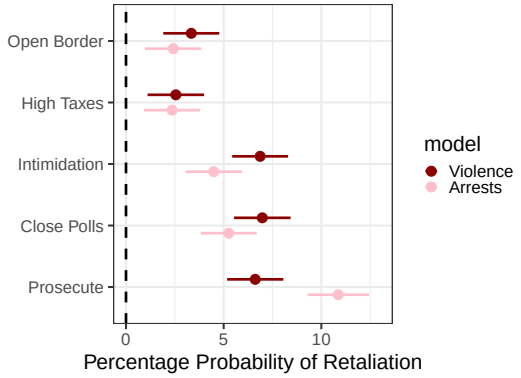
My main preregistered model specifications for the first two experiments are conceptually similar, so are summarized here, with full models in the appendix. All models control for a preregistered vector of covariates and include coefficients for all experimentally manipulated attributes k for individual i and profile j (where k and j $\in [1, 3]$ for Experiment 1 and k and j $\in [1, 5]$). All standard errors across these two experiments are clustered at the individual level to account for the correlation between each subject's multiple responses. For the purposes of displaying results, the baseline level of each attribute is the level that describes fidelity to democratic norms.

3.3 Results

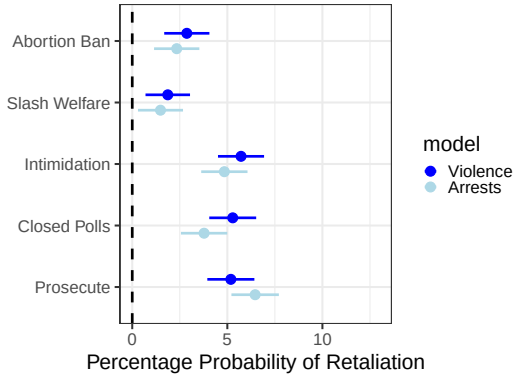
Across both partisan samples, considering violations of democratic norms by co-partisans increased predictions that opposing partisans would violate norms. Figures 1a and 1b report the estimated AMCEs of a given level for each attribute of a given profile , compared to the baseline of no provocation or a moderate policy proposal, as well as the 95% confidence interval for each estimate. These results are separated for Democrats making predictions about the Republican Party and Republicans making predictions about the Democratic Party. Marginal means (Leeper et al., 2020) are presented in Appendix Figure A3.

To give context to these results, the average "baseline" prediction, that is a prediction for cases where no norms were violated, was 30.4% for Democratic predictions of arrests, 37.4% for Democratic predictions of violence, 39.0% for Republican predictions of

[12]The small number of possible scenarios creates the possibility of repeated, identical scenarios in the first experiment, though these are rare and respondents report comparable, though not identical predictions in these cases



(a) Democratic Predictions about Republicans



(b) Republican Predictions about Democrats

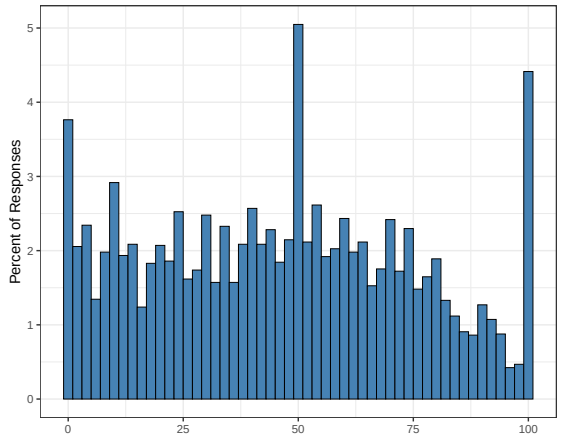
Figure 1: Retaliation predictions across party lines

arrests and 40.9% for Republican predictions of violence. The standard deviation of this baseline outcome is consistently $\approx 30\%$ across party and violation type. The treatment effects of provocations range from $\frac{1}{3}$ to $\frac{1}{10}$ of a standard deviation of the predictions in the "control group" (that is, scenarios with 0 provocations from co-partisans of the respondent). On average, Democrats predict 7 percentage points of retaliation and Republicans predict 5 percentage points. While not pre-registered, interaction models between the parties suggest a statistically significant difference, with Democrats predicting modestly more retaliation.[13]

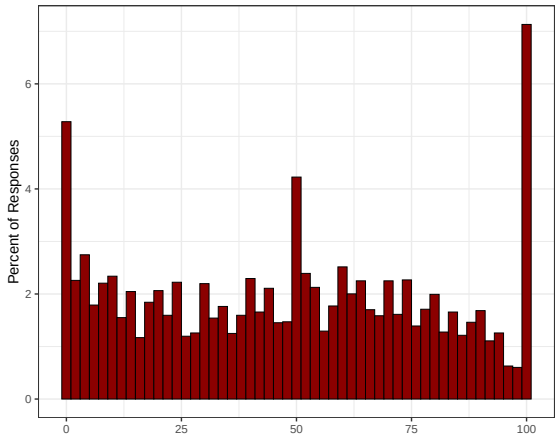
Both Democratic and Republican voters believe in the conditionality of the opposing party's anti-democratic behavior, consistent with H1A. H2A is partially supported: voters do predict their own party's extreme policy positions will promote retaliation, but the magnitude of these effects is quite small, and it is not possible to distinguish between social and economic proposals (see Table A20 for t-tests comparing the coefficients). Evidence about heterogeneity across outcome measures (H1D) is more ambiguous: it appears that subjects treat types of democratic violations comparably, though there are some small, statistically indistinguishable gaps in predictions of direct versus indirect violations. Figures 2a and 2b display the distribution of predictions across all random assignments, showing few

[13]In a simpler version of the same task, observed as a mediator in a future study, control-group respondents predicted a 12% probability of retaliation, as shown in Section 4.4.2. This suggests retaliation predictions are sensitive to topic but substantively remain modest.

respondents at the ceiling for either party and a broad spread of predictions.



(a) Distribution of Democratic predictions about Republicans

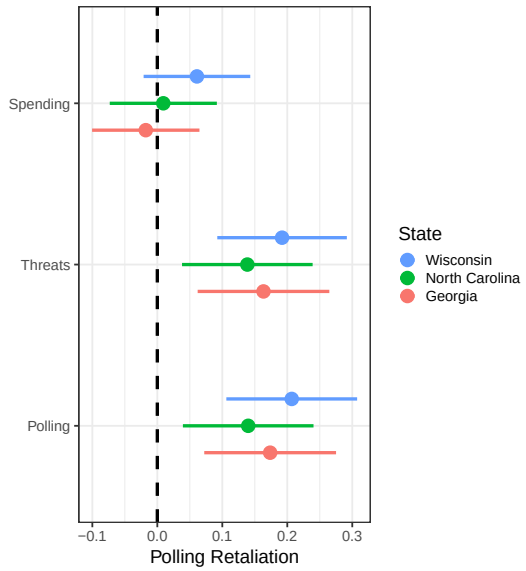


(b) Distribution of Republican predictions about Democrats

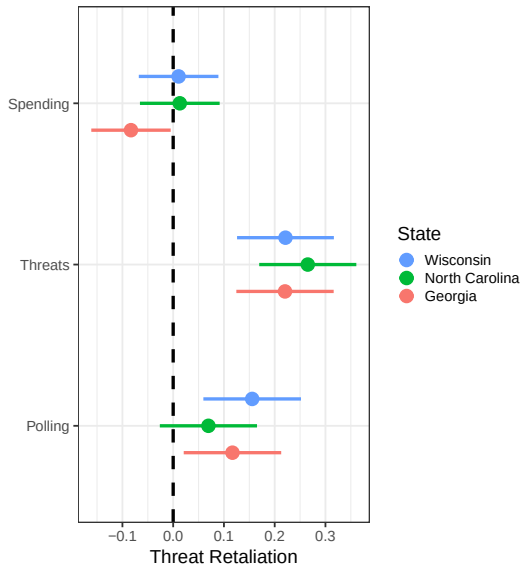
Figure 2: Distribution of predictions across party lines

Building on these retaliation predictions in a hypothetical context, I now extend these results to a more realistic setting. In experiment 2, I again find clear evidence of predicted retaliation, with similarly modest effect sizes. Respondents (pooling across parties) believe that in three key swing states, the Democratic Party is more likely to violate democratic rules if provoked than if unprovoked. The spending treatment, important for benchmarking beliefs about the effectiveness of anti-democratic behavior for a companion study, serves as a placebo test. Unlike democratic violations or the proposal of extreme policies, no existing theory suggests that a spending advantage should produce retaliation[14], although it is possible that participants could have predicted retaliation merely in response to electoral success, perhaps intuiting that those who project electoral defeat are likelier to endorse violence. Ultimately, we see that respondents do not predict Democratic politicians in swing states are more or less likely to retaliate when they have a spending advantage compared to when they do not. This lack of a treatment effect is particularly significant given that the spending treatment did move beliefs about vote share, suggesting that effects

[14]Notably, the predicted effects of the spending treatment on vote share are substantively similar to those of the violation treatments. Respondents are not simply predicting retaliation as a response to electoral success.



(a) Predictions of Electoral Violations



(b) Predictions of Threats

Figure 3: Predictions of Retaliation to Violations in Election Study

on predicted retaliation in the other randomization are driven not by electoral success, but by the violation of democratic norms.

Similar to my first study, there is relatively little evidence of discernment between different norm violations, consistent with my other results. This suggests voters see retaliation as possible through multiple mechanisms, not only direct in-kind responses. Prompts about Republican threats to elected officials have a larger effect on predictions of Democratic Party threats than on Democratic Party changes to electoral rules. However, there is no such gap for changes to electoral rules inspiring retaliation in kind. Even for the outcome of retaliation to the threat prompt, the point estimate is not statistically distinguishable from the expected polling retaliation. Across two separate contexts, hypothetical vignettes and more specific state-level scenarios, voters perceive the threat of retaliation as broad rather than narrow. Again, the substantive magnitude of predicted retaliation is modest, at between $\frac{1}{12}$ and $\frac{1}{6}$ standard deviations of the average prediction in the control group (i.e., the prediction of how likely the Democratic Party is to violate norms in states where it has not been provoked).

While Experiment 1 examined how Democrats and Republicans anticipated the opposing party would behave, Experiment 2 asks all partisans to predict the actions of the Democratic Party. While study 1 found a modest partisan gap, expectations of retaliation are similar in study 2. Across both studies, partisans and independents express roughly similar expectations about the likelihood of retaliation by political opponents.

From here, I consider whether modest retaliation predictions can be explained by a set of explanations common to the theoretical work that I summarized earlier. First, I consider the role of optimism about electoral victory and control over state governments. Second, I investigate second-order beliefs and whether subjects who are generally pessimistic about opponents predict less retaliation. Third, I use the structure of the experiments to explore whether subjects perceive diminishing marginal risks, such that each additional violation of democratic norms by the respondent's party reduces retaliation predictions. Ultimately, I find that these mechanisms play only a limited role in constraining retaliation predictions, with the strongest evidence that observing prior violations from one's own party reduces the marginal cost of future violations.

3.3.1 Do Partisan Optimists Predict Less Retaliation?

I begin by considering the role of beliefs about preemption, that is the perception that winning elections . While my simple formalization folded beliefs about preemption into the broad set of considerations that feed into retaliation predictions, my experiments offer more direct means of assessing beliefs that winning power will leave the opposing party unable to retaliate.[15]

Assessing preemption beliefs experimentally brings challenges so I draw on two features of my prediction experiments and discuss more advanced approaches in the Appendix. First, each experiment includes two types of predicted violation: one that requires control of

[15]Notably, this style of belief is a necessary but not sufficient condition for preemption beliefs which would also require voters and elites to believe that their party's democratic violations are effective

some level of government and one that does not. Second, because the second experiment was embedded in electoral predictions, I can examine how pre-treatment optimism about electoral outcomes correlates with expectations of retaliation. In neither case do I find strong evidence that respondents believe preemption is possible. Respondents are no more likely to predict retaliation for mechanisms such as intimidation in Experiment 1 or threatening election officials in Experiment 2—which do not require control of government—than for democratic violations that do require control of legal institutions (this is confirmed by formal t-tests comparing the coefficients). Similarly, voters more optimistic about their party's chances in Experiment 2 are no more likely to predict that the Democratic Party will retaliate against Republican violations of democratic norms. In the appendix, I discuss an edge case where optimism can undermine retaliation fears regarding court-packing. On the whole however, there is little evidence that voters believe winning elections permanently insulates their party from retaliation.

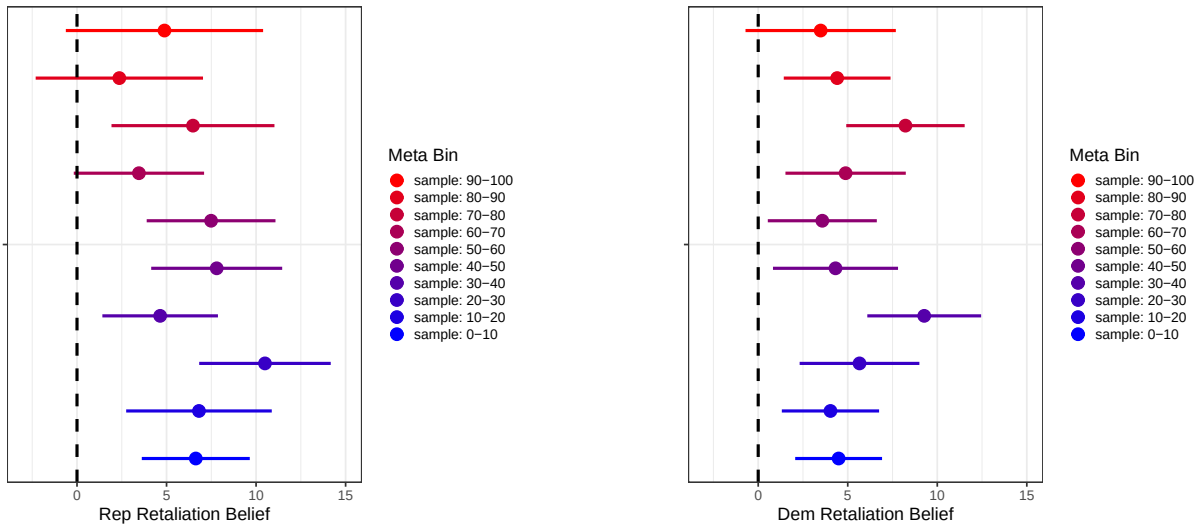
3.3.2 The Role of Second-Order Beliefs

To test H1C, I examine whether pre-treatment meta-perceptions of out-party support for norm violations shape treatment effects in Experiment 1. Prior work shows that such beliefs appear both correlationally and causally related to anti-democratic attitudes (Pasek et al., 2022; Mernyk et al., 2022; Braley et al., 2023). I expected that pessimistic respondents might predict less retaliation due to mechanical ceiling effects (Markovits and Liu, 2024).

