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

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

August 18, 2022

Abstract

Around the world, citizens are voting away the democracies they claim to cherish. Why are they voting against their own values? In this article, we provide evidence that this behavior is driven in part by fear that their opponents might dismantle democracy first. In an observational study (N=1,973), we find that US partisans who most fear the other party’s willingness to subvert democracy are also those most willing to support subverting democracy themselves. In experimental studies (N=2,543, N=1,848), we use an intervention to reduce these often exaggerated fears. With these fears reduced, partisans become more committed to upholding democratic norms. They may also become more willing to vote against candidates of their own party who 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.

[‡]Media Lab, Massachusetts Institute of Technology; Cambridge, MA 02139-4301.

[§]Toronto Metropolitan University; Toronto, ON M5B 2K3.

[¶]MIT Connection Science, 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 cherish. While 78% of the world's population says they want to live in a representative democracy, the past thirty years of democratic advances have been eradicated, leaving 70% of the world's population living in autocracies (1, 2). Citizens in Venezuela, Turkey, and Hungary strongly endorsed democracy while casting votes for authoritarian leaders Chávez, Erdoğan, and Orbán, respectively (3, 4). In fact, in Venezuela, citizen's who claimed to support democracy the most were no more likely to actually vote for a democratic candidate (4). Why are citizens around the world voting against their self-reported preferences?

The puzzle deepens when one considers that the modal form of autocratization today is democratic backsliding (3, 5). In contrast with previous eras, when democracy might be quickly destroyed in a coup, today democracies tend to die a slow death, leaving years for a democracy-loving public to respond and hold their representatives accountable (5, 6). Why, then, is democracy slipping away from the majority of the world's citizens across all regions, cultures, and socioeconomic conditions (2, 7)?

In this paper, we provide experimental evidence in support of one explanation for this puzzle. In order for democracy to sustain itself, partisans must be willing to constrain their politicians from violating key democratic norms. As Weingast writes: "To survive, a constitution must have more than philosophical or logical appeal; citizens must be willing to defend it" (8). But if citizens do not believe that opposing partisans will hold their representatives in check, they have a powerful incentive to give their own politicians broad leeway to do whatever it takes to save democracy from their opponents.

This provides would-be authoritarians with a powerful weapon against democracy, a way to dismantle democracy even with a base of democracy-loving supporters. They stoke fears among their supporters about the other side's commitment to democracy. As their supporters become fearful of the other side defecting, they become more willing to tolerate their leaders' antidemocratic actions. Ironically, supporters of would-be authoritarians grow increasingly fearful that the opposing side will violate democratic norms just as their leaders are doing the same. These democratic oversteps in turn generate real fear in the opposing party leading to an escalation of mutual fears. Under these conditions, even democracy-loving opponents of an aspiring autocrat are placed in the difficult situation: how to save democracy from an aspiring autocrat without escalating this cycle of mutual fear and democratic subversion?

We call this predicament the "subversion dilemma." It has parallels in the "security dilemma" in the realist view of international relations (9–11). Realists attribute the onset of interstate war to the global condition of "anarchy" in which states are not subject to an external sovereign. 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 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 the costs of war.

Because constitutions cannot defend themselves and because citizens serve as the ultimate sovereign in a democratic polity, citizens in a democracy find themselves in a self-help environment like states in the international system. Escalating fear between partisan groups can lead to increasing perceptions of threat. Just as countries may go to war even though they prefer peace due to the slippery slope of the security dilemma, citizens may vote away their cherished democracies due to the slippery slope of the subversion dilemma.

Would-be authoritarians' ability to weaponize the subversion dilemma, we suggest, is part of and may depend upon a larger set of mutually-reinforcing polarizations that contribute to democratic backsliding (12–14). These include heightened partisan identity strength (15–17), divergent policy goals (2, 15, 18), heightened dislike between partisans (19, 20), increased dehumanization between partisans (21, 22), stereotypes of opposing partisans (23, 24), and ethnic antagonism (25). Would-be authoritarians likely find it easier to stoke fears between partisans in societies that are already highly polarized along other dimensions, and the US has become such a society over the last several decades (20, 26).

Although various forms of polarization likely contribute to democratic backsliding, we think the subversion dilemma may play a particularly pernicious and underappreciated role. It is pernicious because it exploits a weakness in the structure of democracy itself: democratic survival depends on a mutual commitment to democratic norms (8, 27, 28). As in the prisoner's dilemma game, if one party suspects the other is defecting, then the best response is to defect.

Critically, both sides need not defect for democracy to decline. It is sufficient if just one partisan group becomes so fearful of the other side that they are no longer willing to act as a check on their representatives. An opposing party with a strong commitment to democracy still faces a difficult dilemma: defect in kind or potentially allow an aspiring autocrat to take over. After all, in the security dilemma, if one country attacks another due to exaggerated fears, the attacked country loses peace even if it refuses to fight.

In the US context, Donald Trump stoked fears about Democrats subverting democracy from the start. Early in his 2016 campaign, his website stated: "Help Me Stop Crooked Hillary from Rigging this Election!" (3). Throughout the 2016 campaign, he repeated, "This is a rigged election" (29). These accusations continued through the 2020 election, and Fox News amplified this message, repeatedly proclaiming the existence of "an all-out effort to depress and suppress the pro Trump vote" (30). For instance, they accused Democrats of discarding Republican ballots in a ditch in Wisconsin (31). At the very same time, they cast Trump as the candidate "trying to protect democracy" (32).

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, CNN portrayed Trump as "manipulating the ballots" and planning to "disregard the popular vote" (33, 34). 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" (35). Despite heightened fears and rhetoric, we have not observed obvious increases in democratic subversion from Democrats in the US. This forbearance is likely upholding democracy in the US for now, yet it is unclear how long forbearance can protect democracy among increasing mutual fears and Republicans'

incremental democratic violations.

To summarize, we think democratic-loving voters may vote away the democracies they cherish in part because (1) would-be authoritarians convince their supporters that the other party is subverting democracy, (2) leading them to begin tolerating backsliding by their leaders, which then (3) prompts legitimate fears in the other party, (4) leading the other party to entertain or tolerate more drastic action by their own leaders, which then, (5) confirms the unfounded fears generated in (1) and ultimately may start a death spiral for democracy.

In this paper, we focus on testing one part of this chain of causal claims, one that is at the heart of the subversion dilemma. Across three studies, we examine whether ordinary partisans allow their representatives to subvert democracy to the degree they fear the other party is willing to subvert it, the causal claim in (2) and (4). Study 1 uses survey data to examine this claim observationally. Study 2 then tests it experimentally, showing that reducing partisans' fear that the other side will violate democratic norms increases partisans' own support for these norms and decreases their willingness to vote for candidates who violate them. That is, when citizens are assured that opposing partisans will keep their representatives in check, they themselves may be more willing to act as a check on democratic backsliding. Finally, Study 3 addresses potential concerns about demand and experimentally-induced social desirability bias by replicating Study 2 in a "seemingly-unrelated surveys" approach.

Study 1: Fear and Support for Backsliding

To examine whether 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 express their level of support for various actions that would benefit their own party at the expense of a democratic norm. We conceptualize democratic norms as common understandings about the necessary areas where political actors must exercise restraint in order to uphold democratic institutions, regardless of issues of legality. We select norms shown to be important to the American public across political affiliations and that are also considered critical for democratic survival by political scientists (15, 36, 37). We avoid using the words "democratic" or "norm" to limit concerns about socially desirable responding. As an example, we ask Democrats: "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."

To assess respondent's beliefs about the other party's willingness to subvert, we ask about the same seven scenarios. As an example, we asked Democrats: "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, ultimately, the defense of democracy depends on citizens' willingness and ability to

constrain potential office holders who themselves are likely to have varying degrees of democratic commitment (8). For conciseness, we describe these beliefs about the opposing party as “fears” and show in Supporting Materials Figure S23 that these beliefs are indeed closely tied to self-reported fear.

