The Subversion Dilemma: Why Voters Who Cherish Democracy Participate in Democratic Backsliding

Alia Braley* Gabriel S. Lenz† Dhaval Adjodah‡ Hossein Rahnama§
Alex Pentland¶

December 08, 2021

Abstract

Around the world, citizens are voting away the democracies they claim to cherish. How is this possible? In this article, we examine whether they do so because they fear that, if they don't, their opponents might do so first. In an observational study (N=1,973), we find that US partisans who most fear the other party's willingness to break democratic norms are also those most willing to support breaking these norms themselves. In an experimental study (N=2,543), we use an intervention to reduce these often exaggerated fears. Less fearful partisan then become more supportive of democratic norms. They also become more willing to vote against candidates of their own party who are willing to break these norms. The findings suggest that we can foster democratic stability by strengthening trust in opposing partisans' commitment to democracy.

*Travers Department of Political Science, University of California, Berkeley; Berkeley, CA 94720-1950. Corresponding Author. Email alia_braley@berkeley.edu

†Travers Department of Political Science, University of California, Berkeley; Berkeley, CA 94720-1950.

‡MIT Connection Science, Massachusetts Institute of Technology; Cambridge, MA 02139-4301.

§Media Lab, Massachusetts Institute of Technology; Cambridge, MA 02139-4301.

¶Media Lab, Massachusetts Institute of Technology; Cambridge, MA 02139-4301.

> *"Democrats are shredding the norms and institutions of our democracy with their full-blown resistance to @realDonaldTrump."* - Ronna McDaniel tweeting as @GOPChairwoman May 31, 2019

> *"Well, it's official: Republicans are now arguing that the US isn't (& shouldn't be) a democracy. This is what they believe. From lobbyists writing their bills to sabotaging our civil rights, the GOP works to end democracy."* - Alexandria Ocasio-Cortez tweeting as @AOC August 27, 2019

Around the world, antidemocratic leaders are convincing their supporters to vote away the democracies they claim to cherish. Citizens in Venezuela, Turkey, and Hungary strongly endorsed democracy while casting votes for authoritarian leaders Chávez, Erdoğan, and Orbán respectively (*1*, *2*). How is this possible?

In this paper, we suggest a simple explanation. The norms that support democracy must generally be supported by all sides. If partisans fear their opponents are preparing to defect from these norms, they will want to subvert them first. What might lead to such fears? Scholars have observed that would-be authoritarian leaders often attempt to foster these fears among their supporters (*1*). Once they convince their supporters to be suspicious of the other side's commitment to democracy, they find their supporters more willing to tolerate their own antidemocratic actions. They can claim that their own antidemocratic actions are merely leveling the playing field. Once a would-be authoritarian leader takes this course, thier antidemocratic actions naturally set off fear in opposing partisans, placing them in the same position in which they will tolerate antidemocratic actions by their leaders in an attempt to level the playing field. Once these mutual exaggerated fears take root, they can be self-reinforcing and highly destructive, as each side has an incentive to preemptively defect from democratic norms before the other does.

We call this predicament the "subversion dilemma." It has parallels in the "security dilemma" in the realist view of international relations (*3*, *4*). Realists attribute the onset of interstate war to the global state of "anarchy" in which states are not subject to an external legal enforcement system. The lack of external protections incentivizes State A to arm itself against potential attackers. State B is likely to interpret this military buildup as a potential threat and is likely to increase its

own military spending. State A reacts in kind. This cycle may continue until one state decides to preemptively attack rather than risk annihilation. Critically, in this account, warfare breaks out despite states universally preferring to avoid war.

Unlike laws, the norms that support democracy within a state also lack external legal enforcement. Just as states may go to war even though they prefer peace due to the slippery slope of the security dilemma, citizens may lose their cherished democracies due to the slippery slope of the subversion dilemma.

In the US context, the subversion dilemma dynamic now appears to be at play. Donald Trump stoked fears about Democrats subverting democracy. Early in his 2016 campaign, his website stated: "Help Me Stop Crooked Hillary from Rigging this Election!"(*1*). Throughout the 2016 campaign, he repeated, "This is a rigged election"(*5*). These accusations continued through the 2020 election, placing Trump's democracy-loving supporters in the subversion dilemma. In the 2020 election, Fox News amplified this message, repeatedly proclaiming the existence of "an all-out effort to depress and suppress the pro Trump vote"(*6*). They accused Democrats of manipulating ballots, for instance, by discarding Republican ballots in a ditch in Wisconsin (*7*). At the very same time, they cast Trump as the candidate "trying to protect democracy"(*8*).

Consistent with the logic of the subversion dilemma, Republican rhetoric prompted mainstream and left-leaning news media outlets to express concern that Republicans would bend the rules to help Trump win. In the 2020 election, for instance, CNN portrayed Trump as "manipulating the ballots" and planning to "disregard the popular vote" (*9*, *10*). During the 2020 election, the dilemma that Democrats faced is evident in a tweet by Senator Elizabeth Warren on September 26th of that year: "Health care. Reproductive Freedom. Workers' rights. Dreamers' futures. Our planet. Democracy. Everything is on the line—so everything is on the table" (*11*).

The subversion dilemma account, we suggest, is one of several contributing factors to democratic backsliding (*12*). No doubt, would-be authoritarians generate distrust in other ways, such as accusing opposing partisans of criminal behavior such as drug-trafficking. Would-be authoritarians also likely find more receptivity to their messages in societies that are already highly politically polarized, and the US has become such a society over the last several decades (*13*, *14*). This polarization likely contributes to a range of negative and exaggerated misperceptions Democrats

and Republicans have about each other (*15–19*), including the misperceptions we document in this article.

Across three studies, we test the core claim at the heart of the subversion dilemma: voters will be willing to subvert democracy to the degree they fear the other party is willing to subvert it. Study 1 uses survey data to examine this claim observationally. Studies 2a and 2b then test this claim experimentally and show that reducing people's fear of the other side leads people to support democratic norms and to vote against same-party candidates who violate them.

Study 1: Fear and Support for Backsliding

To show that fear of the other side drives people to support subverting democracy, we survey a demographically representative sample of 1,973 US residents recruited via Lucid between July 15 and August 6, 2021. We exclude respondents who lack a partisan identification or do not lean towards one of the parties.

To assess partisan willingness to subvert democratic norms, we present participants with seven scenarios where they must choose between an action that benefits their own party at the expense of a democratic norm versus an action that upholds a democratic norm. We select seven norms that are shown to be important to the American public and often undermined in instances of democratic backsliding (*20–22*).

For instance, to assess a Democratic respondent's fear of the other party's willingness to subvert, we ask: "Do you think that MOST REPUBLICANS would support reducing the number of voting stations in towns that support DEMOCRATS?" We ask about "most" opposing partisans rather than opposing partisan elites because our theory is that partisan beliefs about opposing partisan willingness to act as a check against democratic backsliding is an important determinant of one's own willingness to do so. By contrast, we expect beleifs about opposing partisan elites to be more likely to elicit expressive responses and less likely to be subject to change.

To assess the Democratic respondent's own willingness to subvert democratic norms, we ask: "Would YOU support reducing the number of voting stations in towns that support REPUBLICANS?" The other six scenarios ask about banning rallies, ignoring controversial court rulings,

freezing the social media accounts of journalists, changing laws to make it easier for one's own side to get elected, using violence to block laws, and reinterpreting the Constitution to block policies. For each scenario, participants respond on a four-point Likert scale with options: "Never," "Probably Not," "Probably," and "Definitely." In our analyses, we take the simple average of these seven questions and rescale it 0-1.

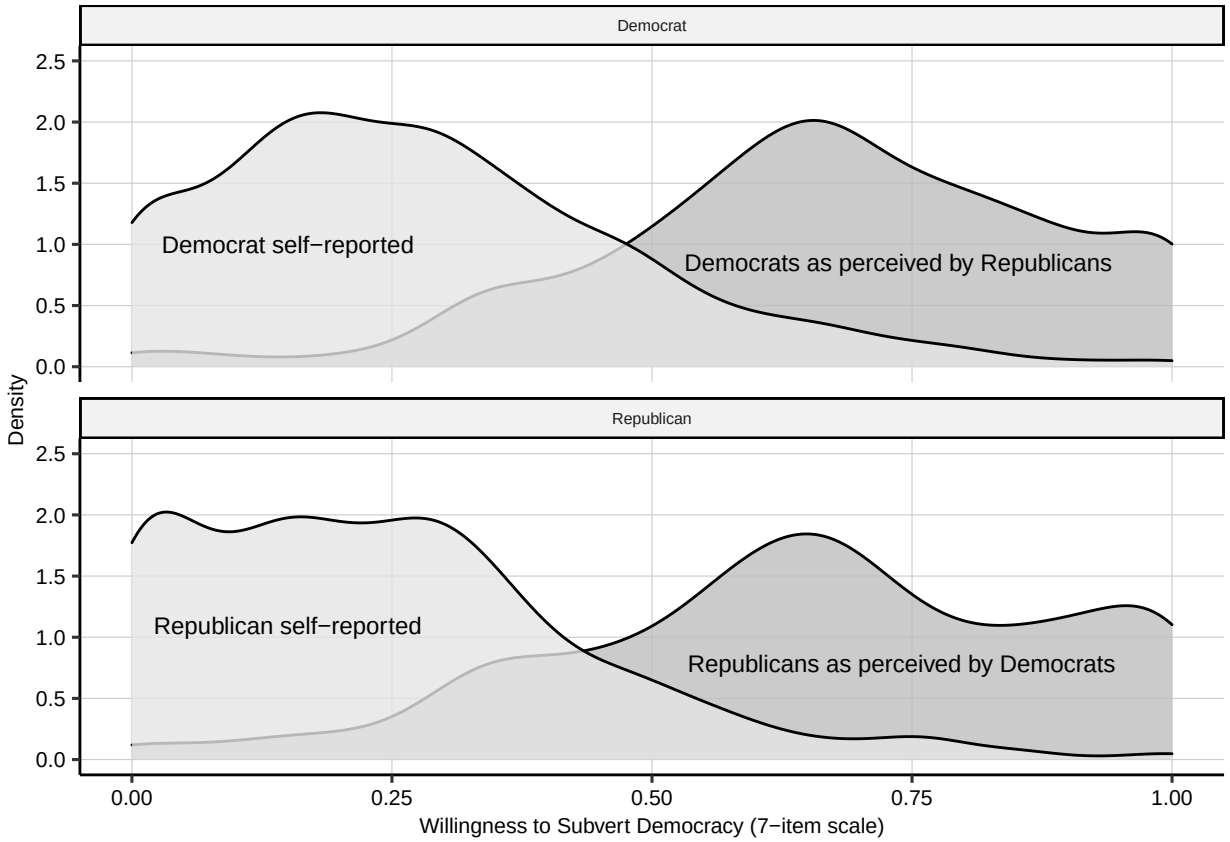


Figure 1: Exaggerated Mutual Fears in Study 1. This figure plots the distributions of the unweighted average of the seven questions we ask respondents about their their perception of opposing partisan willingness to subvert democracy and their own willingness to subvert democracy. The questions vary from reducing polling places in the other party's towns to other-party banning rallies. The figure shows that members of both parties appear to significantly overestimate opposing partisans' willingness to break democratic norms.

Figure 1 presents the distributions of responses to this seven-item index by Republicans and Democrats, respectively. While Democrats believe that Republicans will be willing to subvert democratic norms in 5.2 of the scenarios (mean = 0.67 on the 0-1 scale), Republicans self-report willingness to subvert these norms in only 3.4 of these scenarios (mean = 0.28). Republicans

similarly believe that Democrats will be willing to subvert democratic norms in 5.0 of the scenarios (mean = 0.65) while Democrats self-report willingness to subvert these norms in only 3.1 of the scenarios (mean = 0.28). Supplemental Materials Figures S1 and S2 show the distribution of self-reported willingness to subvert for each item and for each party, respectively.