My results do not support this expectation. As a covariate, meta-perceptions strongly correlate with retaliation predictions: respondents who saw the other party's voters as more supportive of norm violations also anticipated more retaliation, a result I replicate in a purely observational context in the appendix. However, meta-perceptions do not predict retaliation predictions. In pooled linear interaction models, only one interaction approaches significance ($p = 0.14$), but among Republicans all three coefficients are positive, with one marginally significant ($p = 0.08$). Substantively, a one standard deviation increase

in meta-perceptions corresponds to about a 10-point increase in predicted retaliation. Of note, meta-perceptions were highly prognostic of predicted retaliation as a covariate, suggesting they do correlate with pessimism about the actions of the opposing party.

One explanation for the partisan gap in this treatment-by-covariate interaction is that Democrats attribute Republican violations to Donald Trump, viewing retaliation as elite-driven, while Republicans see it as voter-driven. These findings challenge the claim that higher meta-perceptions weaken deterrence by making violations seem inevitable. It remains possible that this finding is an artifact of the relatively severe democratic violations in this study, though I show a similar pattern with regards to gerrymandering in Appendix section 1.11. Finally, to explore possibly non-linear interactions (Hainmueller et al., 2019), I present sub-group treatment effects by binned 10 percentage point meta-perception ranges in Figures 4a and 4b below, showing little evidence for heterogeneity with this specification. I also use causal forests (see (Wager and Athey, 2018) for a discussion) to more flexibly explore interaction effects in Appendix Figure A7.



(a) Democratic Predictions about Republicans (b) Republican Predictions about Democrats

Figure 4: Retaliation predictions by binned beliefs about anti-democratic attitudes among opposing partisans

My results emphasize that there is a distinction between second-order beliefs about

opposing partisans and expectations of how the opposing party will actually behave. Further, conditional beliefs about opponents represent an additional step beyond simple predictions. I confirm these intuitions with an analysis of meta-perception data from a Republican-only sub-sample in the Appendix 1.11, again showing that second-order beliefs about opposing partisans do not predict retaliation expectations.

3.3.3 Beliefs about the Effects of Multiple Violations

I now turn to an exploratory analysis of compounding provocations by examining heterogeneity in the number of democratic norms violated at the vignette level. Prior scholarship on the incremental nature of norm violations suggests that voters may struggle to mobilize in response to minor infractions or violations that are periodically walked back by incumbents (Grillo and Prato, 2023). Similarly, Frederiksen (2025) finds partial evidence that broader fears of democratic fragility can accentuate the damaging implications of a single norm violation while a treatment in Voelkel et al. (2024) explores how the downside risks of violence and chaos can reduce support for incremental undemocratic practices. In international relations, Pauly (2024) emphasizes the importance of "disentangling demands" such that partial escalations will be met by only partial punishments as a vital step in effective coercion. Similarly, from the perspective of voters, reactions to different paces and intensities of violations shape the effectiveness of retaliation in deterring anti-democratic behavior.

To analyze this mechanism, I report interaction effects for each additional violation from experiment 1. Specifically, I estimate Equation 3 where Y_{ij} is participant i's predicted index of responding-party norm violations for scenario j, with $j \in [1, 5]$ and τ_{other} represents the number of additional violations are present in the hypothetical scenario. This model investigates causal treatment-by-treatment interaction effects within individual profiles.

$$Y_{ij} = \beta_1\tau_1 + \beta_2\tau_{other} + \beta_3(\tau_1 \times X\tau_{other}) + \omega\chi + \epsilon_i \tag{3}$$

Table 2: Exploring Diminishing Returns to Retaliation Predictions

	(Prediction)	(Prediction)	(Prediction)
Arrest	7.657***	8.412***	7.602***
	(0.557)	(0.776)	(0.564)
Poll	6.592***	5.850***	5.788***
	(0.780)	(0.549)	(0.554)
Violence	6.172***	6.179***	6.805***
	(0.558)	(0.546)	(0.781)
Poll:Other Violation	−1.491*		
	(0.630)		
Arrest:Other Violation		−1.499*	
		(0.626)	
Violence:Other Violation			−1.377*
			(0.637)
Num.Obs.	17 904	17 904	17 904
R2	0.037	0.037	0.037
Std.Errors	by: Subject	by: Subject	by: Subject

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Models include demographic covariates as well as un-interacted treatment coefficients

These models consistently suggest modest diminishing returns: each additional violation reduces predicted retaliation by 1.37–1.49 percentage points. In the pooled model, violence is predicted to inspire 6.8 percentage points of retaliation in the absence of other violations, 5.4 percentage points if there have already been politicized arrests and 4.0 percentage points if there has been intimidation at the polls. The coefficients on the interaction terms in Table 2 refer to the change in the causal effect of the relevant violation in response to the presence of another violation. The "other violation" variable takes a value between 0 and 2. While modest, these interaction effects suggest one mechanism that mutes retaliation predictions: the observation of prior provocations.

These patterns point to a broader mechanism: as voters repeatedly observe their own party breaking norms, they predict diminishing marginal retaliation risks. These results are not driven by ceiling effects; nearly all predictions remain well below 100%. For instance, in Experiment 1 only 10% of Republican predictions about arrests exceed 95%, rising only

slightly to 10.5% with two provocations. Overall, the findings suggest that accumulated exposure to co-partisan violations over recent years may erode retaliation concerns, with a back-of-the-envelope calculation (assuming linear interaction effects) implying that after five or six violations, voters would expect no additional retaliation from further infractions.

3.4 How Much Should Partisans Fear Retaliation? Benchmarks from Elites

How should these perceptions of retaliation be contextualized? Unlike studies of factual misperceptions (Bursztyn and Yang, 2022; Braley et al., 2023; Ahler and Sood, 2018), there is no ground-truth probability that a violation of democratic norms triggers retaliation. I offer two possible benchmarks: the theoretical maximum retaliation prediction and predictions offered by a sample of political elites.[16]

The first benchmark is the maximum possible retaliation, defined as 100% minus the predicted probability that opponents will violate democratic norms in the absence of provocation.[17] For example, a Republican who believes that the Democratic Party would never violate the rules unprovoked would assign a maximum retaliation potential of 100%, whereas someone who assigns a 30% baseline probability would see a maximum retaliation potential of 70%. The structure of my experiments allows for group-level versions of these estimates.[18] Substantively, this benchmark captures the theoretical ceiling for how much retaliation fears could increase and is the equivalent of believing opponents are playing "grim trigger" in a repeated prisoner's dilemma - though my survey does not capture concerns about indefinite future periods. This benchmark ranges from 65-70 percentage points.

The second benchmark comes from the predictions of partisan elites. I conducted

[16] As an alternative benchmark, in the appendix I explore findings from existing surveys estimating the causal effect on support for a procedural violation of learning about a comparable violation from the opposing party

[17] In theory, the ideal case for deterrence would suggest voters assign a 0% chance of norm violations without provocation and a 100% chance with provocation. However, this is unrealistic given the highly polarized environment and widespread pessimism regarding the opposing party's democratic commitments.

[18] Some within-subject comparisons can also be made among individuals who view both extreme profiles, though such comparisons necessarily restrict the analysis to a subsample.

52 interviews with political elites over the period from October 2024 to September 2025. These elites[19], 33 Democrats and 19 Republicans, worked on campaigns, for interest groups and for think tanks. These individuals made decisions regarding tens of millions of dollars of campaign spending and advised elected officials and campaigns. In addition to open-ended qualitative questions, I asked these respondents to estimate five retaliation probabilities. The full text of these questions is found in Appendix section 1.12. Across more than 200 predictions from this sample, I recorded an average retaliation prediction of 29% percentage points, more than twice the average recorded by subjects in any of the retaliation experiments (and similarly higher than control group retaliation predictions in the final experiment discussed later in this paper). These results suggest that elites subscribe more strongly than voters to the logic of deterrence, though retaliation predictions are again far from the theoretical maximum.

In addition, elites provided qualitative accounts of when and why retaliation is likely, focusing on media attention from the opposing party's partisan outlets and on the manner in which provocations shaped internal party dynamics among opponents. A recurring concern was that provocations would mobilize moderates in the opposing party to support hardball when they otherwise would not have. In addition, elites expressed concerns about two types of personally targeted retaliation for their involvement in controversial tactics 1) Extra-legal targeting through doxxing, swatting, or death threats and 2) Legal targeting by government agencies, including tax investigations as well as regulatory investigations for their or their family's business interests.[20]

Relative to these benchmarks, my predicted retaliation results from survey samples offer three key implications for how citizens consider the strategic nature of democratic violations. First, voters appear to recognize that their own party's violations of democratic norms can provoke retaliation from the opposing party, even in the absence of explicit cues

[19]Which I defined by holding decision-making positions in partisan groups

[20]These interviews and related work are the subject of a companion dissertation paper

and using a design to mitigate social desirability bias (which might otherwise push towards limiting prospective retaliation by claiming that the opposing party's violations are not the responsibility of the respondents' own party) (Horiuchi et al., 2022). This suggests that the existing low levels of support for democratic backsliding (Holliday et al., 2024) may reflect an existing, implicit awareness of retaliatory risks. However, the magnitude of predicted retaliation is modest and is well below the maximum level of concern even when accounting for non-zero baselines and the predictions of elites. Second, voters do not sharply differentiate among types of democratic violations, although they are somewhat more responsive to the prospect of direct (in-kind) retaliation than to indirect forms across both prediction experiments. This pattern implies that concerns about opposition responses are relatively broad and not confined to violations involving state power.

Third and finally, my results offer explanations for *why* retaliation expectations are modest. My finding of diminishing marginal retaliation expectations suggests that learning about prior provocations reduces beliefs about the additional cost of future violations, and observers of recent American politics have witnessed many such violations. Meanwhile, the finding that inflated meta-perceptions of out-party extremism at the mass level do not dampen expectations of retaliation suggests that conditional expectations about opponents are orthogonal to pessimism about the democratic commitments of opposing partisans.

3.5 Does Retaliation Come to Mind?

My prediction experiments explicitly asked respondents to consider the actions of the opposing party. While I took steps to minimize experimenter-demand effects, these experiments did require respondents to think about the opposing party's behavior—considerations that might not immediately come to mind in many contexts. The question remains whether retaliation concerns arise absent prompting.

To answer this, I explored two datasets. First, 3,500 open-ended responses drawn

from my studies preceding treatment administration asked respondents to consider possible downsides to (depending on the study) gerrymandered maps or shuttering polling places in areas with high concentrations of the opposing party. Second, I examined real-world evidence from the comments sections of partisan YouTube videos that discussed democratic violations by the party with which the channel is generally aligned.[21] Full sample characteristics and the coding scheme are described in detail in the Appendix section *Retaliation Concerns in Text*. This yields a sample of $\approx$ 21,000 comments on Republican videos and $\approx$ 17,300 comments on Democratic videos.[22]

Using GPT-5.0 with the pre-registered prompt—"Does this response express fear or concern about this proposal causing others to behave badly or retaliate? Code as 1 if yes and 0 if no."—and following a few-shot prompting approach (100 examples each of the presence and absence of retaliation concerns hand-coded by me and a research assistant), I show that only 9-10% of respondents raised retaliation from the opposing party as a downside of their own party's efforts to revise democratic rules in a bipartisan context. This result is robust to alternative prompting: neither alternative prompts nor hand-coding produced more than a 10% incidence of retaliation concerns. The share drops to near zero in the YouTube comments, even as a percentage of the comments that are germane to the topic of the video. These results are summarized in Table 3. Retaliation concerns occur with some regularity among survey respondents when prompted but are vanishingly rare in the most heated segments of partisan discourse.

[21]The sampling frame was defined as videos that (1) came from clearly partisan channels as identified by Munger et al. (2025), and (2) discussed their own side's efforts to gerrymander a state (California for Democrats, Texas for Republicans), as identified by llm-coding.

[22]To account for occasional comments from members of the other party, I exclude those whose comments who identified them as opposing partisans or who gave irrelevant or incoherent comments

N Comments	Source	Party	% Retaliation Concerns
1,582	Survey	Both Parties	9.5%
1,881	Survey	Republican	10.4%
21,100	YouTube	Republican	0.1%
17,300	YouTube	Democratic	0.3%

Table 3: Summary of comments by source, party, and retaliation share.

4 Testing Warnings of Retaliation

In the prior section, I showed that voters anticipate retaliation, but that these expectations are modest and fall far short of benchmark predictions from elites. Open-ended responses provide some evidence that retaliation comes to mind without explicit prompting when respondents are considering the consequences of changing democratic rules, though not in hyper-partisan contexts. These findings help explain why sanctioning for anti-democratic behavior remains limited: voters are not overly concerned about retaliatory consequences when their party violates norms. Still, because few respondents predict retaliation with certainty and retaliation predictions persist across provocations and respondent types, priming fear of retaliation may increase commitment to democratic norms.

In this section, I use follow-up experiments to test whether warnings of retaliation can bolster democratic commitments by harnessing partisan self-interest. This approach treats warnings of retaliation as a treatment that may reduce support for procedural hardball or undemocratic practices, consistent with (Voelkel et al., 2024) and with the efforts of bridging or depolarization initiatives seeking to strengthen democratic values.