We generate separate scales of the extent to which respondents support violating democratic norms and the extent to which they believe opposing partisans are willing to violate democratic norms by taking the unweighted average of the seven questions for each and rescaling both to 0-1 (Cronbach alphas of 0.89 and 0.82, respectively).

Figure 1 presents the distributions of responses to these seven-item indices by Republicans and Democrats, respectively. Strikingly, Republicans on average believe that Democrats are willing to subvert democratic norms, that is, say “probably” or “definitely,” in 5.0 of the seven scenarios (mean = 0.65 on the 0-1 scale), but Democrats self-report willingness to subvert these norms in only 1.5 of the scenarios (mean = 0.28 on the 0-1 scale). Similarly, Democrats believe that Republicans are willing to subvert democratic norms in 5.2 of the scenarios (mean = 0.67 on the 0-1 scale), but Republicans self-report willingness to subvert these norms in only 1.2 of the scenarios (mean = 0.28 on the 0-1 scale). Supplemental Materials Figures S18 and S19 show the distribution of self-reported willingness to subvert for each item by party.

Although this perception gap seems consistent with the logic of the subversion dilemma, other explanations are possible. For example, people may report these beliefs as a form of expressive responding—they look for opportunities to express negative sentiments about the other party on the survey. Nevertheless, it’s worth noting that the perception gap we find is larger than the perception gaps documented between partisans on issues of ideological and affective polarization (19, 23). 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 examine in this paper is that fear of the other side subverting democracy leads partisans to support subverting it themselves. Using the two multi-item 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 partisans’ 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 also run experiments in Studies 2 and 3). 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, extremism of racial attitudes, 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 (15, 16, 19, 21–23).

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 both Democrats and Republicans. Some of the covariate coefficients are wrongly signed as the model contains too many highly-correlated variables. The one variable that does slightly reduce the core relationship

is Republican perceptions of Democrats' views on race, in particular, on the government making a special effort to help Blacks (see Column 3). Controlling for this perception slightly reduces the coefficient for Republicans' belief that Democrats will subvert on Republicans' own willingness to subvert, though this seems likely to be due to chance as this control coefficient has the wrong sign. In Study 3, we include a four-item ethnic antagonism battery drawn from Bartels (25). We find that at least this set of variables does not account for the strong relationship between fear of the other side subverting and one's own willingness to subvert democratic norms.

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.

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 perception of opposing partisan's willingness to subvert democracy and their own willingness to subvert democracy (kernel densities, $N=1,973$). The questions range from reducing polling places in the other party's towns to banning the other party's rallies. The figure shows that members of both parties appear to significantly overestimate opposing partisans' willingness to break democratic norms.

Figure 2: You Subvert, I Subvert—Study 1. This figure shows each's respondents' willingness to subvert democracy by their perception that the other party supports subverting democracy. Each point shows one respondent (N=1,973). The points are jittered and point coloration reflects respondent's willingness to subvert. The best fit line is loess. This figure shows that partisans support breaking democratic norms more when they believe that opposing partisans are willing to break democratic norms. Both variables are coded to the 0-1 range. See the Figure 1 note for more details on the seven questions we ask respondents.

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. Columns 3 and 6 remove the five-item Policy Extremism and Perceived Policy Extremism of Opposing Partisans scales and replaces them with the single items from those scales that capture racial attitudes. Ordinary least-squares regression estimates showing unstandardized coefficients. Each column shows estimates from a separate regression model.

	<i>Dependent variable:</i>					
	Willingness to Subvert Democracy			Willingness to Subvert Democracy		
	Republicans			Democrats		
	(1)	(2)	(3)	(4)	(5)	(6)
Perceived Willingness of Opposing Partisans to Subvert Democracy	0.24*** (0.03)	0.23*** (0.03)	0.19*** (0.03)	0.29*** (0.03)	0.26*** (0.03)	0.25*** (0.03)
Partisan Identity Strength		0.04* (0.02)	0.04** (0.02)		0.01 (0.01)	0.01 (0.01)
Age		-0.06 (0.03)	-0.07* (0.03)		-0.20*** (0.03)	-0.20*** (0.03)
Education		0.02 (0.02)	0.01 (0.02)		-0.02 (0.02)	-0.02 (0.02)
Policy Extremism		0.05 (0.04)			-0.01 (0.04)	
Dehumanization of Opposing Partisans		0.14*** (0.02)	0.15*** (0.02)		0.08*** (0.02)	0.08*** (0.02)
Perceived Policy Extremism of Opposing Partisans		-0.20*** (0.04)			-0.07* (0.03)	
Perceived Dehumanization by Opposing Partisans		-0.09*** (0.02)	-0.10*** (0.02)		-0.05* (0.02)	-0.05* (0.02)
Feeling Thermometer Reps Minus Dems		0.02 (0.05)	0.01 (0.05)		-0.11* (0.05)	-0.11* (0.04)
General Political Knowledge		-0.05* (0.02)	-0.06** (0.02)		-0.002 (0.02)	-0.01 (0.02)
Gov. Special Effort to Help Blacks Extremism			0.01 (0.02)			-0.01 (0.03)
Perceived Gov. Spe. Eff. to Help Blacks Ext. by Opp. Partisans			-0.06** (0.02)			-0.03 (0.02)
Constant	0.08*** (0.02)	0.21*** (0.03)	0.16*** (0.03)	0.09*** (0.02)	0.26*** (0.04)	0.25*** (0.04)
Observations	913	913	913	1,031	1,031	1,031
R ²	0.09	0.19	0.17	0.09	0.17	0.17
Adjusted R ²	0.08	0.18	0.17	0.09	0.16	0.16
Residual Std. Error	0.19	0.18	0.18	0.19	0.18	0.18

Note:

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

Study 2a: Experiment: Reducing Fear Increases Support for Democratic Norms

Figure 3: Study 2 Flow for Treatment Condition. After answering each question about the opposing party’s support for subversion, respondents in the treatment group receive information from Study 1 about how opposing partisans actually answered the survey items. The control group sees the same flow except for the ‘Tell’ component. We randomize the order of the items within each battery.

We 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.

Successfully manipulating perceptions in survey environments is challenging given the speed with which respondents take surveys, so we turn to an “ask-tell” design, which engages respondents by providing them with feedback. Scholars have successfully employed similar designs to shape perceptions in other studies (23, 24, 38, 39). 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 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 S18 and S19).

After administering two attention checks and determining political party (we drop true Independents), we administer the ask-tell intervention to the 50% randomly assigned to the treatment condition and we ask the control group to answer the same questions about opposing partisans, though no feedback is provided. We then ask all respondents about their own willingness to subvert democratic norms using the same seven-item questionnaire from Study 1 (Figure 3 presents this survey flow visually). As in Study 1, we take the simple average of the seven items and rescale it 0-1. This experiment is a preregistered replication of a pilot experiment. We conducted this study September 15-29, 2021 on a demographically representative sample of 2,545 respondents recruited through Lucid (see Supplemental Materials Figures S33 and S34 for key pilot findings). We find no sign of differential attrition or of imbalances on pretreatment covariates (see Supplemental Materials Table S5).

The ask-tell treatment succeeds at lowering perceptions that opposing partisans are willing to

break democratic norms. Figure 4 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, $t(2,543)=-27.8$, $p=1.3e-148$, two-tailed. 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 receive feedback on the final question. Indeed, when we look at responses only for the final (randomized) question, we see a larger treatment effect with 0.63 for control and 0.33 for treatment.

Figure 4: Reducing Fear—Experimental Manipulation Check for Study 2. These plots show the distributions of the unweighted average of the seven questions we ask respondents about their perception of opposing partisan's willingness to subvert democracy (kernel densities, $N=2,543$). They reveal that, when we inform respondents about opposing partisans' actual level of support for subverting democracy in each of the seven scenarios, they see opposing partisans as less willing 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.