Although this perception gap seems consistent with the logic of the subversion dilemma, other explanations are possible. For example, people may report exaggerated fears as a form of expressive responding—they're upset at the other party and look for any opportunity to express negative sentiments about the other party on the survey. Nevertheless, it's worth noting that the perception gap we find is significantly larger than the perception gaps documented between partisans on issues of ideological and affective polarization (*23*, *24*). On a 0 to 1 scale, we find that partisans inaccurately perceive opposing partisans by an average of 0.09 on policy views, and 0.14 on dehumanization of opposing partisans, while they inaccurately perceive the other side's willingness to subvert democracy *by 0.4.*

The core claim we test in this paper is that fear of the other side subverting democracy leads partisans to support subverting it themselves. Using the two multi-items scales described above, Figure 2 reveals precisely this pattern, showing a strong linear relationship between perceptions of the other side's willingness to subvert democracy and partisan's own willingness to do so. Compared to respondents who don't believe the other party desires to undermine democracy at all (0 on the scale), respondents who believe the other side is fully willing to undermine democracy (1 on the scale) increase their own willingness to undermine democracy by about .25 points (on the 0-1 scale).

To address alternative explanations for this relationship, we statistically adjust for factors that may influence both variables with regression (we further rule out alternative explanations with an experiment in the next study). Since polarization along party and ideological lines could influence both variables of interest, we control for political knowledge, partisan identity strength, extremism of policy views, dehumanization of opposing partisans, and the difference between feeling thermometers for opposing partisans versus copartisans. We also control for two sets of beliefs about opposing partisans: beliefs about how extreme their policy views are and beliefs about how much they dehumanize one's own party members (*20*, *23*–*27*). To measure these constructs,

we include survey items about each one, including multi-item scales about policy views and about perceptions of the other party's policy views.

In Table 1, we control for these variables and for demographics in a series of regression models. These controls leave the association we find in Figure 2 largely unchanged for Democrats and for Republicans. These findings suggest that at least this set of variables cannot account for the strong relationship between fear of the other side subverting and one's own willingness to subvert.

In other analyses, we examine a wide range of potential confounding variables, including ethnicity, employment, household income, household income compared to area mean, perceived economic status, region, urbanity, gini index by zip code, recent economic growth by zip code, religion, big-five personality traits, self-efficacy, general beliefs about the role of government, primary news source, and exposure to partisan news. We also find that the inclusion of these variables in our regression leaves the main variables of interest largely unchanged.

Study 2a: Experiment: Correcting Misperceptions Support for Democratic Norms

We causally test our core claim by experimentally manipulating perceptions of the other sides' willingness to break democratic norms. We examine whether, once we have reduced fears of the other side, partisans become more supportive of democracy. To correct partisans' exaggerated perceptions of opposing partisans' willingness to subvert democratic norms, we use an "ask-tell" design, which scholars have successfully employed to correct misperceptions between partisans in other studies (*15*, *18*, *23*).

In the treatment and control groups, we ask participants the same seven questions that we use in Study 1 to measure beliefs about opposing partisans' willingness to break democratic norms. In the treatment group, after participants answer each question, we tell them how most opposing partisans actually answered the question using data from Study 1 (hence "ask-tell"). Democrats in Study 1 stated that they would "Never" support breaking four of the democratic norms we presented them with and would "Probably Not" support breaking three of them, while Republicans

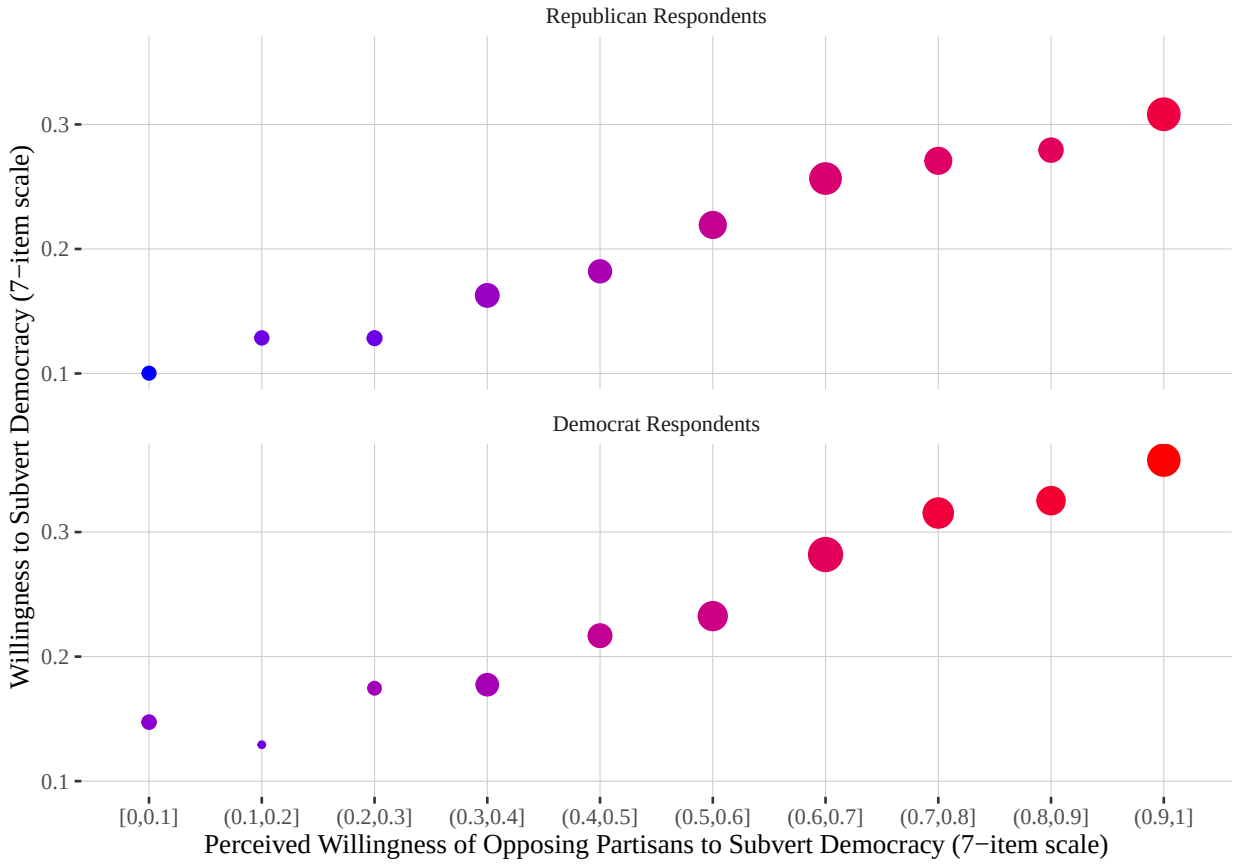


Figure 2: You Subvert, I Subvert—Study 1. This figure shows that partisans support breaking democratic norms more when they believe that opposing partisans are willing to break democratic norms. See the previous Figure 1 note for more details on the seven scenarios we asked respondents about. Point size reflects the number of respondents and coloration reflects willingness to subvert.

Table 1: Explaining Willingness to Subvert Democracy in Study 1. All variables are coded 0-1 and rescaled so that higher values are associated with greater willingness to subvert democracy. This table shows that perceived willingness of opposing partisans to subvert democracy continues to explain respondents own willingness to subvert democracy even when controlling for possible confounders. See the next study for an experiment. Ordinary least-squares regression estimates showing unstandardized coefficients. Each column shows estimates from a separate regression model.

	Dependent variable:			
	Willingness to Subvert Democracy			
	Republicans		Democrats	
	(1)	(2)	(3)	(4)
Perceived Willingness of Opposing Partisans to Subvert Democracy	0.24***	0.23***	0.28***	0.25***
	(0.03)	(0.03)	(0.03)	(0.03)
Partisan Identity Strength		0.04*		0.01
		(0.02)		(0.02)
Age		−0.05		−0.20***
		(0.03)		(0.03)
Education		0.02		−0.02
		(0.02)		(0.02)
Policy Extremism		0.05		−0.01
		(0.04)		(0.04)
Dehumanization of Opposing Partisans		−0.14***		−0.07***
		(0.02)		(0.02)
Perceived Policy Extremism of Opposing Partisans		−0.20***		−0.07*
		(0.04)		(0.03)
Perceived Dehumanization by Opposing Partisans		0.09***		0.05**
		(0.02)		(0.02)
Feeling Thermometer Reps Minus Dems		0.02		−0.12*
		(0.05)		(0.05)
General Political Knowledge		−0.05*		−0.003
		(0.02)		(0.02)
Constant	0.08***	0.26***	0.09***	0.29***
	(0.02)	(0.04)	(0.02)	(0.04)
Observations	907	907	1,016	1,016
R^2	0.09	0.19	0.09	0.17
Adjusted R^2	0.09	0.18	0.09	0.16
Residual Std. Error	0.19	0.18	0.19	0.18

Note: Standard Errors in Parentheses
* p<0.05; ** p<0.01; *** p<0.001

in Study 1 indicated that they would "Never" support breaking five of the democratic norms we presented them with and would "Probably Not" support breaking two of them (see Supplemental Materials Figures S1 and S2). We ask the control group to answer the same questions about opposing partisans, though no feedback is provided. After administering two attention checks and determining political party (we drop true Independents), we administer the ask-tell treatment and then ask about respondents' own willingness to subvert democratic norms using the same seven item questionnaire. As in Study 1, we take the simple average of the seven items and rescale them 0-1. This experiment is a preregistered replication of a preregistered pilot experiment that we conducted September 15-29, 2021 on a representative sample of 2,543 respondents recruited through Lucid (see Supplemental Materials Figures S7 and S8 for key pilot findings).

The ask-tell treatment succeeds at changing reported perceptions, lowering perceptions that opposing partisans are willing to break democratic norms. Figure 3 shows the distribution of these perceptions of opposing partisans for treatment and control conditions by party. Similar to our findings in Study 1, participants in the control condition place opposing partisans' willingness to subvert democratic norms at 0.64 on the 0-1 scale across the seven scenarios. By contrast, participants in the treatment condition place them at 0.4, a highly significant difference (p=7.5e-149). Since the ask-tell treatment is administered across seven scenarios, these results likely underestimate the manipulation effect, since respondents aren't fully treated until they've received feedback on the final question. Indeed, when we look at responses only for the final (randomized) question, we see a larger treatment effect with 0.64 for control and 0.33 for treatment.