My final set of experiments evaluates whether messages that emphasize the conditional nature of democratic norm violations alter beliefs about their real-world consequences. I distinguish this treatment from prior efforts aimed at correcting assumptions about the other party's preferences (Mernyk et al., 2022; Braley et al., 2023; Christensen et al., 2025). The design parallels Corbett et al. (2022), who show that updating beliefs about female candidates' actual performance increases support for women in primary elections, whereas

merely shifting second-order perceptions of gender bias produces null effects. I argue that updating beliefs about retaliation operates in a similar way to shape voter preferences. Drawing on laboratory studies of strategic interaction (Di Tella et al., 2015; Arechar and Rand, 2022)[23] and my broader theoretical framework, I assume partisans can learn in general terms about the likely behavior of the opposing party. By framing retaliation from the opposing party as a tangible cost, I shift voters' strategic calculus in favor of preserving democratic norms. Taken together, these studies contribute to broader debates about voter competence and offer evidence of capacity for strategic reasoning in the mass public.

4.1 Method

To test whether individuals learn about conditional retaliation and update their preferences accordingly, I repeat a simple design across three separate survey samples and five randomized treatment assignments (total N of $\approx$ 5000, $\approx$ 9000 observations). While substantively distinct, these treatments have a common structure: they compare a *control condition* where a policy is proposed using neutral language to a *warning condition* where that proposal is accompanied by a warning that if it is adopted, it will lead to retaliation from the opposing party. Below I briefly describe the samples and designs of these experiments, though for the sake of parsimony, further details are confined to the appendix. All outcomes are in terms of standardized support for the proposal.

First, I randomly assigned a sample of 1,935 self-identified Democratic partisans, recruited via Prolific in July 2024, to either a threat or control condition with equal probability (study 3). In the control condition, participants read a proposal to pack the Supreme Court, described in neutral terms. In the threat condition, they read the same proposal but were additionally informed that Republicans would retaliate against Democratic court-packing—both in kind (by packing the court in response) and more aggressively (by disre-

[23]In this literature, participants can learn from prior histories of play, for example by observing a robot that consistently plays "always defect," and such learning influences future decisions.

garding the decisions made by a packed court). Unlike the violations tested in my earlier experiment, the retaliatory threat requires that the opposing party control the federal government since packing the court requires passing a law. This bundled warning of retaliation builds on findings from Experiments 1 and 2, which show that voters often anticipate retaliation across multiple domains.

Building on this test of a retaliation warning, I conducted a preregistered experiment in April 2025 that was designed to test a broader set of retaliation warnings (study 4). In this study, administered to 2,000 respondents (1,100 Democrats and 900 Republicans) recruited via Cloud Research Connect, I examined whether warnings of retaliation reduce support for three anti-democratic proposals: gerrymandering, disregarding court rulings, and altering Electoral College vote allocation rules to give a party's nominee an additional electoral vote in the 2028 presidential election. I refer to this as the *pooled experiment*. As noted in my analysis plan, I analyze each proposal separately and test for (and reject) spillover effects between the proposal-level randomizations.

As a final experiment (study 5), I examine beliefs about the escalating conflict over mid-decade redistricting 2026 midterm elections. This escalating conflict has included multiple explicit threats of retaliation from both parties. Notably, after Texas announced its intention to redistrict, California, led by Democratic Governor Gavin Newsom, publicly declared plans to retaliate. The experiment, which recruited 2,000 self-identified Republicans via Prolific, began with an open-ended question about respondents' concerns regarding their party's redistricting efforts (this is one of the open-ended questions used in the text analysis section), followed by two probability estimates and then a main outcome assessing support for gerrymandering among Republican respondents. It was conducted in mid-August 2025, when the redistricting fight received peak national media attention.

4.2 Hypotheses

For the final set of experiments, I hypothesized that randomized warnings about retaliation to each proposal would reduce support for anti-democratic behavior (H3A), as measured by a simple two-question index ("Do you approve of this behavior", "Would you be more or less likely to vote for a primary candidate proposing this behavior?") and that the warning conditions would be more effective for respondents who were more optimistic at baseline about the opposing party's commitment to democracy, as measured by beliefs about the opposing party's likelihood of violating norms (H3B). My pre-registrations for the final experiment about gerrymandering included several more specific hypotheses that are described in the linked pre-analysis plan.

4.3 Estimation

My final set of experiments is a series of simple two-arm designs where the only randomization is between a threat and a control condition. Because the pooled experiment involves multiple observations per individual, the pooled model reports clustered standard errors. In contrast, models that assess support for each violation use robust standard errors without clustering, since they involve only a single observation per respondent.

4.4 Results

After assessing baseline retaliation predictions in my first set of experiments, I now investigate whether explicitly priming the possibility of retaliation can reduce support for anti-democratic behavior. In Figure 5, I report standardized treatment effects on support for three partisan-motivated changes to democratic rules from a bipartisan sample, as well as two treatment effects for partisan samples. While some of these behaviors are contested in their democratic legitimacy (Wunsch et al., 2022), my theoretical argument does not depend on voters finding these behaviors normatively offensive, merely on subjects updating their

beliefs about the probability a given behavior inspires retaliation. The pooled estimate for experiment 4 is that a warning of retaliation reduces support for anti-democratic behavior by 0.11 standard deviations (95% CI 0.085-0.135) across the April 2025 outcomes. Notably, these reductions in support are more modest than the 0.29 standard deviation decrease of support for court-packing in experiment 3. In addition to the stronger wording of the warning in experiment 3, two explanations for these smaller effects are 1) that a broader set of pre-treatment questions about conditionality had already primed respondents across treatment and control conditions to consider the possibility of opposing party retaliation and 2) The proposals in this experiment were less popular in the control group (2.5 on a 5-point scale for the control group) than the court-packing scheme from experiment 3 (3.5 on a 5-point scale) suggesting the possibility of floor effects. Figure 5 also includes a precision weighted meta-analysis of these standardized treatment effects.

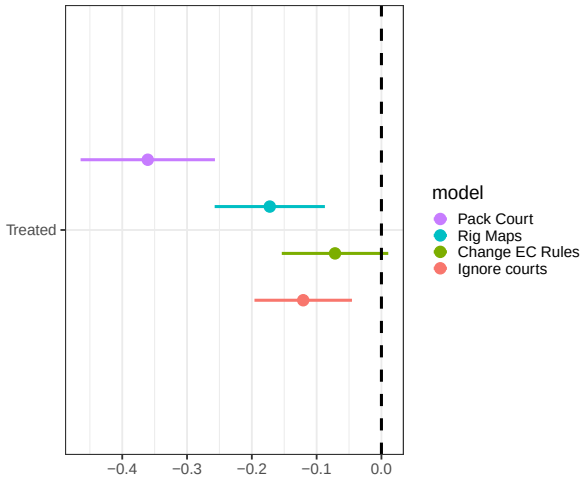


Figure 5: Treatment effects on randomized warnings of retaliation on support for procedural hardball

4.4.1 For Whom do Warnings Matter Most?

Next, I explore a series of preregistered heterogeneous effects analyses in Table 4 to investigate whether my top-line results suggest more complex strategic reasoning. These analy-

ses explore whether treatment effects vary by respondents' time horizons (see Gazmararian (2025) for how time horizons can affect policy preferences in the arena of climate change), risk aversion or pre-treatment beliefs about the behavior of the opposing party. I also explore heterogeneity by party, though I did not preregister directional hypotheses about partisan gaps in responsiveness to warnings of retaliation. I find neither substantively nor statistically significant heterogeneity across three of these dimensions. However, there is some directional evidence that the treatment is less effective among college educated respondents, because more sophisticated audiences might already grasp the logic of conditionality. The magnitude of this effect suggests that while warnings reduce support for anti-democratic behavior among the college-educated by less than 0.05 standard deviations, they reduce support among non-college respondents by 0.16 standard deviations.[24] One implication of these results is that warning treatments may work through priming the threat of retaliation rather than through numeric updating.

[24]These heterogeneous treatment effects are only for the middle three warning experiments fielded in April 2025. The other studies did not contain the same set of attitudinal covariates and are not part of this analysis

Table 4: Heterogeneous Effects

	(1)	(2)	(3)	(4)
Treat	−0.163***	−0.084	−0.095	−0.120***
	(0.032)	(0.067)	(0.067)	(0.034)
College	−0.072+	−0.021	−0.021	−0.021
	(0.044)	(0.034)	(0.034)	(0.034)
Risk Seeking	0.033*	0.033*	0.038*	0.033*
	(0.014)	(0.014)	(0.018)	(0.014)
Longer Time Horizons	−0.013	−0.005	−0.013	−0.013
	(0.018)	(0.023)	(0.018)	(0.018)
Republican	−0.102**	−0.102**	−0.103**	−0.102*
	(0.036)	(0.036)	(0.036)	(0.045)
Treat:College	0.102*			
	(0.049)			
Treat:Time Horizon		−0.015		
		(0.026)		
Treat:Risk Seeking			−0.009	
			(0.020)	
Treat:Republican				−0.002
				(0.048)
Num.Obs.	5918	5918	5918	5918
R2	0.284	0.284	0.284	0.284
Std.Errors	by: cluster	by: cluster	by: cluster	by: cluster

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Models include demographic covariates

4.4.2 How Voters Learn about Opposing Party Conditionality

Finally, I present results from my final experiment (study 5) examining Republican beliefs about gerrymandering in Texas. This Republican-led redistricting effort began after a request from President Trump and culminated with the approval of new congressional maps in August 2025. Unlike in earlier experiments, this study directly measured both respondents' support for rule changes and their beliefs about retaliation within the same survey. Specifically, respondents estimated the probability that the Democratic Party would gerrymander the state of California under two conditions: if Texas had gerrymandered first (Provoke = 1) and if Texas had not (Provoke = 0).

Figure 6 shows that treated respondents were approximately 6 percentage points more likely to expect Democrats to gerrymander in response to Republican provocation, and 19 percentage points less likely to expect gerrymandering when no provocation occurred—yielding a net 25-point increase in predicted retaliation due to the treatment.

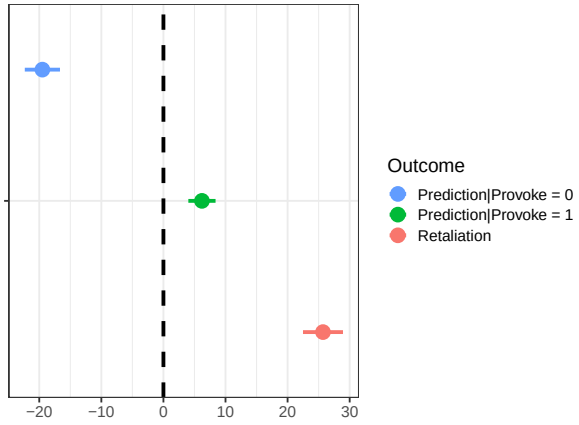


Figure 6: Updating about Democratic behavior in response to Newsom threat

This finding has two important implications. First, Republican beliefs about unprovoked Democratic behavior appear highly malleable. Second, this exercise allows for a straightforward back-of-the-envelope calculation: if a standard warning shifts retaliation beliefs by roughly 25 percentage points, then the treatment effects reported in Figure 5 represent about one-quarter of the potential reduction in anti-democratic behavior that could realistically be achieved through updates to beliefs about retaliation—again assuming linearity, and implicitly assuming an exclusion restriction such that warnings of retaliation affect preferences only through factual updating.

Meanwhile, Figure 7 shows respondents' predictions of the Democratic Party's unprovoked behavior (x-axis) and its provoked behavior (y-axis). Treated subjects are visibly concentrated in the upper-left corner, where predictions are near 100 percent conditional on provocation and near 0 percent otherwise. Notably, I included a meta-perceptions outcome as a placebo check and found no movement on this measure, suggesting that warnings of

retaliation from opposing-party elites do not cause updating (in either direction) regarding the preferences of opposing partisans at the mass level.

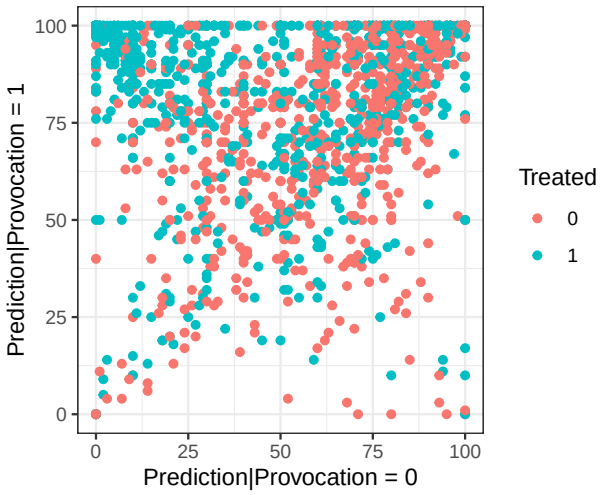


Figure 7: Distribution of Conditional Predictions about Democratic Party Gerrymandering

5 Discussion: The Promise and Limits of Democratic Deterrence

Accounts of the mass public as a check on politicians who seek to violate democratic rules rely on partisans, whether voters or elites, adhering to norms even at the expense of their ideological goals or career ambitions (Graham and Svolik, 2020; Frederiksen, 2024). At the voter level, it is increasingly clear that abstract commitments to democracy do not reliably translate into the punishment of anti-democratic politicians, particularly once elites make the case for norm violations (Clayton et al., 2021; Krishnarajan, 2023) and given the broader decline of "floating voters" in recent American politics (Smidt, 2017). Even when interventions succeed in strengthening public support for democratic norms, they often depend on political elites willing to criticize their own party or praise the opposition. Treatments of this form constitute some of the most successful efforts to reduce anti-democratic attitudes (Voelkel et al., 2024; Weiss et al., 2025). Such strategies expose messengers to reputational backlash (Hussein and Wheeler, 2024) and electoral risk (Bartels and Carnes, 2023). There is no evidence that substantial numbers of partisan elites are willing to disseminate such

pro-normative messages.

In this paper, I examine a mechanism that is instead rooted in partisan self-interest and a style of intervention that can be disseminated by aggressive, often pugilistic partisan elites. I find that in the absence of explicit prompting, retaliation concerns are modest and rarely appear un-prompted in online partisan discourse. My results suggest these patterns are partially attributable to diminishing returns, such that voters believe prior provocations have already convinced the opposing party to violate norms. However, the absence of ceiling effects and the fact that retaliation predictions are not muted for those who are pessimistic about opposing partisans at the mass level suggests that there is room for simple informational treatments to reinforce the risk of retaliation.