If voters vote away the democracies they cherish because they think the other side is already trying to dismantle it, this successful manipulation should increase support for upholding democratic norms. It does. The treatment group is less willing to subvert democracy than the control group. Figure 5 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, $t(2,543)=-9.8$, $p=3.4e-22$, two-tailed. Another way to understand this finding is that respondents in the control condition say

Figure 5: Less Fear, More Support for Democratic Norms—Experimental Effect for Study 2a. These plots show the distributions of the unweighted average of the seven questions we ask respondents about their own willingness to subvert democracy by treatment condition and party identification (kernel densities, $N=2,543$). They reveal that the informational intervention—which targets beliefs about opposing partisans’ commitment to democracy—decreases participants’ own willingness to subvert democracy. They therefore suggest that the observational relationship documented in Study 1 may be causal.

they would “never” support breaking democratic norms in 3.5 of the seven scenarios—a number that increases to 4.7 in the treatment group. Republicans and Democrats exhibit similar responses to the treatment with Democrats falling from 0.26 in control to 0.16 in treatment, and Republicans falling from 0.22 to 0.17.

We find a statistically significant treatment effect across all seven individual items in the self-subversion scale (see Supplemental Materials Figure S24); across all six of our party identification levels, from strong Democrat to strong Republican, though the effect is marginally significant for strong Republicans (see Supplemental Materials Figure S25); and across levels of general political knowledge (see Supplemental Materials Figure S26).

We think the treatment effect is important because it implies that the observational relationship we document in Study 1 may be causal. A key question that follows is how much of the observational relationship may be causal? We can shed light on this question by scaling the treatment effect by the degree of successful manipulation (Complier Average Causal Effect) using the answer participants gave in the final question of the ask-tell battery (order-randomized). Intriguingly, when we do so, the CACE estimate is 0.25—about the same as the regression estimate from Study 1—suggesting that much of the observational relationship may be causal. Of course, CACE estimates depend on numerous assumptions that may not be satisfied in this case (see Supplementary Materials for details and assumptions).

If fear of the other side’s 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 S27). We don’t find that the intervention changes respondents’ own policy views (5-item scale), but it may reduce Democrats’ perception that Republicans are extremists on a number of policy issues (Supplemental Materials Figure S28). It also appears to decrease respondents dehumanization of the other party and decrease their perception that the other party dehumanizes them (Supplemental Materials Figure S29). In a series of experiments, Broockman et al. (40) find that lowering affective polarization fails to increase support for democratic norms, despite these correlating in survey data. The results here suggest that the causal arrow could go in the other direction: changes in the perception of opposing partisans’ support for democratic norms may influence inter-partisan animosity.

We find no sign that respondents guessed the study’s purpose based on an open-ended question at the end of the survey, nor do we see evidence consistent with a demand or social desirability pressure account (see Supplemental Materials section on demand and Study 3). However, we do think that the treatment is likely working on a difficult-to-parse combination of injunctive and descriptive norms (41). In other words, we think the main mechanism behind the treatment effect is cognitive—learning about the attitudes of the other party—but other mechanisms likely contribute, including social pressure to appear democratic.

Study 2b: Fear 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 at the end of Study 2. We then examine the impact of the ask-tell intervention on these voting decisions. Unlike the previous analyses, this analysis is exploratory (not pre-registered), though we replicate it in Study 3. 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 in our study (chosen at random) and another who has “opposed” breaking the norm.

In order to model the real-world rhetoric that we suggest allows for democratic backsliding by aspiring autocrats, we told participants that the candidate who “supported” breaking a democratic norm did so “because they believe opposing partisans (Democrats or Republicans) have done the same.” By including this rhetoric, we also likely overcome a social desirability problem, which is that voters seem unlikely to support a subverting candidate on a survey—even if they would in the real world—unless the choice seems justifiable. Doing so likely contradicts people’s views about what it means to be a good citizen, and respondents likely want to look like good citizens on surveys. Since the ask-tell treatment should decrease their support for these candidates, we also can’t detect an effect if the control group won’t vote for the subverting candidate. Of course, this artificial primary design in no way captures the real world complexity of primary races, but potentially provides a proof of concept.

To summarize, Study 2b simply changes the dependent variable from respondents’ willingness to subvert democracy to respondents’ willingness to vote for a subverting candidate. Everything else is the same as in 2a, including the treatment and the sample.

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, $t(5,050)=6.9$, $p=0.003$, two-tailed (calculated with robust standard errors clustered at the respondent level), as Figure 6 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. Mediation analysis suggests that respondents’ own willingness to subvert democracy mediates the treatment effect (see Supporting Materials Tables S9 and S10) (42, 43). Of course, mediation analysis should be taken with a large grain of salt. This hypothetical candidate choice study in no way captures the complexity of real-world campaigns. It is merely a first step, providing suggestive evidence that reduced fears lead voters to prefer candidates who uphold democratic norms. The effect is also not large, but may be biased downward by social desirability bias. It also may not need to be large to matter in the real world, where elections are often closely contested.

Study 3: Demand and Social Desirability

To further address concerns that the treatment effect arises from respondents guessing the intent of the experiment (demand) or from respondents desiring to appear less subversive after exposure to the treatment (treatment-induced social desirability), we conduct a panel version of the study designed to minimize demand and experimentally induced social desirability. We administer the treatment in a first survey and collect the dependent variable in a second survey. We attempt to

Figure 6: Less Fear, Less Tolerance for Candidates Who Subvert Democracy—Study 2b. This figure shows the treatment-group mean minus the control-group mean with the dependent variable coded as follows: democracy-protecting candidate = 0, democracy-subverting candidate = 1, and neither candidate = 0.5. It reveals that, 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. 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 a handful only cast a vote in one primary). Each point shows the treatment effect estimate from a separate regression. The error bars show 95 percent (thin) and 68 percent (thick) confidence intervals and are calculated from standard errors clustered by respondent.

make the two surveys appear as “seemingly unrelated” as possible and to avoid highlighting the intent of either survey. We do so by not describing the second survey as a follow-up study, by changing the look and feel of the second survey, and by beginning and ending the surveys with other questions. We also administer this study on Mechanical Turk, where respondents typically take many surveys every day.

In a preregistered study, we recruited 3000 Mechanical Turk respondents for the first survey, which took place on May 9 and 10, 2022. After excluding those who failed the attention checks or who didn’t lean towards either party, we administered the experiment to 2,523 respondents. The first survey began with consent, then two attention questions, demographics, party affiliation, political knowledge, the seven questions about the other party’s willingness to subvert democratic norms, and finally more demographics. As in Study 2, half of the sample received feedback on the accuracy of their perceptions after each subversion question about the other party (the “ask-tell” treatment), and the other half did not. We then successfully re-surveyed 1,848 (73.2%) between May 9 and 14, 2022. We find no sign of differential attrition or meaningful imbalances (see Supporting Materials Table S6). We randomly assigned when the second survey became available, resulting in a minimum time between interviews of 1.4 hours, a maximum of 94 hours, and a mean of 47 hours. The second survey began with the seven self-subversion items (the main dependent variable), the seven subversion items about the other party (a manipulation check), the primary choice questions, and four ethnic-antagonism items (25). To check for order effects, about 10% of the sample received the battery of questions about opposing partisans first.

Even with these steps to reduce demand and social desirability, the treatment effect remained similar to the one found in Study 2. The intervention in wave one shifted the mean response on the wave-two seven-item subversion scale from 0.32 to 0.24 on the 0-1 range, a 25% change, and one that is highly unlikely to occur by chance, $t(1,843)=-6.8$, $p=1.1\text{e-}11$, two-tailed. The estimated effect remains the same in the 10% of the sample where the order of the two batteries was reversed; see Supplemental Materials Table S11. When we again scale the treatment effect by the degree of successful manipulation (Complier Average Causal Effect) using the small subset randomly assigned to first answer the questions about the other party’s subversion first, the estimate closely matches the regression estimate from Study 1 and the CACE estimate from Study 2: 0.24 (see Supplementary Materials for details and assumptions). Figure 7 shows the effect over time, revealing a consistent difference between the control group and the treatment group. Treatment and control lines trend upwards over time, but, interestingly, respondents randomly assigned to take the survey later fail to exhibit these upward trends, implying that time since the previous survey does not cause changes in subversion levels. Instead, this pattern appears to result from nonrandom variation in respondents’ tendency to take the survey earlier versus later. As in Study 2a, the treatment effect holds up across levels of general political knowledge and across levels of party identification (see Supplemental Materials Figures S30 and S31).