This successful manipulation does, in fact, appear to increase support for upholding democratic norms. The treatment group is less willing to subvert democracy than the control group. Figure 4 shows this result, plotting the distribution of the willingness to subvert scale for those in the treatment versus control group. The average participant in the treatment group becomes less willing to subvert democratic norms, shifting from a mean of 0.24 to 0.17 on the 0-1 scale, a 29% change, and one that is highly unlikely to occur by chance (p=3.5e-22, calculated from robust standard errors). Another way to understand this finding is that respondents in the control condition say they would "never" support breaking democratic norms in 3.5 of the seven scenarios in the control group - a number that increases to 4.7 in the treatment group. When we scale the treatment effect by

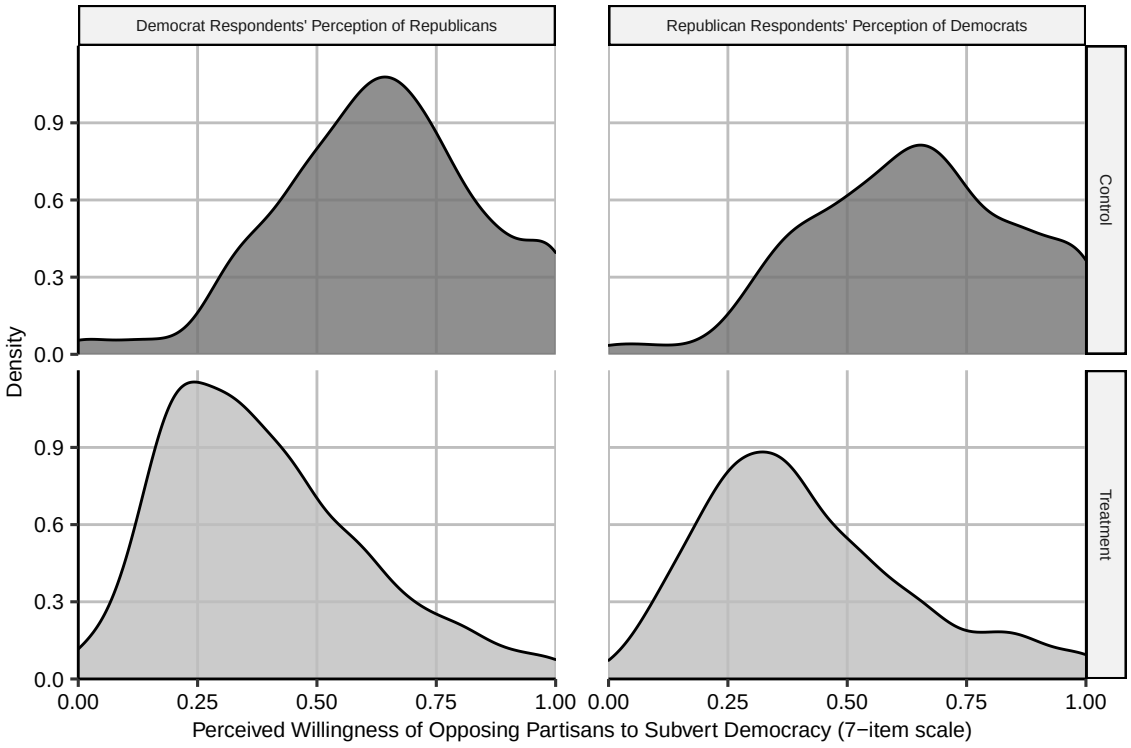


Figure 3: Correcting Misperceptions—Experimental Manipulation Check for Study 2. This figure shows that, when we inform respondents about opposing partisans' actual support for subverting democracy in each of the seven scenarios, they became less fearful of opposing partisans' willingness to subvert democracy. Since we give respondents feedback after they answer each of the seven questions, this figure likely understates the extent of the manipulation effect, since respondents don't receive the full treatment until they receive feedback on the seventh item.

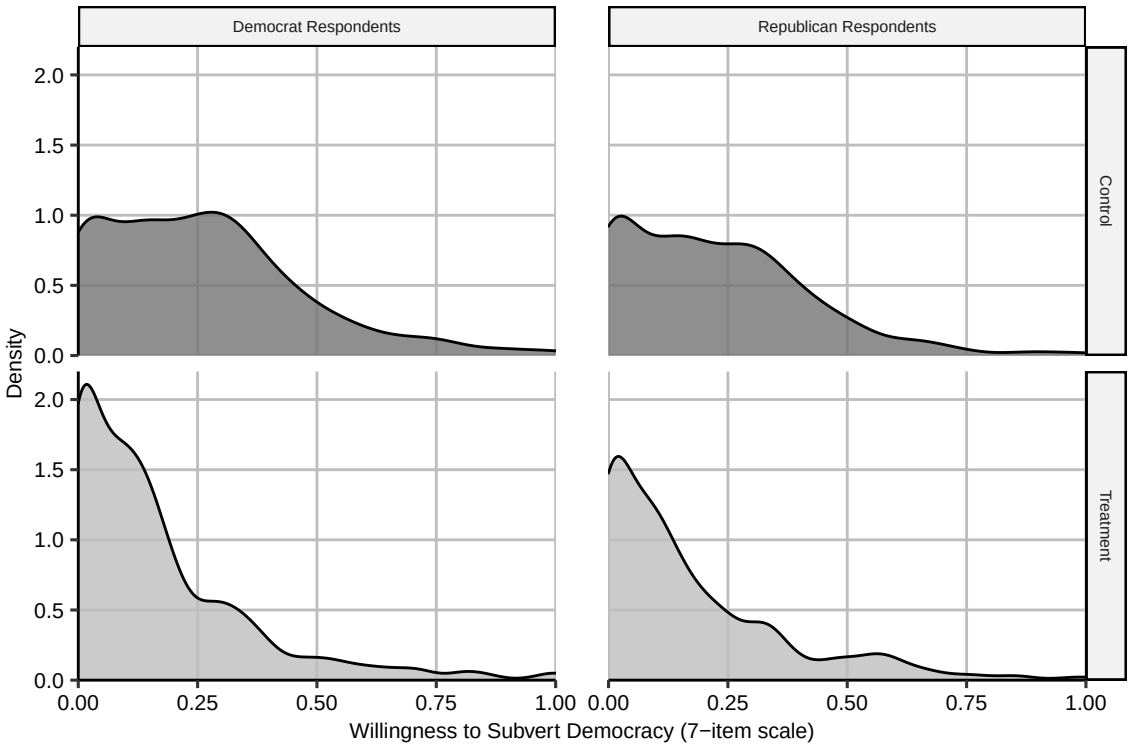


Figure 4: Less Fear, More Support for Democratic Norms—Experimental Effect for Study 2a. The informational intervention that lowers fears that opposing partisans will break democratic norms decreases participants' own willingness to subvert democracy.

the degree of successful manipulation (complier average causal effect) using the final (randomized) question we ask both groups, the estimate closely matches the regression estimate from Study 1: 0.25 (see Supplementary Materials for details and assumptions).

We also find a statistically significant treatment effect for all seven individual items in the seven-item self-subversion scale. All seven have a similar magnitude except for the violence item, which is noticeably more resistant to change (see Supplemental Materials Figure S3).

If fear of the other sides' willingness to dismantle democracy is central to current politics, the treatment should have a wide range of effects on political attitudes. Consistent with this prediction, we find that the informational intervention also increases partisans' positive feelings about each other and decreases the perception that the other party dislikes them (Supplemental Materials Figure S4). We don't find that this policy changes people's own policy views (5-item scale), but it may have reduced Democrats' perception that Republicans are extremists on a number of key policy issues. We also find that the effect appears similar across demographic categories, such as age, except for general political knowledge, where the treatment may have a larger effect on less knowledgeable respondents (see Supplemental Materials Figure S5). We find no sign that respondents were able to guess the study's purpose based on an open-ended question at the end of the survey.

Study 2b: Correcting Misperceptions and Voting Behavior

Does lowering fear that opposing partisans will subvert democratic norms—as the informational intervention appears to do—translate into behaviors that can alter the political landscape? We find that it may shape citizen voting in a way that could limit democratic backsliding.

To examine the consequences for voting decisions, we ask respondents to vote in two hypothetical primary elections. We then examine the impact of the ask-tell intervention on these voting decisions. Unlike the previous analyses, this analysis is exploratory (not preregistered). In each hypothetical primary race, respondents face a choice between two candidates from their own party, one of whom has "supported" breaking one of the democratic norms among the seven scenarios (chosen at random) and another who has "opposed" breaking the norm.

In order to model the real-world rhetoric that we contend produces democratic backsliding, we tell participants that the candidate who "supported" breaking a democratic norm did so because they believe opposing partisans (Democrats or Republicans) have done the same.

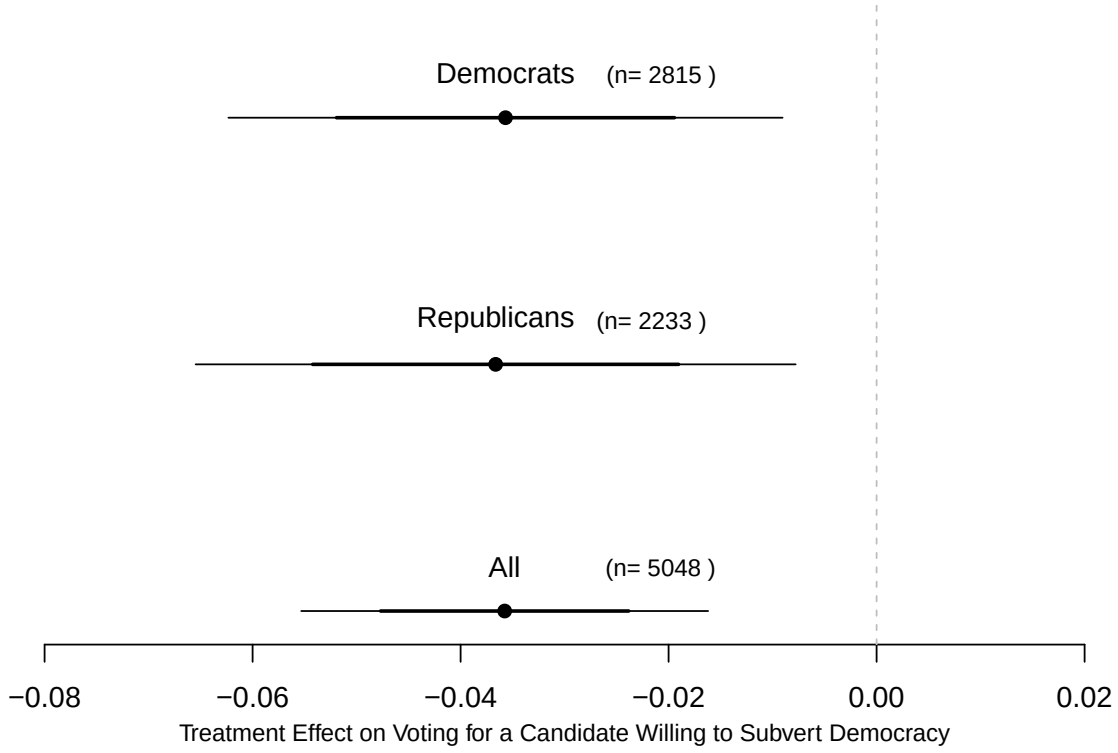


Figure 5: Less Fear, Less Tolerance for Candidates Who Subvert Democracy—Study 2b. When we expose respondents to information about how few members of the other party support subverting democracy, respondents become less willing to vote for a primary candidate from their own party who supports subverting democracy compared to the control group. We code the dependent variable as 0 voting for democracy promoting candidates, 0.5 as voting for neither, and 1 as voting for democracy subverting candidates. The mean on the scale is around .24 for both parties. The Ns represent the total number of votes cast. Since respondents cast votes in two primary races, the number of respondents is half the N (though some respondents only cast a vote in one primary). Each coefficient is from a separate regression and we cluster the standard errors by respondent.

We find that the "ask-tell" treatment decreases partisan willingness to vote for candidates who support breaking democratic norms by about 0.04 on a scale from 0-1 where voting for a democracy protecting candidate is coded as 0, voting for neither is coded as 0.5, and voting for a democracy subverting candidate is coded as 1. The effect is highly statistically significant (p=0.0027, calculated from robust standard errors clustered at the respondent level), as Figure 5 shows. This 0.04 effect represents a shift from 0.26 in control to 0.22 in the treatment group, a 15% change. Republicans

and Democrats exhibit almost identical responses to the treatment with Democrats falling from 0.27 in control to 0.23 in treatment, and Republicans falling from 0.25 to 0.21. The complier average treatment effect is about 0.14.