Consistent with formal accounts across disciplines, skepticism towards opponents can serve a constructive purpose so long as opponents are understood to act conditionally. Importantly, these final experiments suggest that harsh rhetoric from political opponents need not fuel a spiral of escalation. Instead, partisans can recognize conditional threats as distinct from unconditional signals that democratic norms will be violated. The documented rise of threatening rhetoric from American political elites (Zeitzoff, 2023; Kim et al., 2025) need not uniformly contribute to escalation. In fact, my results suggest that awareness of how norm violations fuel escalation can help avert democratic erosion in the mass public.

Even when politicians make vituperative and threatening remarks, couching these threats in conditional language can cause opposing partisans to refrain from provocations. This provides an avenue for elites to subtly reduce tensions without abandoning their partisan commitments. Recent papers have highlighted the challenge of scaling depolarization and have emphasized the importance of shifting elite rhetoric (Holliday et al., 2025). My results suggest that conditional warnings can reduce support for democratic violations; learning about spirals of escalation can help to defuse them.

This paper also provides important evidence of a limited form of strategic reasoning

among the mass public. Compared to elites, my online samples predicted far less retaliation, and elites described a complex framework through which they assessed retaliation risks, which included the scope of media coverage and the role of opposing party moderates in catalyzing attacks on democracy. Despite this, the public is capable of updating in ways consistent with optimistic theoretical accounts. As retaliation becomes more salient, survey respondents become less supportive of provoking opponents by violating democratic norms or trying to manipulate procedural rules for partisan gain. The electorate is sufficiently sophisticated to grasp an important theoretical logic that can protect democracy. Finally, my interviews offer novel evidence on elite beliefs about when and how opposing partisans will retaliate for democratic transgressions. These results shed light on how important actors consider strategic trade-offs in a new era of democratic contestation.

References

Daron Acemoglu. A theory of political transitions. page 48, 2001.

Daron Acemoglu and James A. Robinson. *Economic Origins of Dictatorship and Democracy*. Cambridge University Press, December 2005. ISBN 978-1-139-44695-2. Google-Books-ID: 2eKcAgAAQBAJ.

Christopher Achen and Larry Bartels. Democracy for realists: Holding up a mirror to the electorate. *Juncture*, 22(4):269–275, 2016. ISSN 2050-5876. doi: 10.1111/j.2050-5876.2016.00873.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2050-5876.2016.00873.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2050-5876.2016.00873.x.

Douglas J. Ahler and Gaurav Sood. The Parties in Our Heads: Misperceptions about Party Composition and Their Consequences. *The Journal of Politics*, 80(3):964–981, July 2018. ISSN 0022-3816. doi: 10.1086/697253. URL https://www.journals.uchicago.edu/doi/abs/10.1086/697253. Publisher: The University of Chicago Press.

Antonio A. Arechar and David G. Rand. Learning to be selfish? A large-scale longitudinal analysis of Dictator games played on Amazon Mechanical Turk. *Journal of Economic Psychology*, 90:102490, June 2022. ISSN 0167-4870. doi: 10.1016/j.joep.2022.102490. URL https://www.sciencedirect.com/science/article/pii/S0167487022000083.

Larry M. Bartels and Nicholas Carnes. House Republicans were rewarded for supporting Donald Trump's 'stop the steal' efforts. *Proceedings of the National Academy of Sciences*, 120(34):e2309072120, August 2023. doi: 10.1073/pnas.2309072120. URL https://www.pnas.org/doi/full/10.1073/pnas.2309072120. Publisher: Proceedings of the National Academy of Sciences.

David A. Bateman. Democracy-Reinforcing Hardball: Can Breaking Democratic Norms Preserve Democratic Values? *Comparative Political Studies*, page 00104140241312107, January 2025. ISSN 0010-4140. doi: 10.1177/00104140241312107. URL https://doi.org/10.1177/00104140241312107. Publisher: SAGE Publications Inc.

Tyler Bowen, Michael A. Goldfien, and Matthew H. Graham. Public Opinion and Nuclear Use: Evidence from Factorial Experiments. *The Journal of Politics*, 85(1):345–350, January 2023. ISSN 0022-3816. doi: 10.1086/720329. URL https://www.journals.uchicago.edu/doi/full/10.1086/720329. Publisher: The University of Chicago Press.

Alia Braley, Gabriel S. Lenz, Dhaval Adjodah, Hossein Rahnama, and Alex Pentland. Why voters who value democracy participate in democratic backsliding. *Nature Human Behaviour*, 7(8):1282–1293, August 2023. ISSN 2397-3374. doi: 10.1038/s41562-023-01594-w. URL https://www.nature.com/articles/s41562-023-01594-w. Publisher: Nature Publishing Group.

Leonad Bursztyn and David Yang. Misperceptions About Others | Annual Reviews. 14: 425–452, 2022. URL https://www.annualreviews.org/content/journals/10.1146/annurev-economics-051520-023322.

John Carey, Gretchen Helmke, Mitchell Sanders, Katherine Clayton, and Brendan Nyhan. Who Will Defend Democracy? Evaluating Tradeoffs in Candidate Support Among Partisan Donors and Voters. 2022.

Aaron Christensen, Daniel Markovits, and Andrew Thompson. Meta-Perception Corrections in the Field. *Working Paper*, 2025.

Katherine Clayton, Nicholas T. Davis, Brendan Nyhan, Ethan Porter, Timothy J. Ryan, and Thomas J. Wood. Elite rhetoric can undermine democratic norms. *Proceedings of the National Academy of Sciences*, 118(23):e2024125118, June 2021. doi: 10.1073/pnas.2024125118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2024125118. Publisher: Proceedings of the National Academy of Sciences.

Hayley Cohen. "Primary Concerns: The Lack of Forward-Looking Strategic Voting in Primary Elections", 2025.

Hayley Cohen and Daniel B Markovits. Crossover Voting in Congressional Primaries. 2025.

Elizabeth C Connors, Taylor N Carlson, and Steven W Webster. You're Making Us Look Bad: Can Partisan Embarrassment Dampen Partisanship and Polarization? 2025.

Christianne Corbett, Jan G. Voelkel, Marianne Cooper, and Robb Willer. Pragmatic bias impedes women's access to political leadership. *Proceedings of the National Academy of Sciences*, 119(6):e2112616119, February 2022. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2112616119. URL https://pnas.org/doi/full/10.1073/pnas.2112616119.

Sirianne Dahlum, Torbjørn Hanson, Åshild Johnsen, Andreas Kotsadam, and Alexander Wuttke. Is Support for Authoritarian Rule Contagious? Evidence from Field and Survey Experiments, 2024. URL https://www.ssrn.com/abstract=5049239.

Ernesto Dal Bó, Pedro Dal Bó, and Erik Eyster. The Demand for Bad Policy when Voters Underappreciate Equilibrium Effects. *The Review of Economic Studies*, 85(2):964–998, April 2018. ISSN 0034-6527. doi: 10.1093/restud/rdx031. URL https://doi.org/10.1093/restud/rdx031.

Pedro Dal Bó, Guillaume R. Fréchette, and Jeongbin Kim. The determinants of efficient behavior in coordination games. *Games and Economic Behavior*, 130:352–368, November 2021. ISSN 0899-8256. doi: 10.1016/j.geb.2021.08.010. URL https://www.sciencedirect.com/science/article/pii/S0899825621001172.

Nicholas T. Davis, Kirby Goidel, and Keith Gaddie. *Democracy's Meanings: How the Public Understands Democracy and Why It Matters*. University of Michigan Press, August 2022. ISBN 978-0-472-13312-3.

Rafael Di Tella, Ricardo Perez-Truglia, Andres Babino, and Mariano Sigman. Conveniently Upset: Avoiding Altruism by Distorting Beliefs about Others' Altruism. *The American Economic Review*, 105(11):3416–3442, 2015. ISSN 0002-8282. URL https://www.jstor.org/stable/43821379. Publisher: American Economic Association.

Nicholas C Dias, Laurits F Aarslew, Kristian Vrede Skaaning Frederiksen, Yphtach Lelkes, Lea Pradella, and Sean J Westwood. Correcting misperceptions of partisan opponents

is not effective at treating democratic ills. *PNAS Nexus*, 3(8):pgae304, August 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae304. URL https://academic.oup.com/pnasnexus/article/doi/10.1093/pnasnexus/pgae304/7730165.

James N. Druckman, Suji Kang, James Chu, Michael N. Stagnaro, Jan G. Voelkel, Joseph S. Mernyk, Sophia L. Pink, Chrystal Redekopp, David G. Rand, and Robb Willer. Correcting misperceptions of out-partisans decreases American legislators' support for undemocratic practices. *Proceedings of the National Academy of Sciences*, 120(23):e2301836120, June 2023. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2301836120. URL https://pnas.org/doi/10.1073/pnas.2301836120.

James D. Fearon and David D. Laitin. Explaining Interethnic Cooperation. *American Political Science Review*, 90(4):715–735, December 1996. ISSN 0003-0554, 1537-5943. doi: 10.2307/2945838. URL https://www.cambridge.org/core/journals/american-political-science-review/article/abs/explaining-interethnic-cooperation/CE9BC6184CEB72ECD6E18E17041BAB12.

Kristian Frederiksen. Considering Democracy?, March 2025. URL https://www.researchgate.net/publication/389974564_Considering_Democracy.

Kristian Vrede Skaaning Frederiksen. Do Partisanship and Policy Agreement Make Citizens Tolerate Undemocratic Behavior? *The Journal of Politics*, 86(2):766–781, April 2024. ISSN 0022-3816, 1468-2508. doi: 10.1086/726938. URL https://www.journals.uchicago.edu/doi/10.1086/726938.

Karolin Freitag, Laura Kiemes, and Alexander Wuttke. Replication of "Why voters who value democracy participate in democratic backsliding". Working Paper 235, I4R Discussion Paper Series, 2025. URL https://www.econstor.eu/handle/10419/319603.

Laura Gamboa. *Opposition at the Margins*. Cambridge University Press, November 2022. ISBN 978-1-00-916406-1.

Alexander F. Gazmararian. Valuing the Future: Changing Time Horizons and Policy Preferences. *Political Behavior*, 47(2):553–572, June 2025. ISSN 1573-6687. doi: 10.1007/s11109-024-09965-3. URL https://doi.org/10.1007/s11109-024-09965-3.

Alan S. Gerber and Donald P. Green. *Field experiments : design, analysis, and interpretation*. W.W. Norton, 2012. URL https://cir.nii.ac.jp/crid/1971149384742718888.

Alan S. Gerber, Donald P. Green, and Christopher W. Larimer. Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment. *American Political Science Review*, 102(1):33–48, February 2008. ISSN 0003-0554, 1537-5943. doi: 10.1017/S000305540808009X. URL https://www.cambridge.org/core/product/identifier/S000305540808009X/type/journal_article.

Matthew H. Graham and Milan W. Svolik. Democracy in America? Partisanship, Polarization, and the Robustness of Support for Democracy in the United States. *American Political Science Review*, 114(2):392–409, May 2020. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055420000052. URL https://www.cambridge.org/core/product/identifier/S0003055420000052/type/journal_article.

Edoardo Grillo and Carlo Prato. Reference Points and Democratic Backsliding. *American Journal of Political Science*, 67(1):71–88, 2023. ISSN 1540-5907. doi: 10.1111/ajps.12672. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12672. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12672.

Edoardo Grillo, Zhaotian Luo, Monika Nalepa, and Carlo Prato. Theories of Democratic Backsliding. 2023.

Jens Hainmueller and Daniel J. Hopkins. The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes toward Immigrants. *American Journal of Political Science*, 59(3):529–548, 2015. ISSN 1540-5907. doi: 10.1111/ajps.12138. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12138. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12138.

Jens Hainmueller, Jonathan Mummolo, and Yiqing Xu. How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice. *Political Analysis*, 27(2):163–192, April 2019. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2018.46. URL https://www.cambridge.org/core/product/identifier/S1047198718000463/type/journal_article. Publisher: Cambridge University Press (CUP).

Gretchen Helmke, Mary Kroeger, and Jack Paine. Democracy by Deterrence: Norms, Constitutions, and Electoral Tilting. *American Journal of Political Science*, 66(2):434–450, 2022. ISSN 1540-5907. doi: 10.1111/ajps.12668. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12668. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12668.

Derek E. Holliday, Shanto Iyengar, Yphtach Lelkes, and Sean J. Westwood. Uncommon and nonpartisan: Antidemocratic attitudes in the American public. *Proceedings of the National Academy of Sciences*, 121(13):e2313013121, March 2024. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2313013121. URL https://pnas.org/doi/10.1073/pnas.2313013121.

Derek E. Holliday, Yphtach Lelkes, and Sean J. Westwood. Why depolarization is hard: Evaluating attempts to decrease partisan animosity in America. *Proceedings of the National Academy of Sciences*, 122(39):e2508827122, September 2025. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2508827122. URL https://pnas.org/doi/10.1073/pnas.2508827122.

Yusaku Horiuchi, Zachary Markovich, and Teppei Yamamoto. Does Conjoint Analysis Mitigate Social Desirability Bias? *Political Analysis*, 30(4):535–549, October 2022. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2021.30. URL https://www.cambridge.org/core/product/identifier/S1047198721000309/type/journal_article.

Mohamed A. Hussein and S. Christian Wheeler. Reputational costs of receptiveness: When and why being receptive to opposing political views backfires. *Journal of Experimental Psychology: General*, 153(6):1425–1448, 2024. ISSN 1939-2222. doi: 10.1037/xge0001579. Place: US Publisher: American Psychological Association.

Lisa Janssen, Anna Kern, and Hannah Werner. 'When they go low, we kick them' : affective

polarization and its effects on the support for reciprocal democratic transgressions. In *Working Paper*, 2025. URL https://files.osf.io/v1/resources/pwcz2_v1/providers/osfstorage/68484094d7e96b4d055395bb?format=pdf&action=download&direct&version=2.

Robert Jervis. Cooperation under the Security Dilemma. *World Politics*, 30(2): 167–214, January 1978. ISSN 1086-3338, 0043-8871. doi: 10.2307/2009958. URL https://www.cambridge.org/core/journals/world-politics/article/abs/cooperation-under-the-security-dilemma/C8907431CCEFEFE762BFCA32F091C526.

Robert O. Keohane. *After Hegemony: Cooperation and Discord in the World Political Economy*. Princeton University Press, 1984. ISBN 978-1-4008-2026-9. Google-Books-ID: HnvpdocqT9EC.