In an exploratory analysis (not preregistered), we replicate the results of the primary choice experiment, which follows Study 2 exactly, except that we measure the candidate choice in wave 2 and ask respondents about three hypothetical primary races (Figure 8). Interestingly, we find that Democrat and Republican respondents with higher levels of ethnic antagonism are more likely to support subverting democratic norms, but controlling for ethnic antagonism leaves the key findings in this paper—the observational findings and experimental findings—unchanged (see Supporting Materials Table S12).

In sum, ruling out demand and experimentally induced social desirability is difficult, but this study helps further allay concerns.

Figure 7: Study 3 Wave 1 ‘Ask-Tell’ Treatment Effect on Wave 2 Support for Subverting Democracy by Hours Since Treatment (N=1,845). The figure shows that, despite efforts to reduce demand and social desirability with our seemingly-unrelated studies approach, the treatment effect remains similar to Study 2. Presumably, the longer the time between interviews, the lower the likelihood that demand and experimentally-induced social desirability drive the results, but we don’t see the treatment effect obviously diminishing with time. Treatment and control lines trend upwards over time, but this reflects differences between participants based on when they chose to take the survey. When we analyze the data by when participants were randomly assigned to the survey, we don’t observe these upward trends, implying that time since the previous survey does not cause changes in subversion levels. Lowess lines with 95% confidence interval. The figure shows the treatment effect for all respondents, and this effect is similar among the 10% who first answered the questions about the other party in wave 2.

Figure 8: Less Fear, Less Tolerance for Candidates Who Subvert Democracy—Study 3. This figure shows that 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. This study replicates Study 2b, except that we measure the candidate choice in wave 2 and ask respondents to cast votes in three primary choice elections, rather than just two. See the description of Study 2b for details on the design and the analysis. The figure shows the treatment group mean minus the control group mean with the dependent variable coded as follows: democracy-protecting candidate = 0, democracy-subverting candidate = 1, and neither candidate = 0.5. The mean on the scale is around .34 for both parties. The Ns represent the total number of votes cast. Since respondents cast votes in three primary races, the number of respondents is one-third the N. Each point shows the treatment effect estimate from a separate regression. The error bars show 95 percent (thin) and 68 percent (thick) confidence intervals and are calculated from standard errors clustered by respondent.

Discussion

Across three studies, we examined whether voters support subverting democracy because they fear that the other side is already subverting it—the subversion dilemma. In the observational study, we find that partisans are willing to break democratic norms to the extent that they fear opposing partisans support undermining them, an association that holds even while controlling for the 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. We then further assessed whether this relationship is causal with two experimental studies, and found that reducing fears of the other side increases support for democratic norms. Reducing these fears, we also found, may translate into behavioral outcomes, such as voting for democracy-promoting candidates.

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 already support dismantling democracy.

If such a dynamic is indeed present in the US, we think these fears likely started with elite rhetoric, such as Trump’s claims that Democrats are trying to steal elections, a theme he has repeated relentlessly. This rhetoric, combined with Trump’s overstepping of democratic norms in the name of “saving democracy,” most notably during the January 6th Capitol Riot, has likewise prompted widespread fears among Democrats that Republicans are planning to undermine democracy.

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 norms, even in the face of would-be authoritarian opponents (3). Democratic parties facing authoritarian challenges in Colombia, Ecuador, and Bolivia (until recently) were able to stem democratic backsliding by refraining from calling for anything irregular during the aspiring authoritarian’s term, such as early removal from office (44, 45). Instead, efforts were made to counter specific democratic violations made by the incumbent (45).

When facing aspiring authoritarian leaders, opposing parties might focus on demonstrating their absolute commitment to democratic norms, particularly to the voters of the opposing party, who may then be more likely to keep their leaders in check. In the US context, these findings imply that convincing Republicans that Democrats are committed to democracy would reduce voter tolerance for backsliding among elected officials. This is the most important implication from our study.

Convincing voters that opposing partisans don’t support subverting democracy may not be easy, especially because aspiring authoritarians are likely to continue to foster fear of the other side. Research in international relations on the security dilemma suggests that tools such as third-party observation along with costly signals of good-faith are possible ways to resolve such dilemmas (11).

In the US, to signal to Republicans that they don’t support undermining democracy, Democrats could wait to pursue policies that Republicans perceive as tilting the playing field and contributing to an arms race, such as DC statehood or eliminating the filibuster. Non-partisan public interest groups could also consider acting as third-party guarantors on issues such as election reform and

monitoring. Higher levels of trust could also be pursued at a more grassroots level. In Study 1, we asked respondents whether they would be interested in making 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.” Another possible source of third-party guarantees on democratically normative behavior could come from a strengthened international democracy-promoting regime (46, 47).

This paper provides a novel framework for understanding why democracy-loving citizens nevertheless have voted away their democracies in countries throughout the world. Undoubtedly, policy and affective polarization and other factors previously 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.

Code and Data Availability Statement

The authors will make all data collected and all code used to analyze the data for this study available at the project's OSF link upon publication, or prior per request.

Ethical Compliance and Informed Consent

The authors have complied with all relevant ethical regulations in conducting human subjects research with approval from the IRB Committee at the University of California, Berkeley. The authors obtained informed consent from all participants in this study.

References and Notes

1. R. Wike, K. Simmons, B. Stokes, J. Fetterolf, Globally, broad support for representative and direct democracy. *Pew Research Center* (2017).
2. S. I. Lindberg, V. A. Boese, Autocratization changing nature? *Democracy Report 2022* (2022).
3. S. Levitsky, D. Ziblatt, *How democracies die* (Broadway Books, 2018).
4. M. Svobik, When polarization trumps civic virtue: Partisan conflict and the subversion of democracy by incumbents. *Available at SSRN 3243470* (2018).
5. N. Bermeo, On democratic backsliding. *Journal of Democracy*. **27**, 5–19 (2016).
6. L. Gamboa, Opposition at the margins: Strategies against the erosion of democracy in colombia and venezuela. *Comparative Politics*. **49**, 457–477 (2017).
7. M. S. Fish, Roeder Philip G, Hanson Stephen E, Fish M Steven, R. D. Anderson, Eds. (Princeton University Press, 2021).
8. B. R. Weingast, The political foundations of democracy and the rule of law. *The American Political Science Review*. **91**, 245–263 (1997).
9. R. Jervis, Cooperation under the security dilemma. *World Politics*. **30**, 167–214 (1978).
10. B. R. Posen, The security dilemma and ethnic conflict. *Survival*. **35**, 27–47 (1993).
11. J. D. Fearon, Rationalist explanations for war. *International organization*. **49**, 379–414 (1995).
12. D. Waldner, E. Lust, Unwelcome change: Coming to terms with democratic backsliding. *Annual Review of Political Science*. **21**, 93–113 (2018).
13. P. Pierson, E. Schickler, Madison's constitution under stress: A developmental analysis of political polarization. *Annual Review of Political Science*. **23**, 37–58 (2020).
14. J. Fishkin, A. Siu, L. Diamond, N. Bradburn, Is deliberation an antidote to extreme partisan polarization? Reflections on "america in one room." *American Political Science Review*. **115**, 1464–1481 (2021).