Discussion

In a preregistered observational study and a pregistered experimental study, we find that partisans support subverting democracy to the degree that they fear the other side supports subverting it. We also find exploratory evidence suggesting that correcting these fears may translate into behavioral outcomes, such as voting for democracy-promoting candidates.

Notably, we find that partisan willingness to break democratic norms correlates noticeably more with the degree of fear about the other side's willingness to undermine democracy than with other usual suspects of democratic backsliding including strength of partisan identity, extreme policy preferences, the belief that the other side holds extreme policy preferences, partisan antipathy, dehumanization of the other side, belief that the other side dehumanizes oneself, political knowledge, and a range of demographic and geographic variables.

Taken together, these findings suggest an answer to the question of why voters may vote away the democracies they cherish. They may do so in part because they fear their opponents are already doing so.

This account helps make sense of patterns we observe during backsliding. In particular, it may explain why would-be authoritarians accuse their opponents of undermining democracy—it increases social tolerance for the backsliding they desire. It may also explain how some countries avoid the complete dismantling of their democracies: by taking actions that reinforce democratic institutions, even in the face of would-be authoritarian opponents (*1*). By taking such actions, a party may change opposing partisans' perceptions of them, convincing them that they support democracy, and taking them out of the subversion dilemma. It also suggests that Democrats need to think hard about pursuing policies that Republicans perceive as tilting the playing field and contributing to an arms race, such as DC statehood or eliminating the filibuster.

Since beliefs about opposing partisans' willingness to break democratic norms appear amenable

to change through a simple informational intervention, informational campaigns may have the potential to stem democratic backsliding. Interventions that directly address the subversion dilemma may also have potential.

As with the "security dilemma" in international relations, the lack of an external enforcement mechanism leaves democratic norms fragile to intergroup threat dynamics. Research in international relations suggests that external sources of third-party observation along with costly signals of good-faith are possible ways to resolve this dilemma (*4*). As one example, when we asked participants in Study 1 whether they would make a one-to-one pact with a member of the other party to never vote for a candidate that subverts democracy, 66% said "probably" or "definitely yes."

This paper provides a novel framework for understanding why democratic loving citizens can vote away their democracies. Undoubtedly, polarization and other factors studied by researchers contribute, but the subversion dilemma may help further explain democratic backsliding. It also points the way to interventions that may help those that cherish democracy to prevent themselves from falling victim to the slippery slope of the subversion dilemma.

References and Notes

1. S. Levitsky, D. Ziblatt, *How democracies die* (Broadway Books, 2018).

2. M. Svolik, When polarization trumps civic virtue: Partisan conflict and the subversion of democracy by incumbents. *Available at SSRN 3243470* (2018).

3. B. R. Posen, The security dilemma and ethnic conflict. *Survival*. **35**, 27–47 (1993).

4. J. D. Fearon, Rationalist explanations for war. *International organization*. **49**, 379–414 (1995).

5. D. J. Trump, (2016), (available at https://www.theguardian.com/us-news/video/2016/oct/18/us-presidential-election-rigged-donald-trump-wisconsin-video).

6. L. Ingram, *Fox News*. **September 9** (2020).

7. T. Carlson, *Fox News*. **September 24** (2020).

8. L. Ingram, *Fox News*. **October 1** (2020).

9. E. Burnett, *CNN*. **September 8** (2020).

10. D. Lemon, *CNN*. **September 23** (2020).

11. E. Warren, (2020), (available at https://twitter.com/ewarren/status/1309876174231949312).

12. D. Waldner, E. Lust, Unwelcome change: Coming to terms with democratic backsliding. *Annual Review of Political Science*. **21**, 93–113 (2018).

13. N. McCarty, *Polarization: What everyone needs to know* (Oxford University Press, 2019).

14. S. Iyengar, Y. Lelkes, M. Levendusky, N. Malhotra, S. J. Westwood, The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*. **22**, 129–146 (2019).

15. D. J. Ahler, Self-fulfilling misperceptions of public polarization. *The Journal of Politics*. **76**, 607–620 (2014).

16. M. S. Levendusky, N. Malhotra, (Mis) perceptions of partisan polarization in the american public. *Public Opinion Quarterly*. **80**, 378–391 (2016).

17. A. M. Enders, M. T. Armaly, The differential effects of actual and perceived polarization. *Political Behavior*. **41**, 815–839 (2019).

18. D. J. Ahler, G. Sood, The parties in our heads: Misperceptions about party composition and their consequences. *The Journal of Politics*. **80**, 964–981 (2018).

19. J. Mernyk, S. Pink, J. Druckman, R. Willer, Correcting inaccurate metaperceptions reduces americans' support for partisan violence (2021).

20. M. H. Graham, M. W. Svolik, Democracy in america? Partisanship, polarization, and the robustness of support for democracy in the united states. *American Political Science Review*. **114**, 392–409 (2020).

21. J. M. Carey, G. Helmke, B. Nyhan, M. Sanders, S. Stokes, Searching for bright lines in the trump presidency. *Perspectives on Politics*. **17**, 699–718 (2019).

22. N. P. Kalmoe, L. Mason, in *National capital area political science association american politics meeting* (2019).

23. J. Lees, M. Cikara, Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nature Human Behaviour*. **4**, 279–286 (2020).

24. S. L. Moore-Berg, L.-O. Ankori-Karlinsky, B. Hameiri, E. Bruneau, Exaggerated meta-perceptions predict intergroup hostility between american political partisans. *Proceedings of the National Academy of Sciences* (2020).

25. E. C. Cassese, Partisan dehumanization in american politics. *Political Behavior*, 1–22 (2019).

26. M. H. Graham, Does partisan identity reduce support for electoral fairness? An experimental test (2020).

27. J. L. Martherus, A. G. Martinez, P. K. Piff, A. G. Theodoridis, Party animals? Extreme partisan polarization and dehumanization. *Political Behavior*, 1–24 (2019).

28. N. Kteily, G. Hodson, E. Bruneau, They see us as less than human: Metadehumanization predicts intergroup conflict via reciprocal dehumanization. *Journal of Personality and Social Psychology*. **110**, 343 (2016).

Acknowledgements

Funding: MIT Media Lab Discretionary Funds (HR, AP)

Research funds (GL)

Authors Contributions:

Conceptualization: AB

Methodology: AB, GL

Investigation: AB, GL

Visualization: AB, GL

Funding Acquisition: AB, GL, DA, HR, AP

Project Administration: AB, GL

Supervision: GL, DA, AP

Writing - original draft: AB, GL

Writing - review and editing: AB, GL, DA, HR, AP

Competing Interests: The authors declare that they have no competing interests.

Data and materials availability: All data needed to evaluate the conclusions of this paper are present in the paper and/or the supplementary materials. Replication data will be made available by the authors.

List of Supplementary Materials

Materials and Methods

Tables S1-S5

Figures S1-S8

Supplementary Materials

Materials and Methods

Survey Flow Study 1

After obtaining consent, we administer two attention checks and a party identification question using the American National Action Study (ANES) 7-point Party Identification Scale. Participants who fail an attention check or who are true Independents (do not lean towards either party) are excluded from the study.

We then ask participants in random order either the seven questions about their willingness to subvert democracy or the seven questions about the opposing party's willingness to subvert democratic norms (see question wording below). The results are similar regardless of the order in which the question blocks and individual questions are given. We then ask questions to capture alternative explanations as described in the text. From Lucid, we obtain demographic information on participants including age, gender, household income, ethnicity, education, region, zip code, and state.

In the analyses, we take the uweighted mean of all responses for the multi-item scales, we orient all data so that higher numbers indicate higher levels of the variable (i.e., policy extremism, dehumanization of the out-party), and we rescale all variables 0-1.

Demographic information about the participants in this study can be found in Table S1.

Experimental Intervention Flow Studies 2a and 2b

Study 2 follows much the same logic as Study 1. After obtaining consent, we administer two attention checks and a party identification question using the ANES 7-point Party Identification Scale. We end the survey for participants who fail an attention check or who are true Independents (do not lean towards either party).

Participants are then randomly assigned to either the treatment or control condition. After the consent and attention checks, we tell all participants: "Great Job! Studies have shown that

DEMOCRATS and REPUBLICANS in America DON'T KNOW MUCH about each other. Let's test how much YOU know!" After the party identification scale, we tell all participants: "Now you will be asked about average, everyday [DEMOCRATS/REPUBLICANS] in America. You will get feedback from REAL DATA. Based on how much you know - you will be placed in one of FOUR LEAGUES." We listed the leagues as "silver," "gold," "diamond," and "professional" next to corresponding images of medals.

After answering each question, we tell participants in the treatment condition how most opposing partisans actually answered the question using data from Study 1. All participants are then asked to answer a few more questions as we calculate their score. In order to measure our main DV, willingness to subvert democratic norms, participants are asked the same seven questions designed to measure their willingness to subvert democracy as in Study 1 (randomized).

In order to measure our exploratory DV, participants are asked how they would vote in two primary elections. In each case, the only information that participants are given is that one candidate "opposes" breaking one of the democratic norms featured in this study, and the other candidate "supports" breaking one of the democratic norms featured in this study "because they believe that [opposing partisans] have done the same." All norms and order of candidates are randomized for each respondent, and participants can vote for "Candidate A," "Candidate B," or "Neither."

The remainder of Study 2 measures the same coviariates as in Study 1 in the same manner, including extremism of policy views, beliefs about the extremism of opposing partisan policy views, blatant dehumanization of opposing partisans, the extent to which participants believe opposing partisans dehumanize them, participants' level of political knowledge, and standard feeling thermometer questions to measure level of positive or negative affect toward copartisans and opposing partisans. From Lucid, we obtain demographic information on participants including age, gender, household income, ethnicity, education, region, zip code, and state.

As in Study 1, in our analyses, we use mean values of all question batteries, we orient all data so that higher numbers indicate higher levels of the variable (i.e. policy extremism, dehumanization of the out-party), and we normalize all variables to 0-1 scales. Primary election voting is recorded as 0 if partisans vote for democracy-promoting candidates, 0.5 if partisans vote for neither, and 1 if

partisans vote for democracy-subverting candidates.

Demographic information about the participants in Study 2 can be found in Table S2. The magnitude and type of demographic differences do not seem to threaten the ability to make meaningful inferences about the main variables of interest.

Demand Effects

We have several reasons to believe that demand effects may not be driving the treatment effect we find in the experiment in Study 2. First, the individuals in the Lucid sample are typically used for marketing surveys and therefore would not expect to be part of an experiment. Second, we describe the study as trying to gauge how much Democrats and Republicans know about each other, as noted earlier. Third, the survey asked respondents many questions and it was not obvious to them what the most important items were. Fourth, we don't find any sign of awareness of the purpose of the experiment in an open-ended question we ask. In particular, we randomly selected 1,297 respondents to answer: "What do you think this study was mainly about? (Write one sentence)." A research analyst, Jacob Levy, then coded the following four questions based on their answers: (1) Does the respondent guess that the study was about correcting their misperceptions about how much the other party is willing to subvert democracy? (2) Does the respondent guess that the study was about changing their own willingness to subvert democracy? (3) Does the respondent guess that the study was about changing or altering their sentiment towards the other party? Finally, (4) does the respondent guess that the survey is about misconceptions/misperceptions/bias more broadly? The coder was blind to respondents' treatment assignment, but not to the purpose of the study. Of the 1,297 respondents, the coder concluded that none provided answers that met the first two criteria, 21 provided answers that met the third, and 35 provided answers that met the fourth. We found no substantive difference between the treatment and control groups rates of answering the latter two questions.

Of course, none of this definitively rules out demand effects. We also think it's likely that a form of demand effect increases learning about the other party's position in the treatment group beyond what we might expect in a non-survey environment.