Seo-young Silvia Kim, Jan Zilinsky, and Brian Brew. Donate to help us fight back: Political fundraising and toxic rhetoric online. *Party Politics*, 31(3):549–561, May 2025. ISSN 1354-0688. doi: 10.1177/13540688241235901. URL https://doi.org/10.1177/13540688241235901. Publisher: SAGE Publications Ltd.

Lina Koppel, David Andersson, Magnus Johannesson, Eirik Strømland, and Gustav Tinghög. Comprehension in economic games. *Journal of Economic Behavior & Organization*, 234: 107039, June 2025. ISSN 0167-2681. doi: 10.1016/j.jebo.2025.107039. URL https://www.sciencedirect.com/science/article/pii/S0167268125001581.

Suthan Krishnarajan. Rationalizing Democracy: The Perceptual Bias and (Un)Democratic Behavior. *American Political Science Review*, 117(2):474–496, May 2023. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055422000806. URL https://www.cambridge.org/core/journals/american-political-science-review/article/rationalizing-democracy-the-perceptual-bias-and-undemocratic-behavior/C78EB8AE1CC777B4392EE73727F4F25C.

Thomas J. Leeper, Sara B. Hobolt, and James Tilley. Measuring Subgroup Preferences in Conjoint Experiments. *Political Analysis*, 28(2):207–221, April 2020. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2019.30. URL https://www.cambridge.org/core/product/identifier/S1047198719000305/type/journal_article.

Matthew Levendusky. *Our Common Bonds: Using What Americans Share to Help Bridge the Partisan Divide*. University of Chicago Press, March 2023. ISBN 978-0-226-82469-7.

Steven Levitsky and Daniel Ziblatt. *How Democracies Die - Google Books*. Penguin, New York, 2018. URL https://www.google.com/books/edition/How_Democracies_Die/VZKADwAAQBAJ?hl=en&gbpv=1&dq=how+democracies+die&pg=PA1&printsec=frontcover.

Jack Lucas, Lior Sheffer, Peter John Loewen, Stefaan Walgrave, Karolin Soontjens, Eran Amsalem, Stefanie Bailer, Nathalie Brack, Christian Breunig, Pirmin Bundi, Linda Coufal, Patrick Dumont, Sarah Lachance, Miguel M. Pereira, Mikael Persson, Jean-Benoit Pilet, Anne Rasmussen, Maj-Britt Sterba, and Frédéric Varone. Politicians' Theories of Voting Behavior. *American Political Science Review*, pages 1–18, November 2024. ISSN 0003-

0554, 1537-5943. doi: 10.1017/S0003055424001060. URL https://www.cambridge.org/core/product/identifier/S0003055424001060/type/journal_article.

Noam Lupu, Eli Rau, and Elizabeth J Zechmeister. Public Tolerance for Anti-Democratic Behavior. 2025.

Janet Malzahn and Andrew B. Hall. Election-Denying Republican Candidates Underperformed in the 2022 Midterms. *American Political Science Review*, pages 1–6, October 2024. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055424001084. URL https://www.cambridge.org/core/journals/american-political-science-review/article/electiondenying-republican-candidates-underperformed-in-the-2022-midterms/E42E3513A1DCD141A84708528921 98C7.

Daniel Markovits and Hayley Cohen. Encouraging Crossover Voting in Presidential Primaries. *Working Paper*, 2025.

Daniel Markovits and Patrick Liu. How do scale ceilings affect "compensation demand" survey responses?, 2024. URL https://business.yougov.com/content/51175-scale-ceilings-impact-on-compensation-demand.

Joseph S. Mernyk, Sophia L. Pink, James N. Druckman, and Robb Willer. Correcting inaccurate metaperceptions reduces Americans' support for partisan violence. *Proceedings of the National Academy of Sciences*, 119(16):e2116851119, April 2022. doi: 10.1073/pnas.2116851119. URL https://www.pnas.org/doi/full/10.1073/pnas.2116851119. Publisher: Proceedings of the National Academy of Sciences.

Michael K. Miller. A Republic, If You Can Keep It: Breakdown and Erosion in Modern Democracies. *The Journal of Politics*, 83(1):198–213, January 2021. ISSN 0022-3816. doi: 10.1086/709146. URL https://www.journals.uchicago.edu/doi/full/10.1086/709146. Publisher: The University of Chicago Press.

Kevin Munger, Matt Hindman, Omer Yalcin, Joseph Phillips, and James Bisbee. Pressing Play on Politics: Quantitative Description of YouTube. *Journal of Quantitative Description: Digital Media*, 5, March 2025. ISSN 2673-8813. doi: 10.51685/jqd.2025.006. URL https://journalqd.org/article/view/8674.

Monika Nalepa, Georg Vanberg, and Caterina Chiopris. A wolf in sheep's clothing: Citizen uncertainty and democratic backsliding. *The Journal of Politics*, page 734253, December 2024. ISSN 0022-3816, 1468-2508. doi: 10.1086/734253. URL https://www.journals.uchicago.edu/doi/10.1086/734253.

Douglass C. North and Barry R. Weingast. Constitutions and Commitment: The Evolution of Institutions Governing Public Choice in Seventeenth-Century England. *The Journal of Economic History*, 49(4):803–832, 1989. ISSN 0022-0507. URL https://www.jstor.org/stable/2122739. Publisher: [Economic History Association, Cambridge University Press].

Michael H. Pasek, Lee-Or Ankori-Karlinsky, Alex Levy-Vene, and Samantha L. Moore-Berg. Misperceptions about out-partisans' democratic values may erode democracy. *Scientific Reports*, 12(1):16284, September 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-19616-4.

URL https://www.nature.com/articles/s41598-022-19616-4. Publisher: Nature Publishing Group.

Reid B. C. Pauly. Damned If They Do, Damned If They Don't: The Assurance Dilemma in International Coercion. *International Security*, 49(1):91–132, July 2024. ISSN 0162-2889. doi: 10.1162/isec_a_00488. URL https://doi.org/10.1162/isec_a_00488.

Corwin D. Smidt. Polarization and the Decline of the American Floating Voter. *American Journal of Political Science*, 61(2):365–381, 2017. ISSN 1540-5907. doi: 10.1111/ajps.12218. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12218. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12218.

Vicente Valentim. Norms of Democracy, Staged Democrats, and Supply of Exclusionary Ideology. *Comparative Political Studies*, page 00104140241283009, October 2024. ISSN 0010-4140. doi: 10.1177/00104140241283009. URL https://doi.org/10.1177/00104140241283009. Publisher: SAGE Publications Inc.

Paul A. M. Van Lange, Anthon Klapwijk, and Laura M. Van Munster. How the shadow of the future might promote cooperation. *Group Processes & Intergroup Relations*, 14 (6):857–870, November 2011. ISSN 1368-4302. doi: 10.1177/1368430211402102. URL https://doi.org/10.1177/1368430211402102. Publisher: SAGE Publications Ltd.

Jan G. Voelkel, Michael N. Stagnaro, James Y. Chu, Sophia L. Pink, Joseph S. Mernyk, Chrystal Redekopp, Isaias Ghezae, Matthew Cashman, Dhaval Adjodah, Levi G. Allen, L. Victor Allis, Gina Baleria, Nathan Ballantyne, Jay J. Van Bavel, Hayley Blunden, Alia Braley, Christopher J. Bryan, Jared B. Celniker, Mina Cikara, Margarett V. Clapper, Katherine Clayton, Hanne Collins, Evan DeFilippis, Macrina Dieffenbach, Kimberly C. Doell, Charles Dorison, Mylien Duong, Peter Felsman, Maya Fiorella, David Francis, Michael Franz, Roman A. Gallardo, Sara Gifford, Daniela Goya-Tocchetto, Kurt Gray, Joe Green, Joshua Greene, Mertcan Güngör, Matthew Hall, Cameron A. Hecht, Ali Javeed, John T. Jost, Aaron C. Kay, Nick R. Kay, Brandyn Keating, John Michael Kelly, James R. G. Kirk, Malka Kopell, Nour Kteily, Emily Kubin, Jeffrey Lees, Gabriel Lenz, Matthew Levendusky, Rebecca Littman, Kara Luo, Aaron Lyles, Ben Lyons, Wayde Marsh, James Martherus, Lauren Alpert Maurer, Caroline Mehl, Julia Minson, Molly Moore, Samantha L. Moore-Berg, Michael H. Pasek, Alex Pentland, Curtis Puryear, Hossein Rahnama, Steve Rathje, Jay Rosato, Maytal Saar-Tsechansky, Luiza Almeida Santos, Colleen M. Seifert, Azim Shariff, Otto Simonsson, Shiri Spitz Siddiqi, Daniel F. Stone, Palma Strand, Michael Tomz, David S. Yeager, Erez Yoeli, Jamil Zaki, James N. Druckman, David G. Rand, and Robb Willer. Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity. *Science*, 386(6719): eadh4764, October 2024. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adh4764. URL https://www.science.org/doi/10.1126/science.adh4764.

Stefan Wager and Susan Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113 (523):1228–1242, July 2018. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2017.1319839. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1319839. Publisher: Informa UK Limited.

Barry R. Weingast. The Political Foundations of Democracy and the Rule of Law. *The American Political Science Review*, 91(2):245–263, 1997. ISSN 0003-0554. doi: 10.2307/2952354. URL https://www.jstor.org/stable/2952354. Publisher: [American Political Science Association, Cambridge University Press].

Chagai Weiss, Don Green, and Robb Willer. Politicians' Bipartisan Appeals to Civility and Partisan Divides: A Field Experiment with U.S. Governors, May 2025. URL https://osf.io/5qxyw_v1.

Sean J. Westwood, Justin Grimmer, Matthew Tyler, and Clayton Nall. Current research overstates American support for political violence. *Proceedings of the National Academy of Sciences*, 119(12):e2116870119, March 2022. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2116870119. URL https://pnas.org/doi/full/10.1073/pnas.2116870119.

Jason Willick. Opinion | The weighty lesson from Arizona's 'fake electors' stumble. *The Washington Post*, May 2025. ISSN 0190-8286. URL https://www.washingtonpost.com/opinions/2025/05/23/fake-electors-trump-prosecution/.

WSJ WSJ Editorial Board. Opinion | Should Harvard Be Tax Exempt?, April 2025. URL https://www.wsj.com/opinion/donald-trump-harvard-tax-exemption-irs-supreme-court-congress-94ba5b53. Section: Opinion.

Natasha Wunsch, Marc Jacob, and Laurenz Derksen. The Demand Side of Democratic Backsliding: How Divergent Understandings of Democracy Shape Political Choice. 2022. URL file:///C:/Users/danie/Dropbox/PC%20(2)/Downloads/Wunsch_Jacob_Derksen_understandings_v1.pdf.

Thomas Zeitzoff. *Nasty Politics: The Logic of Insults, Threats, and Incitement*. Oxford University Press, 2023. ISBN 978-0-19-767948-7. Google-Books-ID: GUDAEAAAQBAJ.

Appendix

1.1 Experimental Materials

1.1.1 Experiment 1

Var	Text
Arrests	INPARTY state attorney generals sought to prosecute prominent OUTPARTY without evidence (INPARTY state attorney generals equally investigated election misconduct on both sides)
Polling	INPARTY in key swing states closed polling places in heavily OUTPARTY areas (INPARTY in key states ensured that polling places were open equally for both parties)
Violence	INPARTY observers intimidated Republican voters (There were no reports of intimidation or violence)
Social (for Dems)	Democratic supported weakening immigration enforcement for humanitarian reasons (Democrats supported cracking down on the border to reduce illegal immigration)
Tax (for Dems)	Democrats campaigned on dramatically raising taxes on the wealthy (Democrats campaigned on keeping tax rates mostly where they are)
Social (for Reps)	Republicans campaigned in support of a national abortion ban (Republicans campaigned on returning abortion policy to the states)
Econ (for Reps)	Republicans campaigned on dramatically cutting corporate taxes and slashing social security benefits (Republicans campaigned on slightly reducing corporate taxes and maintaining social security benefits at current levels)

Table A1: Attribute levels for Experiment 1 (Baseline in parentheses)

- What percentage of OPPOSING PARTY voters do you think agree with the following statements? Please answer on the scale below from 0% to 100%. (Meta-perceptions question pre-treatment)
  - "The stakes of politics are so high that we should break the rules to win"
  - "We should ban rallies organized by the e://Field/party Party"
- In this scenario, how likely are Democrats to do the following in the year after the election? Please answer on the scale below where 100 is very likely and 0 is not at all likely. (Main prediction outcome)
  - OUTPARTISANS will engage in violence
  - OUTPARTY will have leading INPARTISANS arrested without evidence

1.1.2 Experiment 2

Cheating Levels	(In STATE, many ordinary Republicans have threatened officials who oversee elections. There have also been threats directed against Democratic elected officials)/(North Carolina Republicans have limited some forms of voting that are used more frequently by Democrats. North Carolina Republicans have limited the number of polling places in heavily Democratic parts of the state.)/(EMPTY)
Spending Levels	(Both campaigns have heavily contested the state but the Harris campaign and its allies have spent substantially more)/(Both campaigns have heavily contested the state and spent heavily to try to persuade voters.)

Table A2: Conditions for the Prediction Experiment

- Prompt: As you may know, the 2024 presidential election is coming up shortly. In the next section we are going to ask you to make a few predictions about how the election will go in different states.
- I want you to think about how Donald Trump will do in the state of STATE. In 2020, Trump narrowly won the state. What percentage of the vote do you think he will get this election in STATE? Please answer on the slider below. To incentivize you to make your best guess, we will give a $1 bonus after the election to the 10% of respondents that get closest to the true answer.
- Earlier you made a guess about what percent of the vote Trump will get in STATE. Now we want you to guess again. Now what percentage of the vote do you think Donald Trump will receive in the state of North Carolina? Please answer on the slider below. This answer will REPLACE your prior guess and you will win a $1 bonus if you are in the 10% of respondents who get closest to the correct answer.
- Now I want you to think about Democrats in the state of STATE. How likely are they do the following things in the year after the election? (Very likely/somewhat likely/neither likely nor unlikely/somewhat unlikely/very unlikely)
  - Send Threats to Republican Officials
  - Try to make it harder for Republicans to vote

1.1.3 Retaliation Warning Experiments

These experiments address diverse topics and use diverse language but their commonality is that they ask respondents to assess their support for a proposal to engage in a behavior that violates procedural or democratic norms, these proposals are randomly presented either with neutral language (control condition) or accompanied by a warning that the proposal will lead to retaliation from the opposing party (warning condition). For parsimony, text that is shown in bulleted form in actual treatments is compressed into a paragraph in Table A3 below.