15. 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).
16. M. H. Graham, Does partisan identity reduce support for electoral fairness? An experimental test (2020).
17. M. Somer, J. L. McCoy, R. E. Luke, Pernicious polarization: Autocratization and opposition strategies. *Democratization*. **28**, 929–948 (2021).
18. M. S. Levendusky, N. Malhotra, (Mis) perceptions of partisan polarization in the american public. *Public Opinion Quarterly*. **80**, 378–391 (2016).
19. 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).
20. 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).
21. E. C. Cassese, Partisan dehumanization in american politics. *Political Behavior*, 1–22 (2019).
22. J. L. Martherus, A. G. Martinez, P. K. Piff, A. G. Theodoridis, Party animals? Extreme partisan polarization and dehumanization. *Political Behavior*, 1–24 (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. 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).
25. L. Bartels, Ethnic antagonism erodes republicans' commitment to democracy. *PNAS*. **117**, 22572–22579 (2020).
26. N. McCarty, *Polarization: What everyone needs to know* (Oxford University Press, 2019).
27. R. Putnam, *Making democracy work* (Princeton University Press, Princeton, 1993).
28. A. Przeworski, Democracy as an equilibrium. *Public Choice*. **123**, 253–273 (2005).
29. D. J. Trump, (2016), (available at <https://www.theguardian.com/us-news/video/2016/oct/18/us-presidential-election-rigged-donald-trump-wisconsin-video>).
30. L. Ingram, *Fox News*. **September 9** (2020).
31. T. Carlson, *Fox News*. **September 24** (2020).
32. L. Ingram, *Fox News*. **October 1** (2020).
33. E. Burnett, *CNN*. **September 8** (2020).
34. D. Lemon, *CNN*. **September 23** (2020).
35. E. Warren, (2020), (available at <https://twitter.com/ewarren/status/1309876174231949312>).

36. 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).
37. N. P. Kalmoe, L. Mason, in *National capital area political science association american politics meeting* (2019).
38. D. J. Ahler, Self-fulfilling misperceptions of public polarization. *The Journal of Politics*. **76**, 607–620 (2014).
39. J. Mernyk, S. Pink, J. Druckman, R. Willer, Correcting inaccurate metaperceptions reduces americans' support for partisan violence (2021).
40. D. Broockman, J. Kalla, S. Westwood, Does affective polarization undermine democratic norms or accountability? Maybe not. *American Journal of Political Science* (Forthcoming).
41. C. Bicchieri, *Norms in the wild: How to diagnose, measure, and change social norms* (Oxford University Press, 2016).
42. K. Imai, L. Keele, D. Tingley, A general approach to causal mediation analysis. *Psychological methods*. **15**, 309 (2010).
43. D. Tingley, T. Yamamoto, K. Hirose, L. Keele, K. Imai, Mediation: R package for causal mediation analysis (2014).
44. L. Gamboa, Opposition at the margins: Strategies against the erosion of democracy in colombia and venezuela. *Comparative Politics*. **49**, 457–477 (2017).
45. M. R. Cleary, A. Öztürk, When does backsliding lead to breakdown? Uncertainty and opposition strategies in democracies at risk. *Perspectives on Politics*. **20**, 205–221 (2022).
46. A. M. Matanock, How international actors help enforce domestic deals. *Available at SSRN 3100500* (2019).
47. S. D. Hyde, Democracy's backsliding in the international environment. *Science*. **369**, 1192–1196 (2020).
48. 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).

Supplementary Materials

Materials and Methods

Study 1

Study 1 is our observational study that includes 1,973 US respondents recruited via Lucid between July 15 and August 6, 2021. The survey was administered in the following order. Note that all question wording and images are included in following sections.

- Introduction Section
 - Informed Consent
 - Welcome Image
 - Attention Check 1 (Survey ended if failed.)
 - Attention Check 2 (Survey ended if failed.)
 - ANES 7-point Party Identification Scale (Survey ended for Independents that do not lean toward either party.)
- Observational Self-Subversion Intro Image
 - Self-Subversion Questions (Block presented in random order with Opposing Partisan Subversion Questions. Question order randomized within-block.)
 - Self-Subversion Comprehension Check Question (Survey ended if failed.)
- Observational Opposing Partisan Subversion Intro Image
 - Opposing Partisan Subversion Questions (Block presented in random order with Self-Subversion Questions. Question order randomized within-block.)
 - Opposing Partisan Subversion Comprehension Check Question (Survey ended if failed.)
- Observational Covariates Intro Image
 - Self Policy Polarization (Two questions randomly selected out of 5.)
 - Opposing Partisan Policy Polarization (Two questions randomly selected out of 5.)
 - Self Blatant Dehumanization of Opposing Partisans
 - Opposing Partisan Dehumanization of Self (meta-dehumanization)
 - Self Intent to vote for Trump
 - Self Political Knowledge Questions
 - Self Fear of Opposing Partisans
 - Self Feeling Thermometer towards Democrats and Republicans

The results are similar regardless of the order in which the question blocks and individual questions are given. 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 unweighted 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.

Studies 2a and 2b

Study 2 follows much the same logic as Study 1. Study 2 is our experimental study that includes 2,545 US respondents recruited via Lucid between September 15 and 29, 2021. The survey was administered in the following order. Note that all question wording and images are included in following sections.

- Introduction Section
 - Informed Consent
 - Welcome Image
 - Attention Check 1 (Survey ended if failed.)
 - Attention Check 2 (Survey ended if failed.)
- Experiment Intro Image
 - ANES 7-point Party Identification Scale (Survey ended for Independents that do not lean toward either party. Random Assignment to Treatment and Control Conditions.)
- Experiment Opposing Partisan Subversion Intro Image
 - TREATMENT GROUP
 - * Opposing Partisan Subversion Ask-Tell Treatment (Question order randomized within-block.)
 - * Opposing Partisan Subversion Comprehension Check Question
 - CONTROL GROUP
 - * Opposing Partisan Subversion Questions (Question order randomized within-block.)
 - * Opposing Partisan Subversion Comprehension Check Question
- Experiment Self-Subversion Intro Image
 - Self-Subversion Questions (Question order randomized within-block.)
 - Self-Subversion Comprehension Check Question
- Primary Candidate Choice Questions (Two)
- Experiment Covariates Intro Image
 - Self Policy Polarization (Two questions randomly selected out of 5.)
 - Opposing Partisan Policy Polarization (Two questions randomly selected out of 5.)
 - Self Blatant Dehumanization of Opposing Partisans
 - Opposing Partisan Dehumanization of Self (meta-dehumanization)
 - Self Political Knowledge Questions
 - Self Feeling Thermometer towards Democrats and Republicans
 - Social Class
 - Demand Check Question (Given to a random 1 out of 20 respondents)
- Experiment Final League Image

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.

Study 3

We make several changes to the experiment in Study 3 to reduce demand. We do not tell participants that they will be “tested” about how much they know, nor do we assign them scores or leagues. We also run this as a panel experiment and add covariate measures for four items on ethnic antagonism based on the four highest loading items in Bartels (25).

We administered the first wave of the survey experiment to 2,523 Mechanical Turk (mTurk) participants on May 9. We randomly assigned respondents to three recontact groups: May 9, 11, and 12, and we successfully resurveyed 1,848 participants.

We used mTurk’s qualification system to open up the second survey to these three groups over the course of the week. We also sent workers an email through mTurk to encourage them to take the study. The subject of the email was: “Take a 3-4 minute survey for \$.50”. The text of the email was: “Hello! Thanks for doing our HIT! To help with our research, could you do a 3-4 minute survey for \$.50? Here’s the HIT (sign in first!): XXXX. If that doesn’t work, search on Mechanical Turk for ‘Lenz’ (the requester) and submit for any hit you qualify for. If you have any trouble, please contact us. We appreciate your help with our research!” The language of this email changed slightly across the three releases to reduce confusion by respondents.

The survey was administered in the following order. Note that all question wording and images are included in following sections.

WAVE 1

- Introduction Section
 - Informed Consent
 - Welcome Image
 - Attention Check 1 (Survey ended if failed.)
 - Attention Check 2 (Survey ended if failed.)
- Demographics
 - Ethnicity
 - Race
 - Age
 - ANES 7-point Party Identification Scale (Survey ended for Independents that do not lean toward either party.)
 - Self Political Knowledge Questions (Random Assignment to Treatment and Control Conditions.)