Scale Construction and Missing Data

In constructing the multi-item scales, we take the average of the non-missing values for each respondent. Since we request responses when respondents attempt to skip a question, very few values are missing and only a handful of respondents have several values missing on any of the multi-item scales (see Tables S2 and Table S3 for descriptive statistics).

So that the regression estimates for Study 1 do not omit respondents because of missing values on control variables, we impute missing values using demographics (using the robust linear model function in the simputation package in R). We never impute values on the key independent or dependent variables (perceptions of the other party's subversion and self-reported subversion). The number of observations imputed on each control variable is small (typically 20-30) and the regression results are substantively identical when we do not impute.

Complier Average Treatment Effect

To scale the experimental estimates by compliance with the treatment so that we can compare them to the observational regression, we estimate the complier average causal effect (CACE). An assumption necessary here is that the treatment can only influence respondents through its effect on perceived willingness of out partisans to subvert, not through other observed or unobserved variables (the exclusion restriction). Table S4 presents the reduced form regression and Table S5 presents the Two-Stage Least-Squares (2SLS) estimates. In the 2SLS estimates, we instrument the final (seventh) "other party's willingness to subvert democracy" item respondents saw with the treatment indicator. We present these items in random order and use the final one because respondents in the treatment group would have mostly received the treatment by that time, as they would have received feedback on six of the seven items. We use the final item for the control group as well. (When we instead use the mean of the full-scale in the 2SLS, we find similar though slightly larger estimates because it underestimates compliance.) The 2SLS estimates reveal a CACE estimate that is strikingly close to the regression estimate from Study 1 (both around 0.25) and one that is precisely estimated. Since we measure many covariates posttreatment, we can include those in the 2SLS estimates to attempt to block those causal paths. When we include measures of

dehumanization, meta-dehumanization, the feeling thermometer difference for the parties, policy extremism, and meta-policy extremism, we find only the slightest decrease in the 2SLS estimate (from about 0.25 to 0.23 and the estimate remains very precisely estimated). Of course, these analyses don't rule out exclusion restriction violations on unobservables.

Preregistration and Findings from Pilot Study

We preregistered the studies here: https://osf.io/vnd4g. The pre-registration listed eight subversion items but we ended up dropping one item before running the study because we had reason to believe it did not represent a widely held democratic norm. In some cases we slightly changed the wording of questions for greater clarity. Because of concerns about attention on Lucid, we added an additional attention check.

After filing this preregistration, we ran a pilot experiment on a representative sample of 668 US residents recruited via Lucid in March of 2021. The findings are almost identical to the ones reported in this study. We present the key experimental results in the last two figures of the Supplemental Materials. Figure S7 shows the manipulation check and Figure S8 shows the treatment effect on willingness to subvert democracy. Both show large and statistically significant differences for Democrats and Republicans.

Question Wording

Self-Reported Subversion. In Studies 1 and 2, we construct a scale from seven items as the main dependent variable measure. Question order is randomized for each participant. Response options are "Never," "Probably Not," "Probably," and "Definitely." For Democrats, we ask: (1) "Would YOU support banning FAR-RIGHT rallies in the state capital?" (2) "Would YOU support ignoring controversial rulings by REPUBLICAN JUDGES?" (3) "Would YOU support freezing the social media accounts of REPUBLICAN JOURNALISTS?" (4) "Would YOU support reducing the number of voting stations in towns that support REPUBLICANS?" (5) "Would YOU support laws that would make it easier for DEMOCRATS (and harder for REPUBLICANS) to get elected?" (6) "Would YOU support using violence to block major REPUBLICAN laws?" (7) "Would YOU support significantly

reinterpreting the Constitution in order to block REPUBLICAN policies?" For Republicans, we ask: (1) "Would YOU support banning FAR-LEFT rallies in the state capital?" (2) "Would YOU support ignoring controversial court rulings by DEMOCRAT JUDGES?" (3) "Would YOU support freezing the social media accounts of DEMOCRAT JOURNALISTS?" (4) "Would YOU support reducing the number of voting stations in towns that support DEMOCRATS?" (5) "Would YOU support laws that would make it easier for REPUBLICANS (and harder for DEMOCRATS) to get elected?" (6) "Would YOU support using violence to block major DEMOCRAT laws?" (7) "Would YOU support significantly reinterpreting the Constitution in order to block DEMOCRAT policies?"

Opposing Partisan Subversion Questions. In Studies 1 and 2, we measure beliefs about opposing partisan willingness to subvert democratic norms. In the treatment group of Study 2, respondents receive feedback after answering each question. Questions are randomized for each participant. Response options are "Never," "Probably Not," "Probably," and "Definitely." For Democrats, we ask: (1) "Would MOST REPUBLICANS support banning FAR-LEFT rallies in the state capital?" (2) "Would MOST REPUBLICANS support ignoring controversial court rulings by DEMOCRAT JUDGES?" (3) "Would MOST REPUBLICANS support freezing the social media accounts of DEMOCRAT JOURNALISTS?" (4) "Would MOST REPUBLICANS support reducing the number of voting stations in towns that support DEMOCRATS?" (5) "Would MOST REPUBLICANS support laws that would make it easier for REPUBLICANS (and harder for DEMOCRATS) to get elected?" (6) "Would MOST REPUBLICANS support using violence to block major DEMOCRAT laws?" (7) "Would MOST REPUBLICANS support significantly reinterpreting the Constitution in order to block DEMOCRAT policies?" For Republicans, we ask: (1) "Would MOST DEMOCRATS support banning FAR-RIGHT rallies in the state capital?" (2) "Would MOST DEMOCRATS support ignoring controversial court rulings by REPUBLICAN JUDGES?" (3) "Would MOST DEMOCRATS support freezing the social media accounts of REPUBLICAN JOURNALISTS?" (4) "Would MOST DEMOCRATS support reducing the number of voting stations in towns that support REPUBLICANS?" (5) "Would MOST DEMOCRATS support laws that would make it easier for DEMOCRATS (and harder for REPUBLICANS) to get elected?" (6) "Would MOST DEMOCRATS support using violence to block major REPUBLICAN laws?" (7) "Would MOST DEMOCRATS support significantly reinterpreting the Constitution in order to block REPUBLICAN policies?" In addition, we

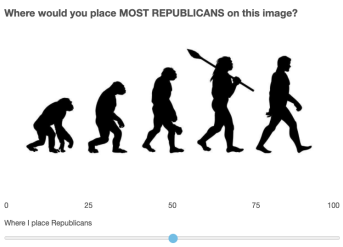
asked a comprehension question and ended the survey when respondents answered incorrectly in both studies. The question gave respondents the following options to describe what the questions in this battery had been about: (1) "What I would support in American politics." (2) "What most Republicans would support in American politics." (3) "What most Democrats would support in American politics."

Self-Reported Policy Questions. To measure policy extremism, Studies 1 and 2 ask a five-item questionnaire from the ANES. Response options are "Strongly Disagree," "Disagree," "Somewhat Disagree," "Neither Agree nor Disagree," "Somewhat Agree," "Agree," or "Strongly Agree." Question order is randomized for each participant. We ask Democrats to respond to the following: (1) "It is important for the government to provide many more services, even if it means an increase in spending." (2) "The government should spend much less money on defense." (3) "The government should make every effort to improve the social and economic position of blacks." (4) "By law, a woman should always be able to obtain an abortion as a matter of personal choice." (5) "There should be a government insurance plan which would cover all medical and hospital expenses for everyone." We ask Republicans to respond to the following: (1) "The government should provide fewer services, even in areas such as health and education, in order to reduce spending." (2) "Government defense spending should be greatly increased." (3) "The government should not make any special effort to help blacks because they should help themselves." (4) "By law, abortion should never be permitted." (5) "Instead of a government insurance plan, medical expenses should be paid by individuals, and through private insurance plans like Blue Cross."

Opposing Partisan Policy Questions. Studies 1 and 2 repeat the ANES battery, with order randomized for each respondent, and we ask Democrats to respond to the following: (1) "MOST REPUBLICANS BELIEVE that the government should provide fewer services, even in areas such as health and education, in order to reduce spending." (2) "MOST REPUBLICANS BELIEVE that government defense spending should be greatly increased." (3) "MOST REPUBLICANS BELIEVE that the government should not make any special effort to help blacks because they should help themselves." (4) "MOST REPUBLICANS BELIEVE that, by law, abortion should never be permitted." (5) "MOST REPUBLICANS BELIEVE that, instead of a government insurance plan, medical expenses should be paid by individuals, and through private insurance plans like Blue Cross. We ask

Republicans to respond to the following: (1)"MOST DEMOCRATS BELIEVE that it is important for the government to provide many more services, even if it means an increase in spending." (2) "MOST DEMOCRATS BELIEVE that the government should spend much less money on defense." (3) "MOST DEMOCRATS BELIEVE that the government should make every effort to improve the social and economic position of blacks." (4) "MOST DEMOCRATS BELIEVE that, by law, a woman should always be able to obtain an abortion as a matter of personal choice." (5) "MOST DEMOCRATS BELIEVE that there should be a government insurance plan which would cover all medical and hospital expenses for everyone."

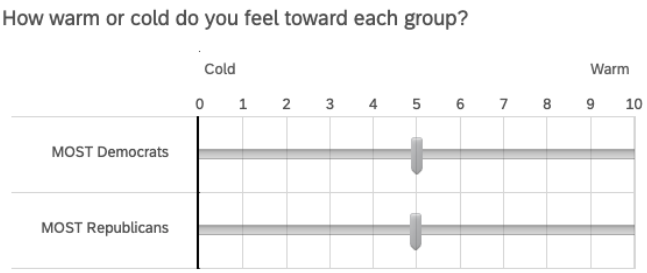
Measuring Partisan Dehumanization of Opposing Partisans. One way that we measure affective polarization is in the form of blatant dehumanization and metadehumanization (the belief that the opposing party sees oneself as less than human), which has been found to be a strong indicator of negative affect between parties in the US (*24*, *28*). This is measured using the "Ascent of Man" scale, developed as a tool to measure blatant dehumanization by Kteily et.al. (2016). Participants see these questions as in the following example for Democrats:



Political Knowledge Questions. To gauge political knowledge, we ask participants in Studies 1 and 2 three political knowledge questions in random order. First, "On which of the following does the U.S. federal government currently spend the least?" Response options are "Foreign aid," "Medicare," "National defense," and "Social Security." Second, "Do you happen to remember who was Hillary Clinton's running mate for Vice President in 2016?" Response options are "Joe Biden," "Elizabeth Warren," "Tim Kaine," "Newt Gingrich," and "John Kerry." Third, "For how many years is a member of the United States Senate elected - that is, how many years are there in one full term of office for a US Senator?" Response options are "Two years," "Four years," "Six years," and "Eight years."

Partisan Animosity Questions. To gauge levels of partisan animosity, all participants in Study 1 and

Study 2 are asked to use a standard "feeling thermometer" ranging from 0 "cold" to 100 "warm" to rate how they feel toward MOST Democrats and MOST Republicans respectively. Participants see the question as follows:



Primary Election Questions. In Study 2, we add an exploratory measure of partisan willingness to vote for candidates in a primary election that either support or oppose breaking democratic norms. Respondents face a choice between two candidates from their own party, one of whom has "supported" breaking one of the democratic norms featured in our study and another who has "opposed" breaking that same norm. Each norm is randomly selected without replacement.

In order to model the real-world rhetoric that we contend produces democratic backsliding, participants are told that the candidate who "supported" breaking a democratic norm did so because they believe opposing partisans (Democrats or Republicans) have done the same.

A Republican would see the question in the following form:

> Now, please tell us how you would vote in two Republican primary elections.
>
> **Republican Primary Election**
>
> **Candidate A**: Has SUPPORTED reinterpreting the Constitution to block DEMOCRAT policies because believes DEMOCRATs have done the same.
>
> **Candidate B**: Has OPPOSED reinterpreting the Constitution to block DEMOCRAT policies.
>
> Which candidate would you vote for?