Group	Description
Treatment (SCOTUS)	Now I want you to consider some details about a proposal made by some Democrats to pack the Supreme Court by adding more liberal justices. *(1) The Supreme Court is currently controlled by conservatives with a 6-3 majority. (2) Democrats have made proposals to add between 0 and 9 more justices (3) Republicans have said that if Democrats do this they will retaliate by ignoring Supreme Court Rulings. (4) Republicans have also threatened to add more justices themselves if Democrats do it first*
Control (SCOTUS)	Now I want you to consider some details about a proposal made by some Democrats to pack the Supreme Court by adding more liberal justices.
Treatment (Gerrymander)	The election for the House of Representatives is likely to be very close, the last 3 elections have been decided by fewer than 10 seats. Because the election is so close, both sides have explored redrawing congressional maps to make it easier to win more seats. *Democrat Gavin Newsom, the Governor of California, has said that if Republicans redraw their maps, California will as well. But if Republicans keep the same maps, so will California. In reference to his plans, Newsom said "They stop, we stop. Simple as that." A proposed law in California allows redrawing to benefit Democrats only if others states make changes to their congressional maps first. This means that if Republicans redraw maps to their benefit in Texas, Democrats will retaliate and cancel out any advantage Republicans might get.*
Control (Gerrymander)	Before, please read some information about the midterms: The election for the House of Representatives is likely to be very close, the last 3 elections have been decided by fewer than 10 seats.
Treatment (EC)	Legislators in [YOUR PARTY] are considering a proposal to allocate electoral votes in [STATE] such that their candidate will win an extra electoral vote. If the [PARTY] takes this step, the [OPPOSING PARTY] is likely to retaliate by changing the rules in [OPPOSING PARTY STATE] to give their candidate an extra vote. This pattern is fairly common in fights over presidential elections. When one side changes the rules, the other side strikes back.
Control (EC)	Legislators in [YOUR PARTY] are considering a proposal to allocate electoral votes in [STATE] to gain an automatic extra vote for their candidate
Treatment (Courts)	e://Field/opp judges routinely issue orders that block policies supported by prominent politicians of the e://Field/party.Governors of states run by the e://Field/party are considering a proposal to ignore court orders from judges that are loyal to the e://Field/opp. Leaders of the e://Field/opp have said that if the e://Field/party violates court orders, governors belonging to the e://Field/opp will retaliate by doing the same in the future. This pattern is fairly common in fights over the courts. When one side breaks the rules, the other side strikes back.
Control	e://Field/opp judges routinely issue orders that block policies supported by prominent politicians of the e://Field/party. Governors of states run by the e://Field/party are considering a proposal to ignore court orders from judges that are loyal to the e://Field/opp.
Treatment	Leaders of the e://Field/party are considering a proposal to aggressively redraw congressional maps at the next possible opportunity, so as to win up to a dozen more house seats. Experts warn that this approach will lead the e://Field/opp to retaliate by re-drawing maps in the states that it controls. This pattern is common in fights over congressional maps. When one side breaks the rules, the other side strikes back.
Control	Leaders of the e://Field/party are considering a proposal to aggressively redraw congressional maps at the next possible opportunity, so as to win up to a dozen more house seats.

Table A3: Treatment and Control Conditions for Warning Experiments

SCOTUS (1)	Give this information, do you approve of disapprove of proposals for Democrats to add seats to the court?
Maps (1)	Do you agree or disagree: the $e://Field/party should re-draw maps in the states it controls so that it wins more congressional seats?
Maps (2)	Would you be more or less likely to vote for a congressional candidate in a $e://Field/party primary in your state who proposes re-drawing the electoral maps so that the $e://Field/party wins more seats?
EC (1)	Do you agree or disagree: Members of the $e://Field/party should change the rules in $e://Field/state so that the $e://Field/party presidential candidate is certain to get another electoral vote in the 2028 presidential election?
EC (2)	Would you be more or less likely to vote for a congressional candidate in a $e://Field/party primary in your state who proposes changing electoral rules so that the nominee of the $e://Field/party is guaranteed to win an additional electoral vote in 2028?
Courts (1)	Do you agree or disagree: $e://Field/party elected officials should sometimes consider ignoring court decisions when the judges who issued those decisions were appointed by $e://Field/opp presidents?
Courts (2)	Would you be more or less likely to vote for a congressional candidate in a $e://Field/party primary in your state who proposes ignoring court orders from judges appointed by $e://Field/opp presidents?
Maps Texas	(To what extent do you agree with the following statements? (Republicans in Texas should redraw maps so Democrats win fewer seats)/ (Republican primary candidates should always support redrawing maps to advantage the Republican Party)/ (Republicans should never compromise with Democrats on the topic of drawing fair maps for congressional elections)

Table A4: Outcome Measure for Warning Experiments

1.2 Sample Characteristics and Balance Tables

Because this paper includes multiple experiments, I report descriptive statistics for the five experimental samples and balance tests showing no concerning imbalances across experimental conditions. Some experiments were embedded in larger surveys, so available covariates differ across studies. I present balance and descriptive statistics for binary indicators of college education, gender, and white ethnicity for all conditions and experiments and include study-specific covariates that are theoretically relevant as available.

1.2.1 Experiment 1

Because Experiment 1 is a fully randomized, single-profile conjoint, balance is assessed across profiles. Specifically, I compare average respondent characteristics for each condition, weighting by the number of profiles with that condition that each respondent viewed. Although the sample was not designed to be nationally representative, its gender, education, and racial distributions are broadly comparable to those of the national electorate.

Table A5: Balance Across Arrest Conditions

	Arrest = 0		Arrest = 1			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
poll	0.49	0.50	0.51	0.50	0.01	0.01
violence	0.50	0.50	0.50	0.50	0.01	0.01
meta_average	43.83	31.83	44.28	32.28	0.45	0.51
age	41.64	13.54	41.60	13.51	-0.04	0.20
female	0.56	0.50	0.55	0.50	-0.01	0.01
white	0.75	0.43	0.75	0.43	0.01	0.01
college	0.53	0.50	0.54	0.50	0.00	0.01

1.2.2 Experiment 2

1.2.3 Retaliation Warning Experiments

1.3 Full Models - With Index and Specific Prediction Outcomes

Table A6: Balance Across Democratic Violation Conditions, Experiment 2

	control (N=3214)		polls (N=3178)		threats (N=3211)	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Spend	0.5	0.5	0.5	0.5	0.5	0.5
Female	0.5	0.5	0.5	0.5	0.5	0.5
College	0.7	0.5	0.7	0.5	0.7	0.5
Prior Prediction	52.5	12.9	52.2	13.0	52.4	12.7

Table A7: Balance Across Spending Conditions, Experiment 2

	0		1			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
Female	0.5	0.5	0.5	0.5	0.0	0.0
College	0.7	0.5	0.7	0.5	0.0	0.0
Prior Prediction	52.2	12.8	52.5	13.0	0.2	0.3

	0		1			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
female	0.60	0.49	0.59	0.49	-0.01	0.02
college	0.63	0.48	0.63	0.48	0.00	0.02
white	0.68	0.47	0.66	0.47	-0.02	0.02
age	39.25	12.78	39.10	12.71	-0.15	0.57
affpol	52.73	28.44	53.89	28.64	1.16	1.28
optimism	0.56	0.50	0.55	0.50	-0.02	0.02

Table A8: Balance Table for court-packing Assignment

	0		1			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
female	0.56	0.50	0.55	0.50	-0.01	0.02
college	0.41	0.49	0.42	0.49	0.01	0.02
white	0.62	0.49	0.63	0.48	0.01	0.02
age	43.56	16.47	43.27	16.12	-0.29	0.73
risks_1	3.02	1.27	2.93	1.26	-0.10	0.06
norms	3.55	1.98	3.50	1.92	-0.05	0.09
assignM	0.49	0.50	0.51	0.50	0.03	0.02
assignE	0.52	0.50	0.48	0.50	-0.04	0.02

Table A9: Balance Table for Court Ignoring Assignment

	0		1			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
female	0.55	0.50	0.56	0.50	0.00	0.02
college	0.39	0.49	0.44	0.50	0.05	0.02
white	0.63	0.48	0.62	0.49	-0.01	0.02
age	43.15	16.43	43.68	16.15	0.53	0.73
risks_1	2.98	1.27	2.97	1.26	-0.02	0.06
norms	3.51	1.97	3.55	1.93	0.04	0.09
assignC	0.52	0.50	0.48	0.50	-0.04	0.02
assignM	0.49	0.50	0.51	0.50	0.02	0.02

Table A10: Balance for Electoral College Change Assignment

	0		1			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
assignE	0.49	0.50	0.51	0.50	0.02	0.02
female	0.55	0.50	0.56	0.50	0.00	0.02
college	0.42	0.49	0.41	0.49	-0.01	0.02
white	0.62	0.48	0.62	0.48	0.00	0.02
age	43.29	16.47	43.54	16.12	0.25	0.73
risks_1	2.96	1.26	2.99	1.28	0.02	0.06
norms	3.59	2.03	3.47	1.87	-0.12	0.09
assignC	0.49	0.50	0.51	0.50	0.03	0.02

Table A11: Balance for First Gerrymandering Assignment

	0		1			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
female	0.60	0.49	0.60	0.49	0.00	0.02
white	0.87	0.34	0.84	0.36	-0.02	0.02
college	0.54	0.50	0.53	0.50	-0.01	0.02
age	46.27	59.14	44.61	13.53	-1.66	1.95
Aff-pol	44.04	33.78	44.25	31.23	0.21	1.48
Norms Pre	2.78	0.98	2.76	0.90	-0.02	0.04
Prior Metas	31.02	19.46	30.77	18.39	-0.25	0.86

Table A12: Balance for Final Warning Experiment

Table A13: Experiment 1, Main Results with Index Outcomes, Full Models

	All
(Intercept)	16.894***
	(2.659)
arrest	6.522***
	(0.431)
poll	4.828***
	(0.419)
violence	5.235***
	(0.413)
econ	2.387***
	(0.415)
social	2.580***
	(0.410)
age	−0.062*
	(0.031)
raceHispanic/Latino	1.223
	(2.104)
raceOther	3.112
	(1.930)
raceWhite	0.593
	(1.435)
College (such as BA, BS)	−2.383+
	(1.385)
educKindergarten through grade 11	4.792
	(4.737)
educMaster's degree or higher	−3.198*
	(1.609)
educNo schooling completed	−3.124
	(10.809)
educRegular high school diploma or GED	−1.419
	(1.710)
educSome college credit but no degree	−2.282
	(1.506)
meta_average	0.433***
	(0.014)
Num.Obs.	15 731
R2	0.273
Std.Errors	by: ResponseId

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Models include demographic covariates

Table A14: Main Results with Index and Specific Prediction Outcomes, Republican Sample, Experiment 1

	Index	Arrests	Violence
Prosecute	5.454***	6.066***	4.842***
	(0.557)	(0.602)	(0.606)
Closed Polls	4.190***	3.358***	5.021***
	(0.546)	(0.595)	(0.596)
Intimidation	4.809***	4.426***	5.192***
	(0.539)	(0.583)	(0.586)
Slash Welfare	2.185***	2.035***	2.335***
	(0.529)	(0.586)	(0.577)
Abortion Ban	2.895***	2.614***	3.176***
	(0.526)	(0.575)	(0.576)
Num.Obs.	9770	9770	9770
R2	0.278	0.237	0.256
Std.Errors	by: Subject	by: Subject	by: Subject

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Models include demographic covariates

Table A15: Main Results with Index and Specific Prediction Outcomes, Democratic Sample Experiment 1

	Index	Arrests	Violence
Prosecute	8.196***	10.868***	6.615***
	(0.683)	(0.804)	(0.734)
Close Polls	5.932***	5.257***	6.978***
	(0.659)	(0.730)	(0.739)
Intimidation	5.790***	4.494***	6.868***
	(0.648)	(0.740)	(0.730)
High Taxes	2.667***	2.362**	2.552***
	(0.680)	(0.731)	(0.738)
Open Border	2.105**	2.420**	3.343***
	(0.657)	(0.735)	(0.731)
Num.Obs.	5961	6616	6616
R2	0.240	0.055	0.050
Std.Errors	by: Subject	by: Subject	by: Subject

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Models include demographic covariates

Table A16: Full Models with All Randomizations, Experiment 2

	Reps	Dems	Inds
Polls	0.182***	0.101**	0.193*
	(0.048)	(0.033)	(0.076)
Threats	0.257***	0.163***	0.134+
	(0.047)	(0.033)	(0.074)
Harris Spend	−0.003	−0.027	0.065
	(0.039)	(0.027)	(0.061)
Num.Obs.	3948	4306	1251
R2	0.047	0.129	0.126
Std.Errors	IID	IID	IID

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001
Models include demographic covariates

Table A17: Full Models with All Randomizations, Experiment 2, Pooling Across All Respondents

	Index	Poll Retaliation	Threat Retaliation
Polls	0.117***	0.179***	0.149***
	(0.028)	(0.029)	(0.027)
Threats	0.235***	0.166***	0.201***
	(0.028)	(0.029)	(0.027)
Harris Spend	−0.024	0.017	−0.004
	(0.023)	(0.024)	(0.022)
Prior Guess	0.008***	0.011***	0.010***
	(0.001)	(0.001)	(0.001)
Num.Obs.	9524	9505	9494
R2	0.234	0.267	0.276
Std.Errors	IID	IID	IID

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001
Models include demographic covariates

Table A18: Effect of Retaliation Warning on Predictions of Democratic Gerrymandering