- Panel Wave 1 Experiment Intro Image and Panel Wave 1 Opposing Partisan Subversion Intro Image
 - TREATMENT GROUP
 - * Opposing Partisan Subversion Ask-Tell Treatment (Question order randomized within-block.)
 - * Opposing Partisan Subversion Comprehension Check Question
 - CONTROL GROUP
 - * Opposing Partisan Subversion Questions (Question order randomized within-block.)
 - * Opposing Partisan Subversion Comprehension Check Question
- Panel Wave 1 Final Image

WAVE 2

- Introduction Section
 - Informed Consent
 - Self Political Knowledge Questions
- Panel Wave 2 Self-Subversion Text “Please let us know what YOU would support in American Politics. Your REAL OPINION matters! Watch out for attention checks!” (No Image.)
 - Self-Subversion Questions (Question order randomized within-block.)
 - Self-Subversion Comprehension Check Question
- Panel Wave 2 Opposing Partisan Subversion Text “We will now ask what you think most [OTHER PARTYs] would support in American politics.” (No Image.)
 - Opposing Partisan Subversion Questions (Question order randomized within-block.)
 - Opposing Partisan Subversion Comprehension Check Question
- Primary Candidate Choice Questions (Three)
- Experiment Covariates Section
 - Ethnic Antagonism Questions
- Thanks Page and Open-Ended Space for Feedback (No Image.)

Note that to check for order effects, about 10% of the sample received the “Opposing Partisan Subversion Questions” first, and that the data from this randomly chosen group was not included in the main analysis of Study 3. As in Studies 1 and 2, 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.

Demographic information about the participants in Study 3 can be found in Table S4. The magnitude and type of demographic differences do not seem to threaten the ability to make meaningful inferences about the main variables of interest.

The Ask-Tell Treatment

The flow of the ask-tell treatment is depicted in Figure 3 in the main text. Respondents in the treatment condition in Studies 2 and 3 are given feedback after they answer each question about the extent to which they believe opposing partisans support a given democratic subversion.

For example, a Republican in the treatment condition would see a question as in Figure S1.

Do **MOST DEMOCRATS** support freezing the social media accounts of **REPUBLICAN JOURNALISTS**?

☐ Never
☐ Probably Not
☐ Probably
☐ Definitely

Figure S1: Example Question of Opposing Partisan Subversion (Republican Example).

If the respondent gets the correct answer, they would see Figure S2. If the respondent gets the incorrect answer, they would see Figure S3. In the unusual case that the correct answer is “probably not” rather than “never,” and the respondent underestimates (rather than overestimates) opposing partisan subversion, they would see Figure S4.

This is repeated for all seven questions in the opposing-partisan subversion battery. The answers to these questions are generated by the modal response we received from partisans on each question in observational Study 1. Democrats answered “Never” to four self-subversion questions and “Probably Not” to three, while Republicans answered “Never” to 5 self-subversion questions and “Probably Not” to two.

Figure S2: Example of a Correct Answer Treatment Image (Republican Example).

Figure S3: Example of an Incorrect Answer Treatment Image (Republican Example).

Figure S4: Example of an Underestimate Answer Treatment Image (Republican Example).

Demand Effects and Social Desirability

Study 3 addresses demand and social desirability effects. Besides the evidence in Study 3, there are 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 asks 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 group's rates. Of course, none of this definitively rules out demand effects.

Besides the evidence in Study 3, there are also reasons to believe experimentally induced socially desirable responding is not driving the result in Study 2. Although the treatment influences respondent perceptions of outpartisans, they nevertheless still see outpartisans as more willing to subvert democracy than they report being themselves. This is true even when we look only at the last item we asked them about (so they would have received 6/7 of the ask-tell treatments). On this last item, mean perception on the seven-item scales is about twice as high for the other party than for themselves (.39 versus 0.17 on the 0-1 scale). So, respondents could admit to more willingness to subvert and still keep considerable distance between themselves and the other party. Second, we find a very similar observational and experimental relationship between these variables (a linear relationship of about 0.25 on the 0-1 range). This similarity seems inconsistent with a social desirability account, since we are not inducing social pressure in the observational study, and yet it nevertheless yields a similar estimate.

We also think it is 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. Our goal here is to experimentally examine this relationship in a survey context, which we find is consistent with our theory despite respondent inattention and the fast pace of survey taking.

Scale Construction

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 S4 for descriptive statistics).

Missing Data

So that the regression estimates for Study 1 do not omit respondents because of missing values on the control variables in Table 1, we impute missing values using demographics (using the robust linear regression with M-estimators from 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.

Significant Testing

For the significant tests of the main experimental effects in Study 2a and 3, we use HC3 standard errors. Given the sample sizes and the effect sizes for these tests, the choice of standard errors is inconsequential.

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 outpartisans to subvert, not through other observed or unobserved variables (the exclusion restriction). Table S7 presents the reduced form regression and Table S8 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 presented these items in random order to respondents, and we 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 post-treatment, 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 between 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 Study 1 and 2 at <https://osf.io/vnd4g> and Study 3 at https://aspredicted.org/7PJ_6TQ.

After filing the first preregistration, we ran a pilot experiment on a demographically 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 S33 shows the manipulation check and Figure S34 shows the treatment effect on willingness to subvert democracy. Both show large and statistically significant differences for both Democrats and Republicans.

Slight differences between our pre-registration and Study 1 and 2 are as follows. Our pre-registration listed eight subversion items, but we ended up dropping one item before running the experimental portion of the study. This question asked “Would YOU support a [OWN PARTY] governor ruling by executive order if [OTHER PARTY]s don’t cooperate?” Responses to this question were so positive (high levels of support), that this indicated to us that this question was not understood or the action was not likely perceived as normatively prohibited in the US democratic context.

Additionally, we changed the wording of two other questions for simplicity. We preregistered “Would YOU support a [OWN PARTY] governor banning far-[OTHER SIDE] group rallies in the state capital?” and “Would YOU support a [OWN PARTY] governor ignoring controversial court rulings by [OTHER PARTY] judges?” We simplified these questions to make them easier to understand by removing the governor and asking about direct support for the policy. So the questions instead read: “Would YOU support banning FAR-[OTHER SIDE] rallies in the state capital?” and “Would YOU support ignoring controversial rulings by [OTHER PARTY] JUDGES?”

Because we changed these questions, there was no other question that mentioned a governor except for the question about executive orders. So we also removed this question for the sake of simplicity, so that respondents are asking about their support for certain actions rather than their support mediated through a hypothetical representative.

The preregistration notes that the index for subversion and opposing partisan subversion would include 8 questions coded 0-3 for a total index range of 0-24. In reality we have 7 questions coded 0-3 which are added together and normalized to 0-1 for ease of interpretation.

Because of concerns about attention on Lucid, we also added an additional attention check.

Another note about our preregistration is that it doesn’t explicitly list the covariates that we measure in order to test alternate hypotheses in our observational survey and to test moderators in our experiment. These include intent to vote for Trump and fear of opposing partisan subversion in Study 1, as well as policy polarization, opposing partisan policy polarization, blatant dehumanization, opposing partisan dehumanization, level of political knowledge, and feeling thermometer towards Democrats and Republicans in Study 1 and 2.

In order to reduce the length of our survey and experiment in Study 1 and 2, we gathered all demographic variables that we preregistered except for Party Identification and Strength of Partisanship directly from Lucid rather than from respondents.

Due to uncertainties about our project budget, we conservatively estimated that we would include 500 people in Study 1 and 1000 in Study 2, though due to increased funds we were able to include about 2,000 (1,973) in Study 1 and about 2,5000 (2,545) in Study 2.

In our pre-analysis plan we wrote we would estimate the CACE by subtracting the numerical level of subversion the participants assigned to opposing partisans in the final question minus the answer they gave in the first question. We decided that a more statistically sound estimate of the CACE would use only the answer the participants gave on the final question. For more information about the CACE see the section on “Complier Average Treatment Effect.”

Question Wording

Self-Reported Subversion. In Studies 1, 2, and 3, 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”. We ask the following questions where “OWN SIDE” or “OTHER SIDE” can be “LEFT” or “RIGHT” and “OWN PARTY” or “OTHER PARTY” can be “DEMOCRAT” or “REPUBLICAN”.