Options: “Candidate A,” “Candidate B,” and “Neither.”

The set of possible democracy-supporting candidate statements include: (1) “Has OPPOSED freezing the social media accounts of [OTHER PARTY] journalists.” (2) “Has OPPOSED banning far-[OTHER WING] rallies in the state capital.” (3) “Has OPPOSED reinterpreting the Constitution to block [OTHER PARTY] policies.” (4) “Has OPPOSED ignoring controversial court rulings by [OTHER PARTY] judges.” (5) “Has OPPOSED reducing the number of polling stations in areas that support [OTHER PARTY]s.” (6) “Has OPPOSED laws that would make it easier for [OWN PARTY]s (and harder for [OTHER PARTY]s) to get elected.” (7) “Has OPPOSED using violence to block major [OTHER PARTY] laws.”

The set of possible democracy-subverting candidate statements include: (1) “Has SUPPORTED freezing the social media accounts of [OTHER PARTY] journalists because believes [OTHER PARTY] has done the same.” (2) “Has SUPPORTED banning far-[OTHER WING] rallies in the state capital because believes [OTHER PARTY] has done the same.” (3) “Has SUPPORTED reinterpreting the Constitution to block [OTHER PARTY] policies because believes [OTHER PARTY] has done the same.” (4) “Has SUPPORTED ignoring controversial court rulings by [OTHER PARTY] judges because believes [OTHER PARTY] has done the same.” (5) “Has SUPPORTED reducing the number of polling stations in areas that support [OTHER PARTY]s because believes [OTHER PARTY] has done the same.” (6) “Has SUPPORTED laws that would make it easier for [OWN PARTY]s (and harder for [OTHER PARTY]s) to get elected because believes [OTHER PARTY] has done the same.” (7) “Has SUPPORTED using violence to block major [OTHER PARTY] laws because believes [OTHER PARTY] has done the same.”

Tables and Figures

Table S1: Demographic Information for Study 1 and 2 Participants

	Study 1	**Studies 2a and 2b**	**Benchmark**
Sample	Lucid	Lucid	ACS 2019
N	1,973	2,543	12,686,854
Gender			
Female	54	52	52
Age			
18-24	11	13	9
25-34	13	17	15
35-49	26	24	22
50-64	32	28	27
65+	18	17	26
Race			
Non-Hispanic White	72	68	70
Black	8	10	10
Asian + Other	7	8	8
Hispanic	13	14	13
Education			
No HS Degree	3	5	11
HS Graduate	20	24	28
Some College / 2-year Degree	33	32	31
Bachelor's Degree	28	22	19
Graduate Degree	15	13	12
Income			
<$20k	25	27	12
$20k-$39k	22	25	15
$40k-$59k	16	16	14
$60k-$79k	11	11	12
$80k+	26	21	47

Note: Except for the N row, cell entries provide the percentage of each sample present in each demographic category. In rare cases age and income categories were inconsistent across surveys and were either combined or averaged across other categories. We excluded respondents from the ACS 2019 under 18.

Table S2: Descriptive Statistics for Study 1

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Willingness to Subvert Democracy (7-item scale)	1,973	0.263	0.197	0.000	0.095	0.381	1.000
Perceived Willingness of the Other Party to Subvert (7-item scale)	1,973	0.662	0.223	0.000	0.524	0.810	1.000
Party Identification (7-point)	1,973	0.474	0.402	0	0	0.8	1
Party Identification Strength (7-point,folded)	1,973	0.640	0.401	0	0.5	1	1
Policy Extremism (5-item scale)	1,973	0.644	0.214	0.000	0.500	0.800	1.000
Policy Extremism of Opposing Partisans (5-item scale)	1,944	0.740	0.192	0.000	0.633	0.867	1.000
Age	1,973	0.389	0.219	0.000	0.218	0.564	1.000
Education	1,973	0.424	0.324	0.000	0.000	0.500	1.000
Dehumanization of Other Party	1,940	0.398	0.333	0.000	0.100	0.710	1.000
Perception of the Other Party's Dehumanization of Your Party	1,940	0.537	0.345	0.000	0.240	0.850	1.000
General Political Knowledge (3-item scale)	1,936	0.433	0.329	0.000	0.000	0.667	1.000

Table S3: Descriptive Statistics for Study for Studies 2a and 2b

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Treatmemt	2,543	0.497	0.500	0	0	1	1
Willingness to Subvert Democracy (7-item scale)	2,543	0.205	0.200	0.000	0.048	0.333	1.000
Perceived Willingness of the Other Party to Subvert (7-item scale)	2,543	0.519	0.244	0.000	0.333	0.714	1.000
Party Identification (7-point)	2,543	0.454	0.397	0.000	0.000	0.833	1.000
Party Identification Strength (7-point,folded)	2,543	0.631	0.398	0	0.5	1	1
Policy Extremism (5-item scale)	2,514	0.627	0.209	0.000	0.467	0.800	1.000
Perceived Policy Extremism of Opposing Partisans (5-item scale)	2,510	0.677	0.214	0.000	0.533	0.833	1.000
Age	2,543	0.386	0.239	0.000	0.176	0.581	1.000
Education	2,543	0.393	0.326	0	0	0.5	1
Dehumanization of Other Party	2,506	0.807	0.171	0.000	0.653	0.950	1.000
Perception of the Other Party's Dehumanization of Your Party	2,506	0.747	0.183	0.000	0.598	0.910	1.000
Vote for Subverting Candidate in Primary 1	2,527	0.239	0.369	0.000	0.000	0.500	1.000
Vote for Subverting Candidate in Primary 2	2,521	0.245	0.375	0.000	0.000	0.500	1.000

Table S4: Explaining Willingness to Subvert Democracy Reduced Form—Study 2. All variables are coded 0-1 and rescaled so that higher values are associated with greater willingness to subvert democracy. We only include pretreatment covariates in this regression, which excludes most of the questions we asked in the survey. We impute a handful of missing values on Partisan Identity Strength with demographics as described earlier.

	Dependent variable:	
	Willingness to Subvert Democracy (7-item index)	
	(1)	(2)
Treatment	-0.076^{***}	-0.077^{***}
	(0.008)	(0.008)
Partisan Identity Strength		0.072^{***}
		(0.010)
Age		-0.160^{***}
		(0.016)
Education		0.005
		(0.012)
Constant	0.243^{***}	0.258^{***}
	(0.005)	(0.010)
Observations	2,543	2,543
R^2	0.036	0.088
Adjusted R^2	0.036	0.086
Residual Std. Error	0.197	0.191
Note:		Standard Errors in Parentheses
		* p<0.05; ** p<0.01; *** p<0.001

Table S5: Explaining Willingness to Subvert Democracy 2SLS—Study 2. This table uses two-stage least-squares to estimate the complier causal average effect (CACE). We instrument the last item (7/7) of the Perceived Willingness of Opposing Partisans to Subvert Democracy scale with an indicator variable for treatment group. We use the last item each respondent saw (item 7/7) because respondents had received most of the treatment by that point. All variables are coded 0-1 and rescaled so that higher values are associated with greater willingness to subvert democracy. We only include pretreatment covariates in this regression, which excludes most of the questions we asked in the survey. We impute a handful of missing values on Partisan Identification Strength with demographics as described earlier.

	Dependent variable:	
	Willingness to Subvert Democracy (7-item scale)	
	(1)	(2)
Perceived Willing. of Oppo. Partisans to Subvert Dem. (item 7/7)	0.247***	0.249***
	(0.025)	(0.024)
Partisan Identity Strength		0.054***
		(0.010)
Age		−0.177***
		(0.016)
Education		0.008
		(0.012)
Constant	0.086***	0.117***
	(0.013)	(0.014)
Observations	2,543	2,543
R^2	0.051	0.102
Adjusted R^2	0.050	0.100
Residual Std. Error	0.195	0.190

Note: Standard Errors in Parentheses
* p<0.05; ** p<0.01; *** p<0.001

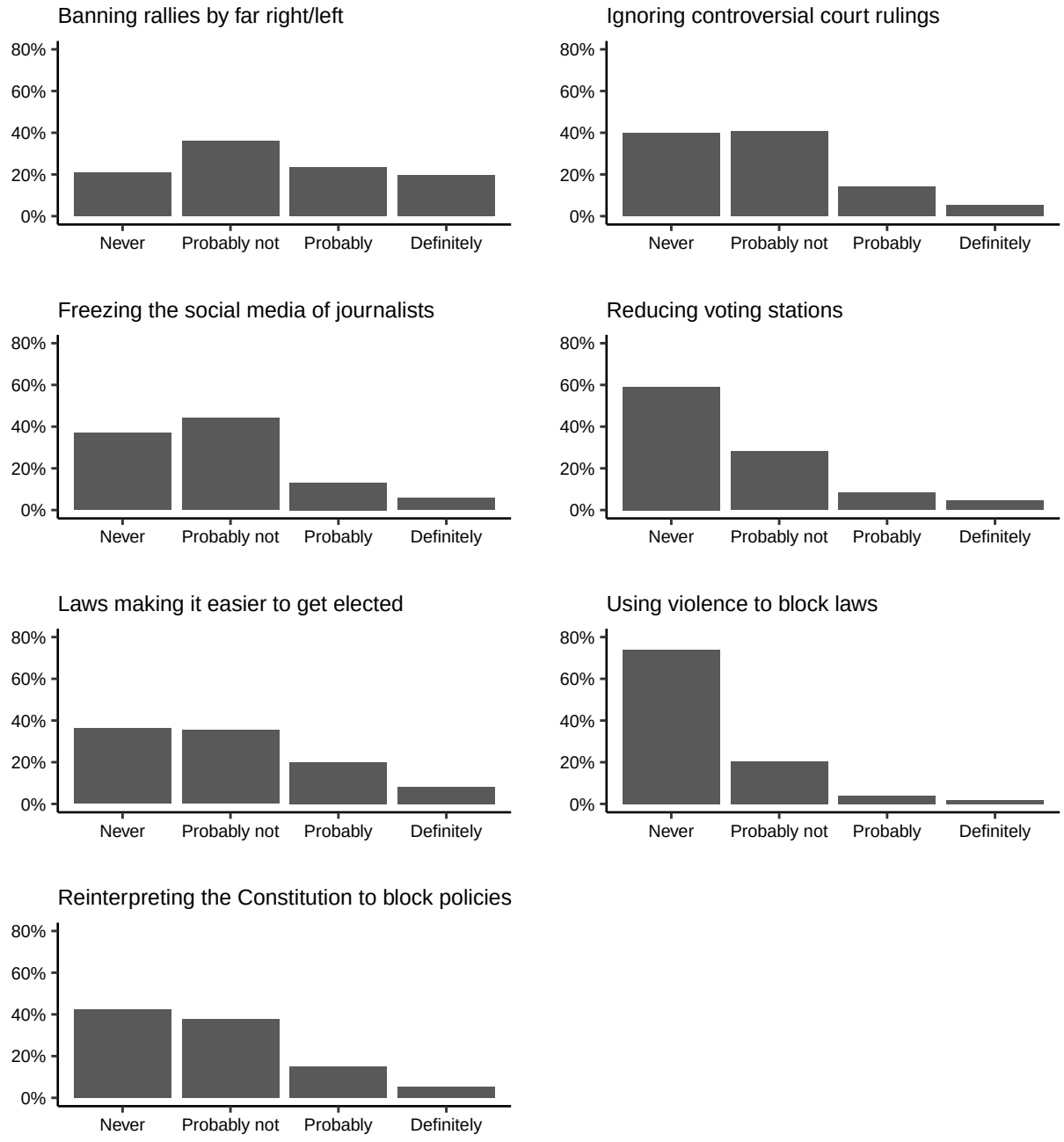


Figure S1: Response distributions for DEMOCRATS' willingness to subvert democracy in the seven scenarios in Study 1. Each question is phrased in terms of advantaging your party/disadvantaging the other party.