	Retaliation	Prediction—Provoke = 1	Prediction—Provoke = 0
(Intercept)	35.872***	75.819***	39.948***
	(3.759)	(2.563)	(3.305)
treated	25.473***	6.322***	−19.151***
	(1.497)	(1.021)	(1.316)
college	5.881***	1.360	−4.521***
	(1.530)	(1.043)	(1.346)
norms_pre	−7.243***	−2.812***	4.431***
	(0.888)	(0.606)	(0.781)
prior_metas	−0.036	0.195***	0.231***
	(0.040)	(0.028)	(0.035)
female	−1.959	−1.003	0.956
	(1.568)	(1.069)	(1.379)
white	−2.113	1.692	3.805*
	(2.173)	(1.482)	(1.911)
conservativeExtreme Conservative	−1.168	0.446	1.615
	(2.084)	(1.421)	(1.832)
conservativeModerate	3.152+	−2.115+	−5.267***
	(1.811)	(1.235)	(1.592)
age	−0.002	0.006	0.008
	(0.017)	(0.012)	(0.015)
affpol	−0.080**	0.039*	0.119***
	(0.028)	(0.019)	(0.024)
Num.Obs.	1886	1886	1886
R2	0.189	0.059	0.192
Std.Errors	IID	IID	IID

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Models include demographic covariates

Table A19: Effect of Threat on Support for Court-packing, Full Model

	(1)
Warning	−0.345***
	(0.052)
Electoral Optimism	0.270***
	(0.054)
Affective Polarization	0.528***
	(0.055)
College Education	−0.100+
	(0.054)
White	−0.070
	(0.057)
Age	−0.011***
	(0.002)
Female	0.089+
	(0.053)
Num.Obs.	1935
R2	0.099

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Models include demographic covariates as well as uninteracted treatment coefficients

1.4 Additional Model Specifications

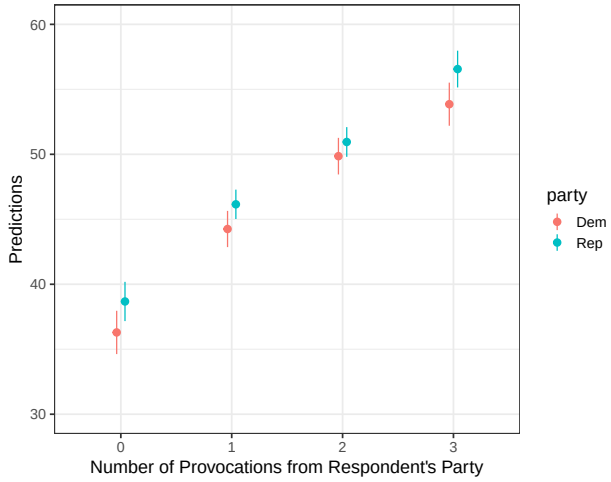


Figure A1: Marginal Means of predictions by number of violations and party, Experiment 1

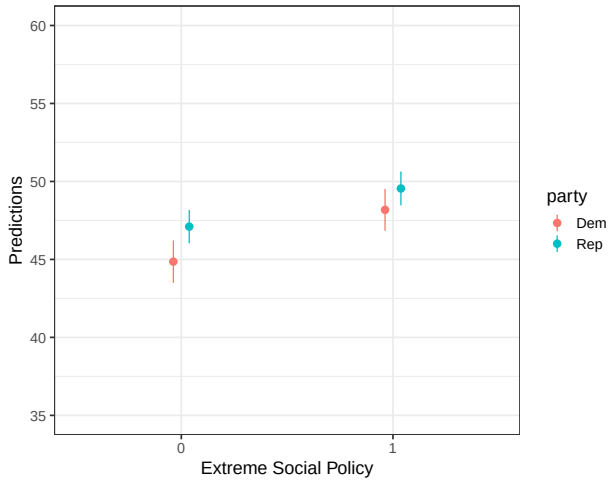


Figure A2: Marginal Means of predictions by social extremism and party, Experiment 1

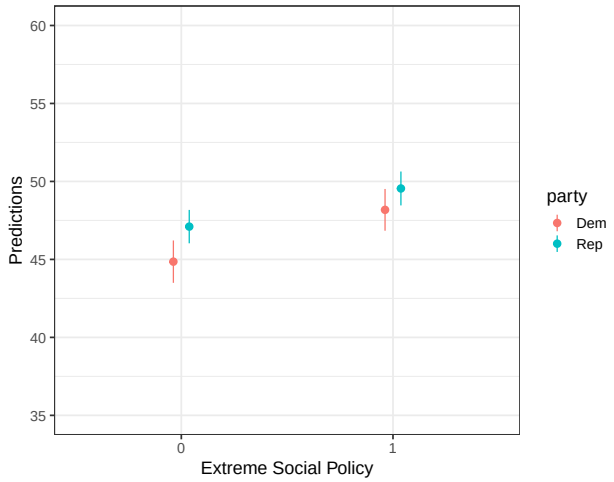


Figure A3: Marginal Means of predictions by economic policy extremism and party, Experiment 1

Table A20: Linear Hypothesis Tests Between Coefficients, Experiment 1

	(1)	(2)	(3)	(4)	(5)
Econ = Social	−0.637				
	(0.625)				
Arrest = Violence		1.420*			
		(0.627)			
Arrest = Poll			1.783**		
			(0.625)		
Violence = Poll				0.363	
				(0.624)	
Arrest = Econ					4.848***
					(0.623)
Num.Obs.	17 904	17 904	17 904	17 904	17 904
R2	0.033	0.033	0.033	0.033	0.033
Std.Errors	HC2	HC2	HC2	HC2	HC2

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001
Models include demographic covariates

1.5 Exploring State-to-State Spillovers in Experiment 2

In Figure A4, I examine how information about the Republican Party's behavior in earlier states influences predictions about the Democratic Party's behavior in later states in Experiment 2. The results provide evidence of "spillovers" across states: learning that Republicans violated a democratic norm in the first or second state shifts beliefs about

Democratic retaliation in the second or third state, though these effects are smaller than the direct treatment effects. Importantly, this analysis is a pre-registered approach to testing a potential violation of the within-subjects stable-unit-treatment-value assumption (SUTVA) Gerber and Green (2012). In this context, an individual's potential outcomes (predictions) in one state are shaped not only by the information assigned for that state but also by information assigned for other states. To account for this risk, I control for prior randomizations in main models as appropriate (the models estimating effects for the second states control for the assignment status of the first state and the models estimating effects for the third state include controls for the assignment of the first two states).

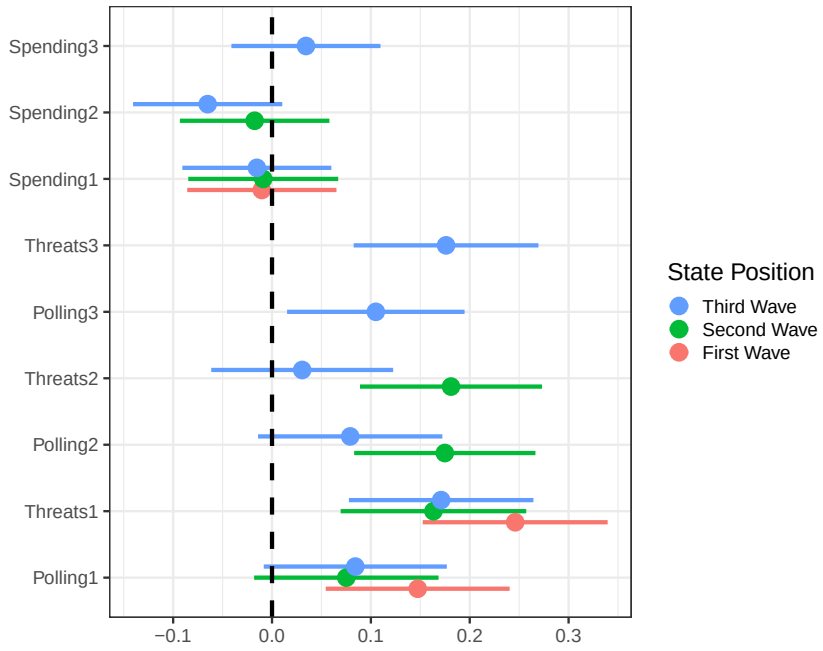


Figure A4: Spillovers in retaliation predictions across state-level estimates

1.6 Testing Perceptions of Preemption

As discussed in the main text, one mechanism that could limit predictions of retaliation is the belief that violating rules and then winning office would prevent the opposing party from retaliating. I use three distinct approaches to probe this belief. Ultimately, I find little evidence that voters view such preemption as likely, except for a narrow edge case involving Supreme Court reform.

First, in each prediction experiment there are two types of predicted violations: one that requires control of state government and one that does not. I find no clear differentiation between these outcomes in either Experiment 1 or Experiment 2. Second, the court-packing experiment included a pretreatment measure of electoral optimism; as expected, more optimistic Democrats were modestly less deterred by retaliation warnings. Third, the second prediction experiment was embedded in forecasts of Trump's vote share in swing states. Each prediction of retaliation was preceded by a pretreatment prediction of Trump's vote share. As shown in Table A21, there is neither a statistically significant nor a substantively meaningful interaction between this pretreatment measure and beliefs about the Democratic Party's response. As discussed in the text, winning the presidency need not rule out state-level retaliation. Still, assumptions about unified control—or general optimism about the Republican Party's prospects—could, in theory, blunt expectations of retaliation. The Democratic Party's state-level prospects also varied (it controlled no branch of state government in Georgia but had partial control in North Carolina and Wisconsin), yet retaliation predictions did not vary by state, suggesting that respondents may not have incorporated these institutional features when forming their expectations.

Finally, the timing of the Supreme Court experiment allows an exploratory look at the role of electoral optimism. Court-packing is a useful test: winning the election reduces the risk of the threatened retaliation—Republicans could not carry it out without unified federal control. Table A22 shows that optimistic Democrats were directionally less deterred by the warning treatment, though the interaction effect was not statistically significant. This pattern suggests that the credibility of future deterrence may diminish when voters expect their party to retain control of key levers of power. As noted above, this logic is most applicable to norm violations that require unified federal control—more clearly satisfied in the Supreme Court case than in the other retaliation warnings analyzed here.

1.7 Attitude Stability of Predictions

One counterargument to the importance of these attitudes is that this type of strategic reasoning is alien to most voters, who tend to have weak priors and unstable beliefs. This

Table A21: Interaction between Electoral Expectations and Retaliation Prediction

	Dem Sample	Rep Sample	All
Predicted Trump Share	0.008**	0.008**	0.009***
	(0.002)	(0.003)	(0.002)
Polling	0.153	0.401*	0.205+
	(0.161)	(0.204)	(0.117)
Threats	0.303+	0.126	0.126
	(0.164)	(0.212)	(0.122)
Spending	−0.022	−0.016	−0.003
	(0.028)	(0.038)	(0.022)
Predicted Trump: Threats	−0.001	−0.004	−0.001
	(0.003)	(0.004)	(0.002)
Predicted Trump: Polling	−0.003	0.002	0.001
	(0.003)	(0.004)	(0.002)
Num.Obs.	4219	3900	9366
R2	0.158	0.052	0.280
Std.Errors	by: participantId	by: participantId	by: participantId

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001
Models include demographic covariates

Table A22: Effect of Threat on Support for Court-packing

	Standardized Support	Standardized Support
Threat	−0.345***	−0.432***
	(0.052)	(0.078)
Optimism	0.270***	0.191*
	(0.054)	(0.075)
Threat:Optimism		0.156
		(0.105)
Num.Obs.	1935	1935
R2	0.104	0.105

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001
Models include demographic covariates as well as uninteracted treatment coefficients

is why I investigate elite beliefs through interviews and an ongoing elite survey in follow-up work. However, compared to related attitudes (notably meta-perceptions), I show that beliefs about the opposing party's commitment to democracy are relatively more stable. Examining data from a recent working paper (Markovits et al., 2025), I find that beliefs about the other party's willingness to violate democratic norms exhibit substantially

greater stability than meta-perceptions of democratic beliefs among out-partisans at the mass level. One explanation for this gap is that partisan media extensively discusses the opposing party's willingness to violate democratic rules, and both Democratic and Republican politicians have made explicit and repeated claims that their opponents seek to engage in political prosecutions and the stifling of civil liberties. In contrast, discussion of the mass public's beliefs about the opposing party's actual behavior is limited and infrequent. In Experiment 2, my repeated observations of retaliation predictions across states allow me to estimate features of attitude stability, and I show in Figure A5 that there is a very high correlation in predictions across states.

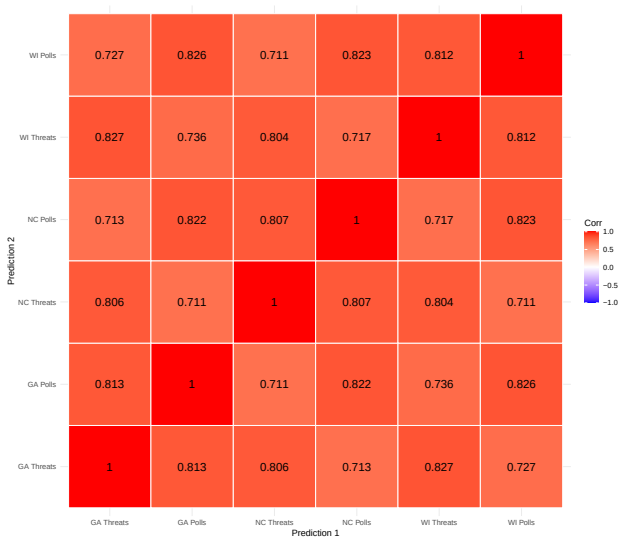


Figure A5: Correlation between state-level predictions of Democratic Party retaliation

1.8 Text Analysis and Retaliation Fears

The text samples are as follows:

- representative sample of 1931 Americans (646 Democrats, 711 Independents, 552 Republicans) were asked to respond to two open-ended questions, one considering the upsides of a proposal to restrict polling access in areas where opponents are popular and one considering the downsides.
- Convenience sample of 1900 self-identified Republicans

- Comments on Republican YouTube channels
- Comments on Democratic YouTube channels

My prediction experiment explicitly asked respondents to consider the actions of the responding party. However, it is unclear what comes to mind when considering anti-democratic actions of co-partisans. To investigate this question, I conducted a series of open-ended analyses using text data. First, I asked respondents in a separate survey sample to consider the upsides and downsides of anti-democratic behavior. Then, responses were coded with *gpt-4-turbo* and *gpt-5* with the following preregistered prompt: "The text after a colon is a Republican (Democrat) describing his or her reaction to a proposal to unfairly help Republicans (Democrats) in an election. Please output a 1 if the text expresses fear or concern about this proposal causing others to behave badly or retaliate and 0 otherwise:" I then hand-coded the full sample of 1,582 responses. As an example, responses that were coded as expressing fear of retaliation included 20 that expressed concerns about "riots" or "violence" while 11 used the word "retaliation". In contrast, concerns about fairness, democracy or other normative considerations were coded as not mentioning retaliation.