- (1) “Would YOU support banning FAR-[OTHER SIDE] rallies in the state capital?”
- (2) “Would YOU support ignoring controversial rulings by [OTHER PARTY] JUDGES?”
- (3) “Would YOU support freezing the social media accounts of [OTHER PARTY] JOURNALISTS?”
- (4) “Would YOU support reducing the number of voting stations in towns that support [OTHER PARTY]s?”
- (5) “Would YOU support laws that would make it easier for [OWN PARTY]s (and harder for [OTHER PARTY]s) to get elected?”
- (6) “Would YOU support using violence to block major [OTHER PARTY] laws?”
- (7) “Would YOU support significantly reinterpreting the Constitution in order to block [OTHER PARTY] policies?”

Opposing Partisan Subversion Questions. In Studies 1, 2, and 3, we measure beliefs about opposing partisan willingness to subvert democratic norms. In the treatment group of Studies 2 and 3, respondents receive feedback after answering each question. Questions are randomized for each participant. Response options are “never”, “probably not”, “probably”, and “definitely”. We ask the following questions where “OWN SIDE” or “OTHER SIDE” can be “LEFT” or “RIGHT” and “OWN PARTY” or “OTHER PARTY” can be “DEMOCRAT” or “REPUBLICAN”.

- (1) “Would MOST [OTHER PARTY]s support banning FAR-[OWN SIDE] rallies in the state capital?”
- (2) “Would MOST [OTHER PARTY]s support ignoring controversial court rulings by [OWN PARTY] JUDGES?”
- (3) “Would MOST [OTHER PARTY]s support freezing the social media accounts of [OWN PARTY] JOURNALISTS?”
- (4) “Would MOST [OTHER PARTY]s support reducing the number of voting stations in towns that support [OWN PARTY]s?”
- (5) “Would MOST [OTHER PARTY]s support laws that would make it easier for [OTHER PARTY]s (and harder for [OWN PARTY]s) to get elected?”
- (6) “Would MOST [OTHER PARTY]s support using violence to block major [OWN PARTY] laws?”

- (7) “Would MOST [OTHER PARTY]s support significantly reinterpreting the Constitution in order to block [OWN PARTY] policies?”

Attention Checks. At the beginning of each survey we ask participants two attention check questions. The first question asks “For our research, please click BOTH ‘I understand’ AND ‘I don’t understand’.” Participants are disqualified unless they select both options. We then say “In order to proceed, we request that you complete this attention check. Please select ‘I am a robot’ below.” Participants are disqualified if they instead select “I am not a robot”.

Comprehension Checks. In all instances where we ask the seven self-subversion or opposing-partisan subversion questions, we include a comprehension check randomly placed among the questions. The comprehension check asks “These questions are asking me about:” to which possible answers are “What I would support in American politics,” “What most Democrats would support in American politics,” and “What most Republicans would support in American politics.” In observational Study 1, participants are excluded from the survey for answering one of these questions incorrectly. In experimental Studies 2 and 3, participants are given a warning to pay better attention for answering these questions incorrectly.

Self-Reported Policy Questions. To measure policy extremism, Studies 1 and 2 ask respondents to answer two randomly selected questions from 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 of policy extremism, asking instead how respondents believe most opposing partisans would respond to two randomly selected questions. Question order is randomized for each respondent. Response options are “Strongly Disagree”, “Disagree”, “Somewhat Disagree”, “Neither Agree nor Disagree”, “Somewhat Agree”, “Agree”, or “Strongly Agree”. 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 meta-dehumanization (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 (19, 48). This is measured using the “Ascent of Man” scale, developed as a tool to measure blatant dehumanization by Kteily et.al. (2016). Figure S5 presents an example of these questions for Democratic respondents. Metadehumanization uses the same image but instead it would ask Democrats, “Where would MOST REPUBLICANS place YOU on this image?”

Political Knowledge Questions. To gauge political knowledge, we ask participants in Studies 1 and 2 the following political knowledge questions in random order.

Where would you place **MOST REPUBLICANS** on this image?

Figure S5: Example of Dehumanization question.

- “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.”
- “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.”
- “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. Figure S6 shows the format of these questions.

Figure S6: Feeling thermometer questions.

Primary Election Questions. In Studies 2 and 3, we add an exploratory measure of partisan willing-

ness 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 SIDE] 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 SIDE] 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.”

Ethnic Antagonism Questions. In Study 3 we include four items on ethnic antagonism based on the four highest-loading items in Bartels (25). These are “Immigrants contribute a great deal to American society and culture” and “Discrimination against whites is as big a problem today as discrimination against blacks and other minorities,” with the response options “strongly agree,” “agree,” “neither; unsure,” “disagree,” and “strongly disagree.” And, “How favorable or unfavorable do you feel towards these groups?” for “immigrants” and “Muslims” with sliders from 0 to 10 and the endpoints labeled “extremely unfavorable feelings” and “extremely favorable feelings.”

Social Class Questions. We included social class questions in previous pilot observational surveys and did not find any meaningful correlations with our variables of interest. We did tack on one social class question in Study 2 as a pilot for another study. The question was asked after all other data was gathered and is as follows: “Which social class do you identify with?” Answer choices include “Poor,” “Working class,” “Middle class,” or “Upper Class.” If the respondent chooses “Working Class” then they are asked “How important is being working class to you?” with the answer choices “Very important,” “Somewhat important,” “Neither important nor unimportant,” “Somewhat unimportant,” or “Not important at all.”

Survey Images

The following images are used in the survey and survey experiments as noted in the previous descriptions of Study 1, 2, and 3. The use of the cartoon character throughout the surveys is meant to increase attention and experimental realism with participants. The cartoon was not used in Wave 2 of Study 3 (the panel experiment) as to maintain the greatest possible difference between Wave 1 and Wave 2 of the experiment.

Hello and welcome!

First we need to ask you a few questions to see if you are eligible to take the survey.

Please, please **read carefully** or you won't be eligible for this survey.

Figure S7: Welcome Image

We will now ask about what YOU would support in American politics.

READ THE QUESTIONS CAREFULLY TO THE END.

Failure to answer **TRICK QUESTIONS** correctly will **DISQUALIFY** you from the survey.

Your **REAL OPINION** matters!

Figure S8: Observational Self-Subversion Intro Image

We will now ask what you think **MOST DEMOCRATS** would support in American politics.

READ THE QUESTIONS CAREFULLY TO THE END.

Failure to answer **TRICK QUESTIONS** correctly will **DISQUALIFY** you from the survey.

Your **REAL OPINION** matters!

Figure S9: Observational Opposing Partisan Subversion Intro Image (Example for Republicans)

Please let us know what you think about a few more questions.

Figure S10: Observational Covariates Intro Image

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!

Figure S11: Experiment Intro Image

Figure S12: Experiment Opposing Partisan Subversion Intro Image (Example for Republicans)

Figure S13: Experiment Self-Subversion Intro Image

Figure S14: Experiment Final League Image (Example for Silver)

Figure S15: Panel Wave 1 Experiment Intro Image

Figure S16: Panel Wave 1 Opposing Partisan Subversion Intro Image (Example for Republicans)

Figure S17: Panel Wave 1 Final Image

Table S1: Demographic Information for Study 1, 2, and 3 Participants

	Study 1	Studies 2a and 2b	Study 3	Benchmark
Sample	Lucid	Lucid	mTurk	ACS 2019
N	1,973	2,545	2,523	12,686,854
Gender				
Female	54	52		52
Age				
18-24	11	13	6	11
25-34	13	17	32	15
35-49	26	24	28	22
50-64	32	28	28	27
65+	18	17	6	25
Race				
Non-Hispanic White	72	68	74	70
Black	8	10	9	10
Asian + Other	7	8	10	8
Hispanic	13	14	13	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 a few 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. For Study 3, we only asked about age, race, and ethnicity to keep the survey brief. We use the 2019 ACS, rather than the 2020, because of Census warnings about Covid-19 interfering with the 2020 ACS.

Tables and Figures

Table S2: Descriptive Statistics for Study 1. All variables rescaled to 0-1.