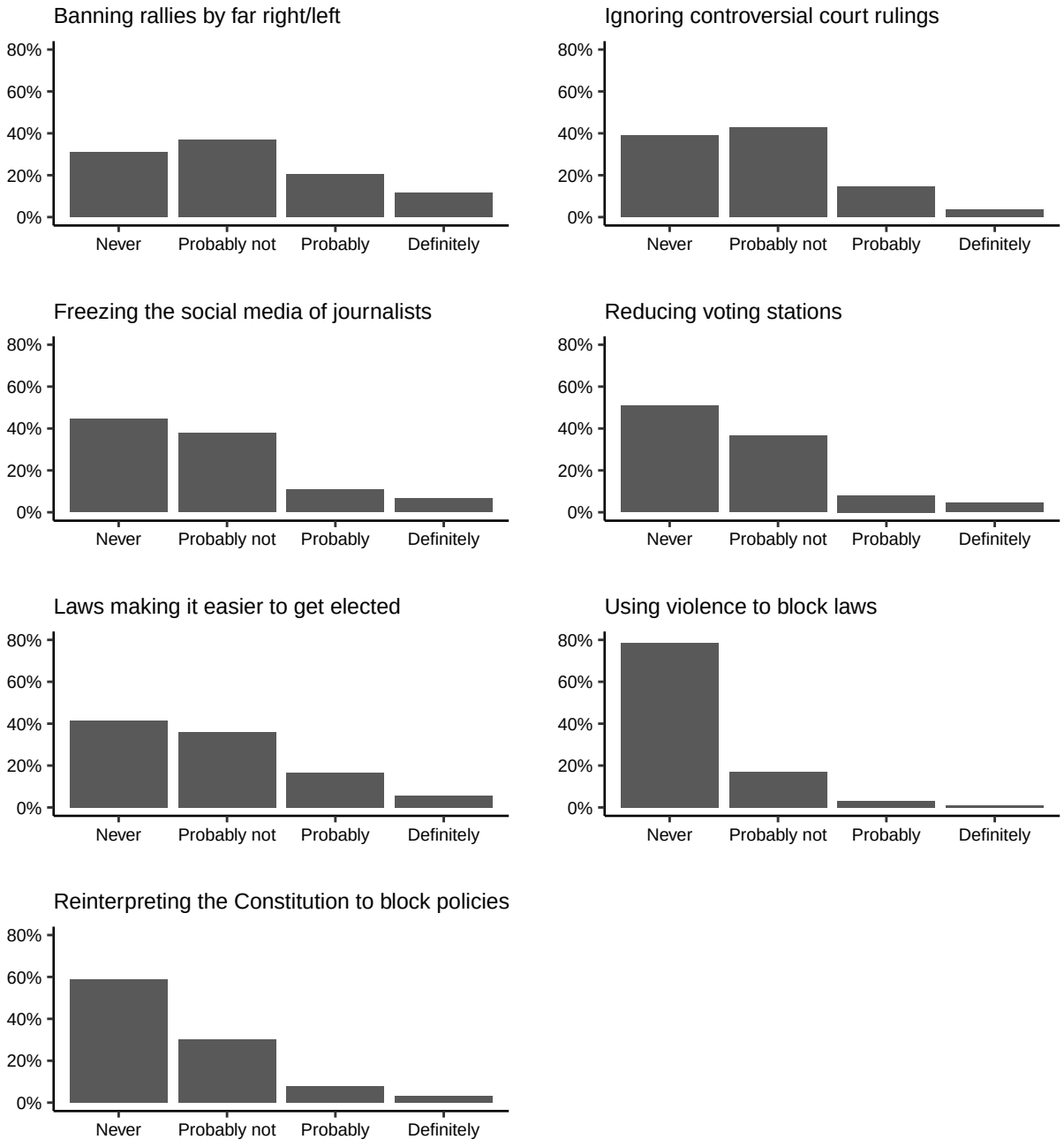


Figure S2: Response distributions for REPUBLICANS' willingness to subvert democracy in the seven scenarios in Study 1. Each question is phrased in terms of advantaging your party/disadvantaging the other party.

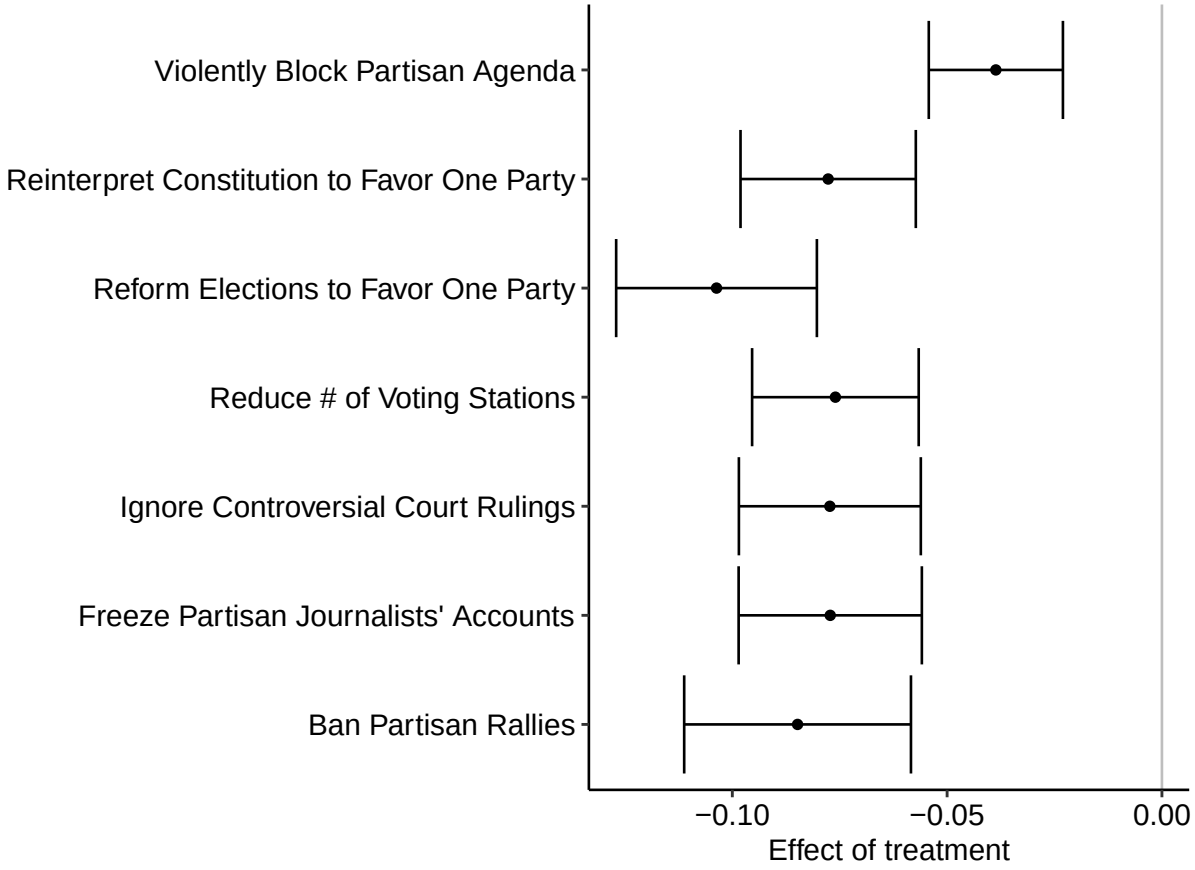


Figure S3: Study 2 treatment effect for each of the seven items used in the willingness to subvert scale. Higher values indicate more support for subverting democracy on each of the seven items.

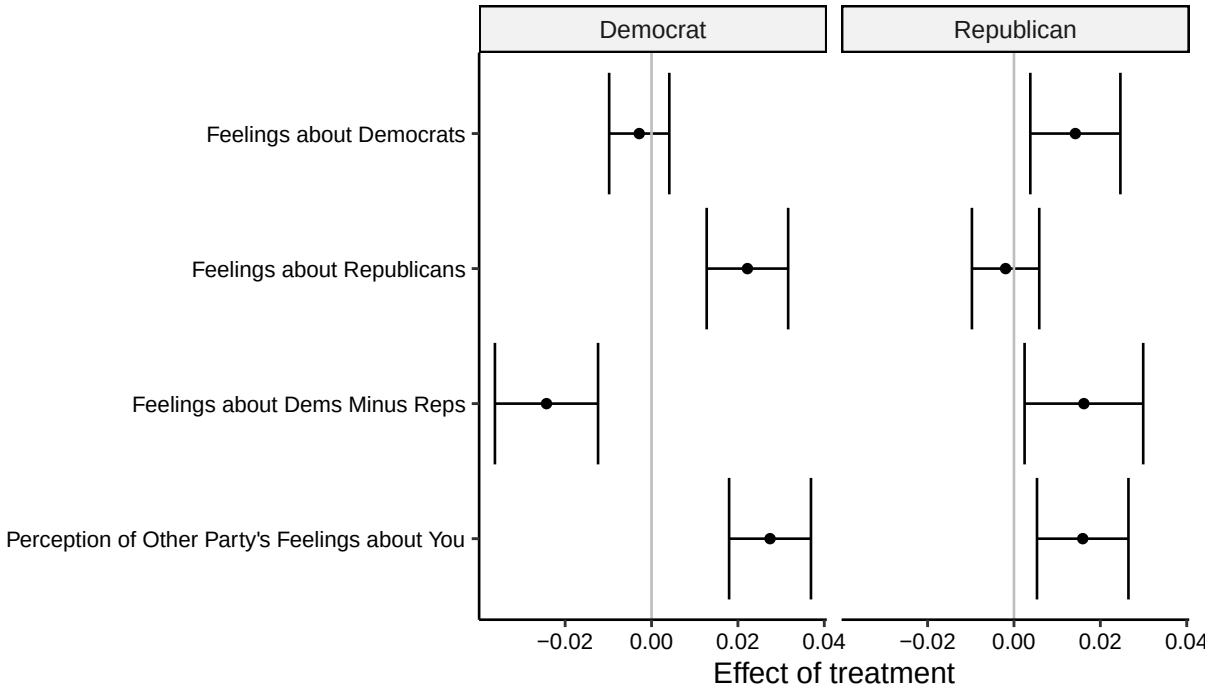


Figure S4: Study 2 treatment effect on the feeling thermometers and perception of the other party's feeling about you.