Second, I explore comments on YouTube videos from partisan channels that describe violations of democratic norms from co-partisans. I first identified 114 videos with a total of 47,000 comments from the leading 100 partisan YouTube channels identified by Munger et al. (2025) and downloaded through the Google API, accessed via the tuber package in R. I then coded the full set of the survey comments and a random sample of 1,000 comments from the YouTube comments by hand. For the YouTube comments, I used both a zero-shot and few-shot learning approach, in the latter case providing a balanced set of hand-coded comments (100 with retaliation fears and 100 without) as training data. Results are substantively unchanged across both approaches.

1.9 Additional Heterogeneous Effect Models Across Experiments

Here I show exploratory interaction models as well as machine learning approaches from Wager and Athey (2018) in order to investigate both different substantive treatment-by-covariate interactions and address concerns about non-linearity that are not explored in my pre-registered linear interaction models.

1.9.1 Experiment 1

In line with Hainmueller et al. (2019), I replicate the interaction models for experiment 1 by using dummy, binned versions of continuous variables and then through causal forests using the grf package in R (Wager and Athey, 2018) to allow for non-linear interaction models.

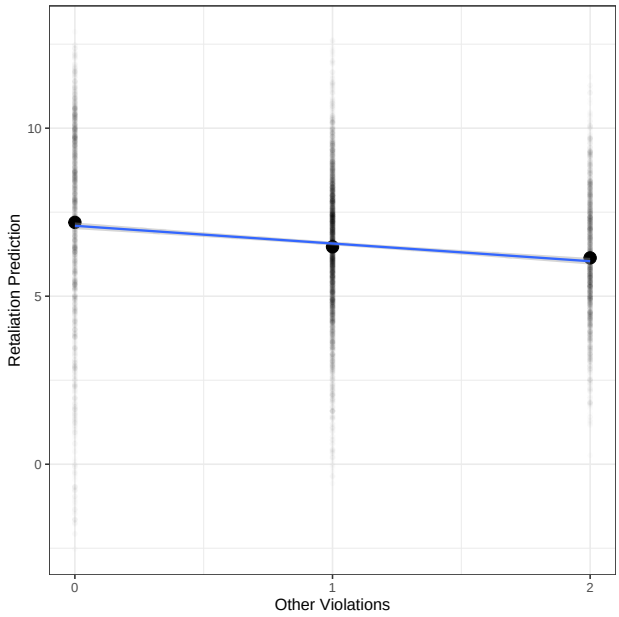


Figure A6: Heterogeneous effects of violation on retaliation outcome by number of other violations

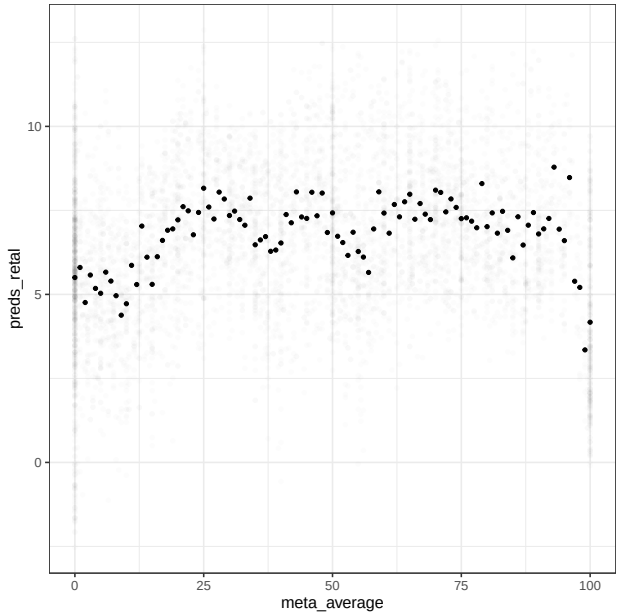
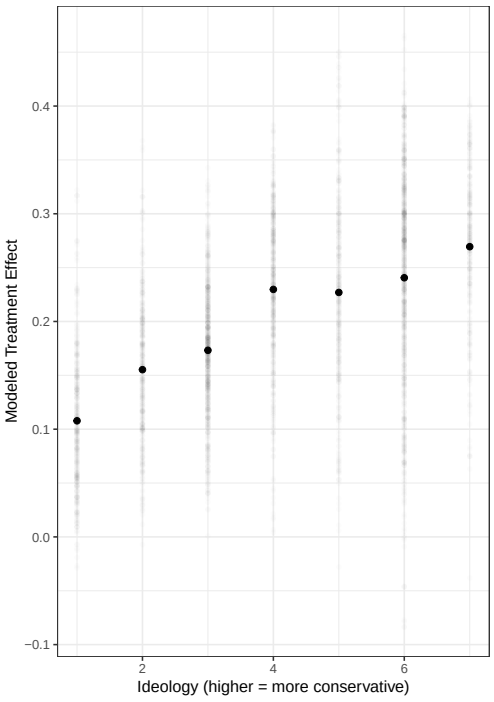
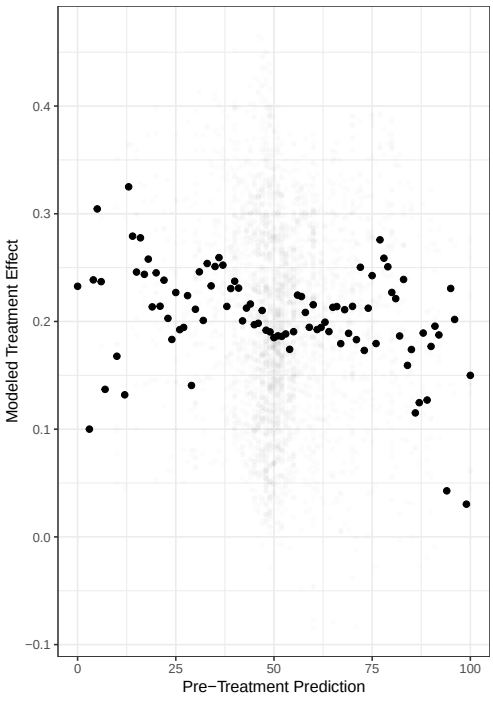


Figure A7: Heterogeneous effects of violation on retaliation outcome by meta-perceptions

1.9.2 Experiment 2



(a) Heterogeneous effects by Numerical Ideology (Higher = More conservative)



(b) Heterogeneous effects by Vote Share Prediction

Figure A8: Causal Forest Estimates of Heterogeneous Effects for Retaliation Predictions

1.10 Retaliation Expectations from Prior Experiments

As an alternative benchmark, I draw on prior experiments in which respondents were randomly exposed to information about opponents' anti-democratic behavior or were given corrections to exaggerated meta-perceptions (beliefs that the opposing party's voters reject democracy). While these earlier studies measured an immediate attitudinal response rather than long-term behavioral change, they identified a clear causal effect. My prediction tasks similarly asked respondents to estimate a causal parameter regarding real-world behavior. Because these past experiments do not match my exact set of norm violations, I aggregate their results across multiple studies and present these as an alternative to the benchmarks in the main text.

Studies measuring survey-based retaliation vary in form, ranging from observed violations **?**, to updating about mass-level support for violations from opposing partisans (Braley et al., 2023; Mernyk et al., 2022; Druckman et al., 2023; Christensen et al., 2025). I briefly summarize these results in the table below and offer a simple, meta-analytic estimate of their results. Some of the studies offer multiple treatment effect estimates and those are presented separately, though I include index outcomes as one treatment effect estimate rather than dis-aggregating them by individual outcomes. In these experiments, I treat the most optimistic condition (either the treatment arm without a provocation or the treatment condition for meta-perception corrections for papers studying second-order beliefs) as the baseline and explore by how much respondent's willingness to violate democratic norms increases in the more pessimistic condition. For example, the baseline is then the treatment group in (Braley et al., 2023), but is the control group in Janssen et al. (2025).[25]

I express these outcomes as: increased support for democratic backsliding, as a share of support in the control group. This measure is required because the underlying data and the standard deviation estimates are not available for the most recent working papers. Similarly, I weight by sample size because precision metrics are not available for

[25]What makes this comparison more challenging for the papers assessing meta-perception changes is the magnitude of the correction is inconsistent, while the papers exploring the effects of informational updating about a real-world violation are most substantively relevant to my theoretic context

the more recent working papers among this set of studies.

Table A23: Summary of Provocation Findings

Study	Treat	Baseline	Optimism	ATE as Share of Control	N
Freitag et al (2025)	Correction	0.25	0.19	31.5%	2188
Braley et al (2023)	Correction	0.24	0.17	41.1%	2645
Lupu et al (2025a)	Provocation	37%	24%	54.2%	1494
Lupu et al (2025b)	Provocation	38%	31%	22.5%	1494
Janssen et al (2025a)	Provocation	22.4%	11.5%	94.8%	3202
Janssen et al (2025b)	Provocation	25.2%	10.9%	131.1%	3276
Precision-Weighted average	Provocation	25.2%	10.9%	131.1%	3276

The precision-weighted average[26] of these existing estimates is 71.7% of the results in optimistic condition. This provides another benchmark against which to compare my estimates. Of note, none of my estimates exceed 20% of the estimates, suggesting wide under-estimates, which are even larger compared to the maximum possible retaliation, which ranges from 63% to 70%. These results suggest that expectations of retaliation are both (1) lower than reasonable benchmarks and (2) far from their theoretical limit. One limitation of this comparison is that these studies measure a micro-foundation for retaliation (greater support for anti-democratic actions among co-partisans) rather than observing or making predictions about actual behaviors, as my elite sample-based benchmark attempts.

1.11 Further Exploring Meta-Perceptions

At several points in this paper, I referenced the beliefs that are alternatively described in the literature as second-order beliefs or meta-perceptions: the attitudes that respondents attribute to opposing partisans regarding democratic norms. My main finding from Experiment 1 was that meta-perceptions were orthogonal to retaliation predictions. In other words, while beliefs about opposing partisans at the mass level were predictive of pessimism about the opposing party's actions, they did not predict the extent to which respondents believed the opposing party would behave conditionally. This null interac-

[26]I weight by sample size because standard errors and raw data are unavailable for some of these analyses

tion effect is robust to a number of alternative specifications as well as machine-learning approaches for detecting treatment effect heterogeneity.

Here, I again examine the relationship between meta-perceptions and predictions by showing how they shape beliefs about the opposing party in the real world in my final gerrymandering warning experiment.[27] Table A24 shows that meta-perceptions do not predict retaliation predictions in the control group because they are correlated with higher expectations of gerrymandering both with and without provocation. Confirming the findings from my first experiment, meta-perceptions appear wholly orthogonal to retaliation predictions even as they predict negative expectations about opponents. However, negative meta-perceptions do modestly blunt the effects of treatments: for every 10 percentage points of opposing partisans that a respondent believes hold anti-democratic views, there is a 1.66 percentage point reduction in updating about retaliation.

Table A24: Interactions with Meta-Perceptions in Final Warning Experiment

	Retaliation Predictions	Provoked	Un-Provoked
Warning	30.132***	11.485***	−18.647***
	(2.897)	(1.970)	(2.561)
Meta-Perceptions	0.029	0.277***	0.248***
	(0.055)	(0.038)	(0.049)
Warning:Meta-Perceptions	−0.151+	−0.165**	−0.014
	(0.079)	(0.054)	(0.070)
Num.Obs.	1886	1886	1886
R2	0.187	0.062	0.182
Std.Errors	IID	IID	IID

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Models include demographic covariates

1.12 Details of Elite Interviews

Before the qualitative portion of each interview, I asked respondents to estimate 5 causal quantities, the effect on the other party's behavior of a potential provocation, with the substantive issues presented shifting over time as circumstances changed. The pool of

[27] Of note, this final experiment treats retaliation predictions as a single outcome so cannot causally estimate retaliation predictions

questions is below. Notably, these questions are not randomized, I am directly asking the elites to estimate the causal quantities.

1. Committees
    - How likely is it that OUTPARY strips members of INPARTY of congresssional committee assignments?
    - Now consider if INPARTY first strips OUTPARTY of congressional committee assignments. Now how likely is it that OUTPARY strips members of INPARTY of congresssional committee assignments?
2. Filibuster
    - How likely is it that OUTPARTY will abolish the filibuster next time they are in a position to do so?
    - Now consider if INPARTY first abolishes the filibuster. Now how likely is it that OUTPARTY will abolish the filibuster?
3. Gerrymandering
    - How likely is it that OUTPARTY will gerrymander states X
    - Now consider if INPARTY first gerrymanders state Y. Now how likely is it that OUTPARTY will gerrymander state X?
4. Violence
    - How likely is it that over the next four years OUTPARTISANS will engage in an fact of serious violence where at least 10 members of INPARTY are seriously injured or killed?
    - Now consider if INPARTY first engages in such an act of violence. Now how likely is it that OUTPARTISANS will engage in an fact of serious violence where at least 10 members of INPARTY are seriously injured or killed?
5. Arrests
    - How likely is it that over the next four years OUTPARTISANS will arrest a prominent sitting politician on INPARTY and charge them without evidence?
    - Now consider if INPARTY first attempts such a step (at state level if Dem respondent). How likely is it that over the next four years OUTPARTISANS will arrest a prominent sitting politician on INPARTY and charge them without

evidence?

6. Media
    - How likely is it that over the next four years OUTPARTISANS will launch investigations of a media outlet that supports INPARTY?
    - Now consider if INPARTY first attempts such a step (at state level if Dem respondent). How likely is it that over the next four years OUTPARTISANS will launch investigations of a media outlet that supports INPARTY?

1.13 Pre-Analysis Plans

Pre-analysis plans for the studies are linked below. In some cases, the PAPs are embargoed until September 30th, 2025 after which they will be publicly available.

- Experiment 1: Estimating Retaliation Predictions
- Experiment 2: Estimating Retaliation Predictions
- Retaliation Experiments