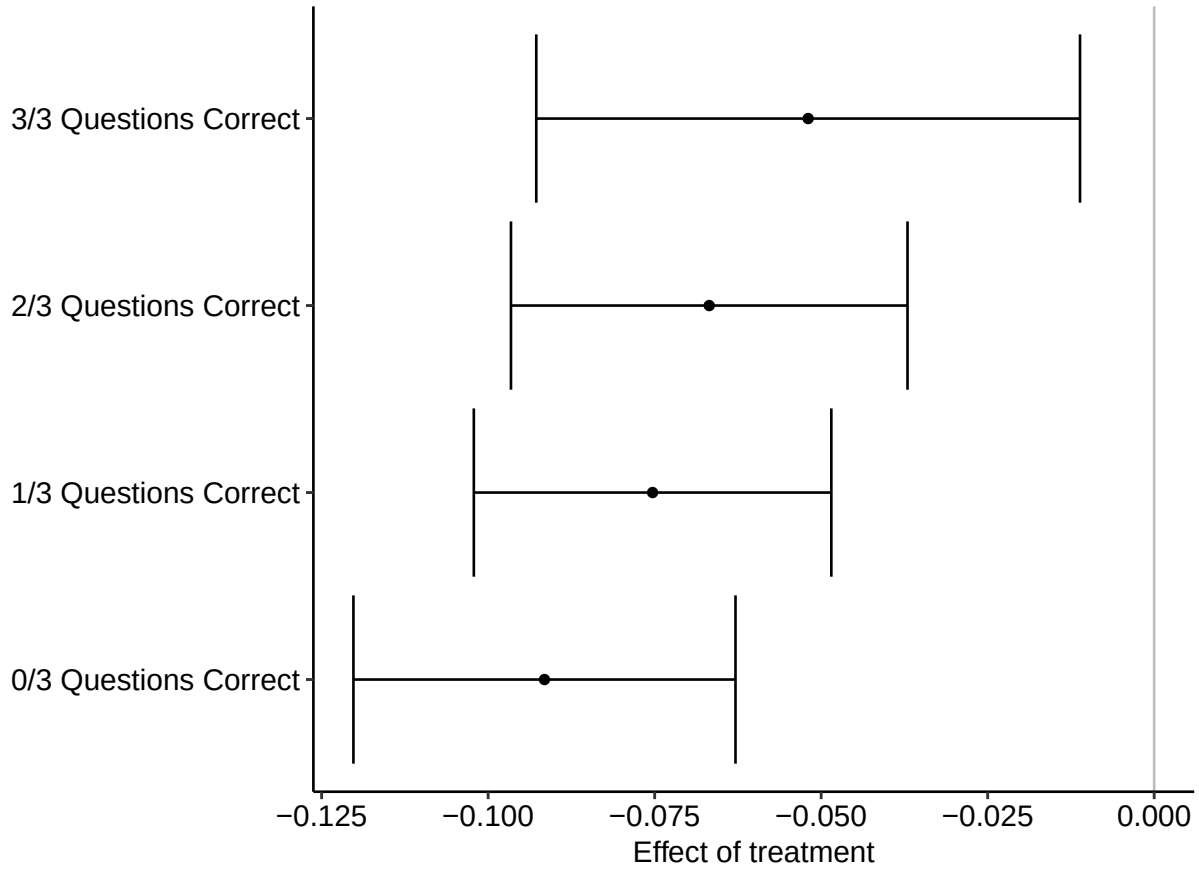


Figure S5: Study 2 treatment effect on the seven-item self subversion scale by number of items correct on the 3-item general political knowledge scale. The figure shows that lower knowledge individuals may exhibit a larger treatment effect. Note that the political knowledge scale was asked posttreatment.

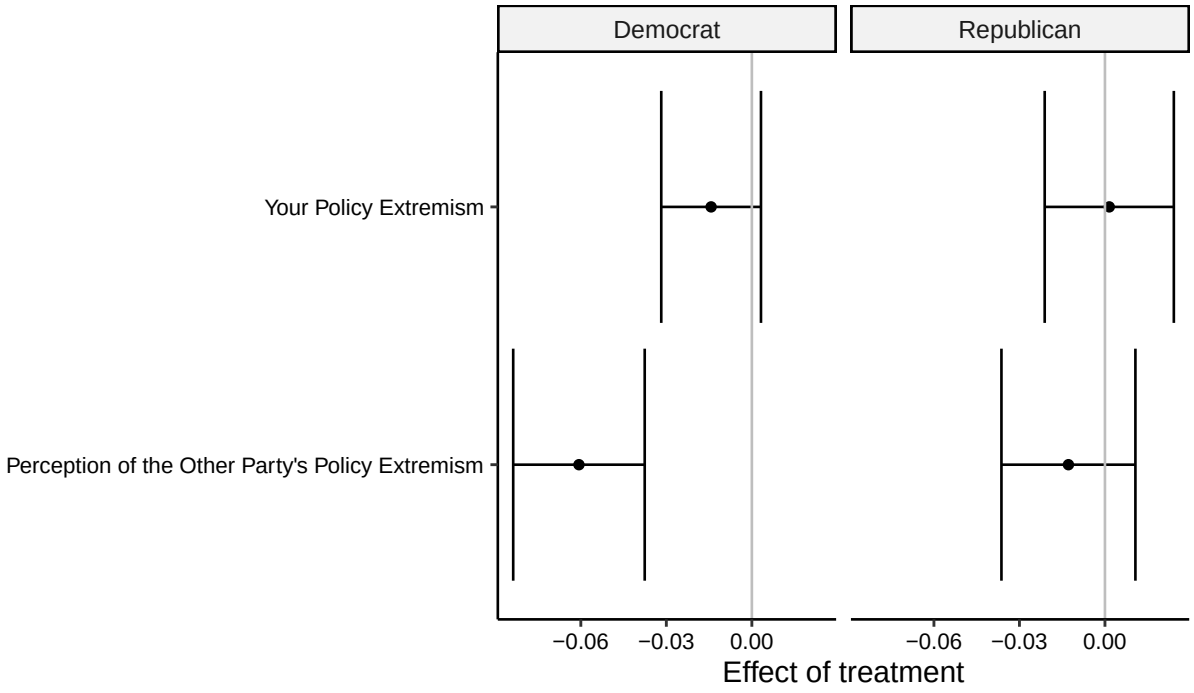


Figure S6: Study 2 treatment effect on respondent's policy extremism and their perception of the opposing party's policy extremism Both measures take the mean of five ANES items. Higher values indicate more extreme policy views or perception of more extremism. The figure shows that Democrats tend to see Republicans as less extreme after they have been exposed to the treatment. Republicans also perceive Democrats as less extreme, but the estimate is smaller and imprecise. There is a hint that the treatment makes Democrats less extreme themselves, but the estimate is imprecise

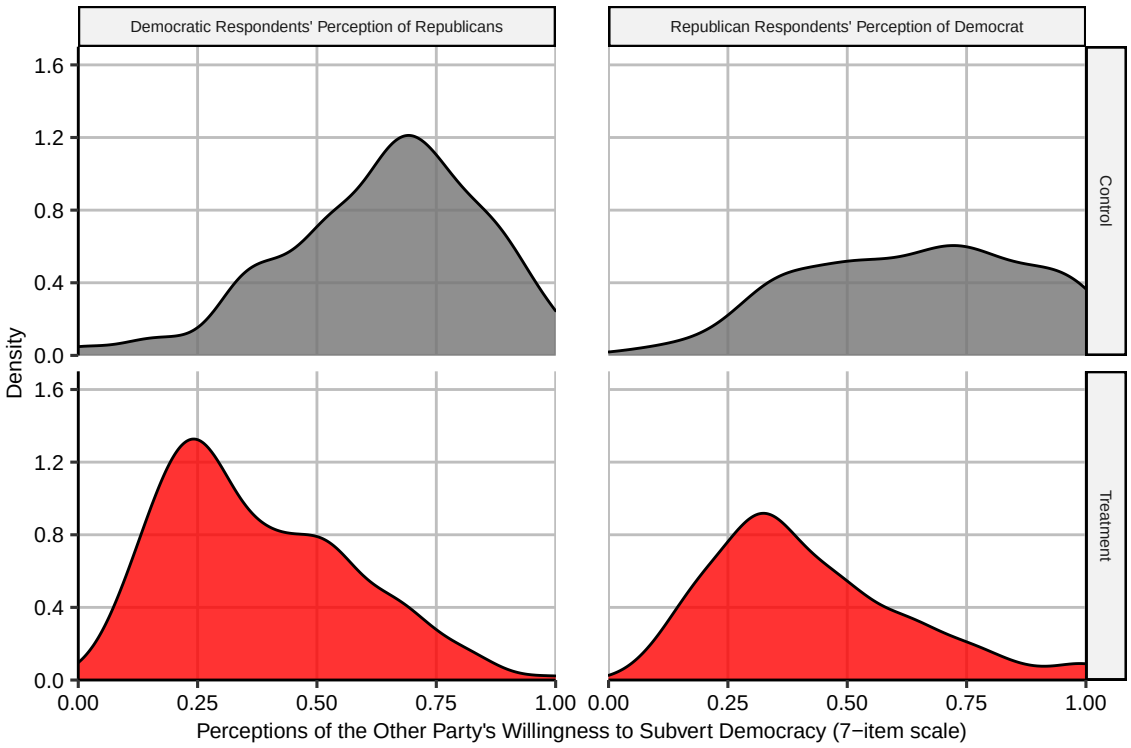


Figure S7: Pilot Study: Correcting Misperceptions - Manipulation Check. This figure shows that an informational intervention effectively lowers partisan overestimates of opposing partisan willingness to break democratic norms. Since the information was given in between each of the seven questions, this figure likely understates the actual treatment effect, since respondents don't received the full treatment until they reach the seventh item.

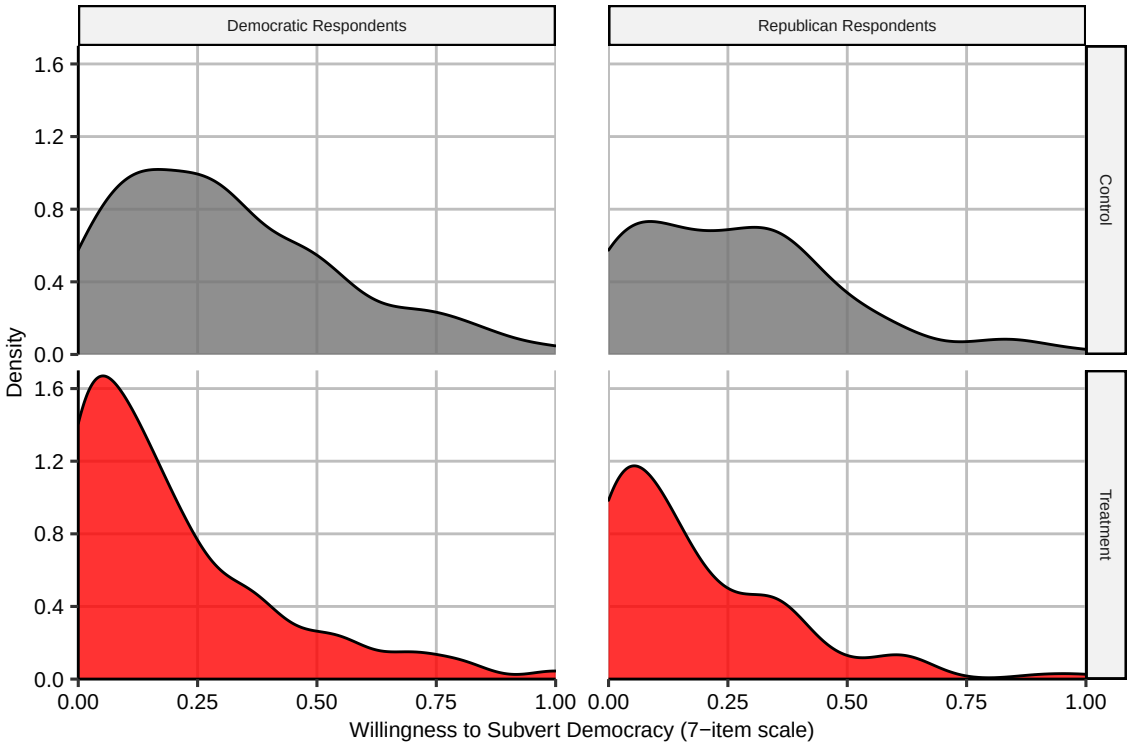


Figure S8: Pilot Study: Support for Democratic Norms - Experimental Effect. The informational, ask-tell intervention, which lowers expectations that opposing partisans will break democratic norms, substantially decreases participant's own willingness to subvert democracy. The effect is highly statistically significant for all respondents (df=665,p=3.1e-09) and for each party.