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 Studies 2a and 2b. All variables rescaled to 0-1.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Treatment	2,545	0.496	0.500	0	0	1	1
Willingness to Subvert Democracy (7-item scale)	2,545	0.205	0.200	0.000	0.048	0.333	1.000
Perceived Willingness of the Other Party to Subvert (7-item scale)	2,545	0.519	0.244	0.000	0.333	0.714	1.000
Party Identification (7-point)	2,545	0.454	0.397	0.000	0.000	0.833	1.000
Party Identification Strength (7-point,folded)	2,545	0.631	0.398	0	0.5	1	1
Policy Extremism (5-item scale)	2,516	0.627	0.209	0.000	0.467	0.800	1.000
Perceived Policy Extremism of Opposing Partisans (5-item scale)	2,512	0.677	0.214	0.000	0.533	0.833	1.000
Age	2,545	0.386	0.239	0.000	0.176	0.581	1.000
Education	2,545	0.393	0.326	0	0	0.5	1
Dehumanization of Other Party	2,508	0.194	0.172	0.000	0.050	0.347	1.000
Perception of the Other Party's Dehumanization of Your Party	2,508	0.253	0.183	0.000	0.090	0.402	1.000
Vote for Subverting Candidate in Primary 1	2,529	0.239	0.368	0.000	0.000	0.500	1.000
Vote for Subverting Candidate in Primary 2	2,523	0.245	0.375	0.000	0.000	0.500	1.000

Table S4: Descriptive Statistics for Study 3. All variables rescaled to 0-1. To minimize attrition, we asked a limited set of demographics.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Treatment	2,523	0.506	0.500	0	0	1	1
Willingness to Subvert Democracy (Wave 2, 7-item scale)	1,845	0.279	0.239	0.000	0.095	0.429	1.000
Perceived Willingness of the Other Party to Subvert (Wave 1, 7-item scale)	2,519	0.478	0.244	0.000	0.286	0.667	1.000
Perceived Willingness of the Other Party to Subvert (Wave 2, 7-item scale)	1,814	0.508	0.278	0.000	0.286	0.714	1.000
Hours between Wave 1 and Wave 2 Surveys	1,848	46.882	22.560	1.420	30.174	67.793	94.008
Vote for Subverting Candidate in Primary 1 (Wave 2)	1,770	0.313	0.427	0.000	0.000	0.500	1.000
Vote for Subverting Candidate in Primary 2 (Wave 2)	1,773	0.300	0.426	0.000	0.000	0.500	1.000
Vote for Subverting Candidate in Primary 2 (Wave 2)	1,757	0.297	0.422	0.000	0.000	0.500	1.000
Age	2,522	0.413	0.265	0.000	0.184	0.592	1.000
White	2,518	0.851	0.356	0.000	1.000	1.000	1.000
Hispanic	2,523	0.128	0.335	0	0	0	1

Table S5: Balance Table for Study 2. All variables rescaled to 0-1. We find no sign of differential attrition: p.value = 0.69.

	Control (N=1282)		Treatment (N=1263)		Diff. in Means	p
	Mean	Std. Dev.	Mean	Std. Dev.		
Party Identification (7-point)	0.458	0.397	0.449	0.398	-0.009	0.576
Party Identification Strength (7-point,folded)	0.630	0.394	0.633	0.402	0.003	0.862
Age	0.387	0.235	0.385	0.244	-0.002	0.874
Education	0.397	0.322	0.390	0.329	-0.007	0.583
Household Income	0.171	0.192	0.166	0.183	-0.005	0.513
Female	0.512	0.500	0.526	0.500	0.013	0.504
White	0.677	0.468	0.675	0.469	-0.002	0.894
Hispanic	0.147	0.355	0.137	0.344	-0.010	0.451

Table S6: Balance Table for Study 3. To minimize attrition, we asked a limited set of demographics. All variables rescaled to 0-1. We find no sign of differential attrition: p.value = 0.38.

	Control (N=1246)		Treatment (N=1277)		Diff. in Means	p
	Mean	Std. Dev.	Mean	Std. Dev.		
Party Identification (7-point)	0.389	0.383	0.394	0.392	0.005	0.731
Party Identification Strength (7-point,folded)	0.755	0.257	0.771	0.257	0.016	0.115
Age	0.404	0.263	0.422	0.266	0.019	0.078
White	0.847	0.360	0.854	0.353	0.007	0.636
Hispanic	0.121	0.326	0.135	0.342	0.014	0.283

Table S7: 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.

	<i>Dependent variable:</i>	
	Willingness to Subvert Democracy (7-item index)	
	(1)	(2)
Treatment	−0.076*** (0.008)	−0.077*** (0.008)
Partisan Identity Strength		0.072*** (0.010)
Age		−0.160*** (0.016)
Education		0.004 (0.012)
Constant	0.243*** (0.005)	0.258*** (0.010)
Observations	2,545	2,545
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 S8: 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.

	<i>Dependent variable:</i>	
	Willingness to Subvert Democracy (7-item scale)	
	(1)	(2)
Perceived Willing. of Oppo. Partisans to Subvert Dem. (item 7/7)	0.248*** (0.025)	0.249*** (0.024)
Partisan Identity Strength		0.055*** (0.010)
Age		−0.177*** (0.016)
Education		0.008 (0.012)
Constant	0.086*** (0.013)	0.116*** (0.014)
Observations	2,545	2,545
R ²	0.050	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

Table S9: Study 2b Models for Mediation Analysis with Pretreatment Covariates. In Model 1, the dependent variable is Willingness to Subvert Democracy. In Models 2-5, the dependent variable is Vote for Subverting Candidate. As described in the text, Vote for Subverting Candidate is coded 'No' 0, 'Neither' 0.5, and 'Yes' 1. The Willingness to Subvert Democracy is the seven-item scale, which is always rescaled to 0-1 and so that higher values indicate greater willingness to subvert democracy. Unlike the main estimates in the paper, the standard errors in this table are not clustered by respondent for the purposes of the mediation analysis. All variables are rescaled 0-1. The mediation analysis uses Models 1 and 5. See the next table for the mediation analysis. The results of the mediation analysis are substantively the same when we exclude pretreatment covariates. Intercept omitted.

	<i>Dependent variable:</i>				
	Willingness to Subv. Dem. Mediator Model		Vote for Subv. Candidate Outcome Model		
	(1)	(2)	(3)	(4)	(5)
Willingness to Subv. Dem. (7-item scale)				0.637*** (0.025)	0.601*** (0.026)
Treatment (0/1)	−0.077*** (0.005)	−0.035*** (0.010)	−0.037*** (0.010)	0.013 (0.010)	0.009 (0.010)
Republican (0/1)	0.004 (0.006)		0.007 (0.011)		0.005 (0.010)
Education Years	0.004 (0.008)		−0.039* (0.016)		−0.041** (0.015)
White	−0.028*** (0.007)		−0.060*** (0.014)		−0.043** (0.014)
Party Ident. Strength (7-point,folded)	0.070*** (0.007)		0.091*** (0.013)		0.049*** (0.013)
Age	−0.151*** (0.012)		−0.161*** (0.023)		−0.071** (0.022)
Household Income	−0.026 (0.014)		−0.028 (0.028)		−0.013 (0.026)
Female	−0.017** (0.005)		−0.011 (0.010)		−0.0004 (0.010)
Hispanic	−0.018 (0.010)		−0.047* (0.018)		−0.035* (0.017)
Observations	5,052	5,052	5,052	5,052	5,052
R ²	0.093	0.002	0.029	0.115	0.124
Adjusted R ²	0.091	0.002	0.027	0.115	0.122
Residual Std. Error	0.191	0.371	0.367	0.350	0.348

Note:

Standard Errors in Parentheses

* p<0.05; ** p<0.01; *** p<0.001

Table S10: Study 2b Mediation Analysis. This table reports model-based mediation analysis from the R 'mediation' package by Tingley et al. It examines whether the ask-tell effect on voting for the subverting primary candidate is mediated by the ask-tell effect on respondents' willingness to subvert democracy. See the previous table for the mediator (Column 1) and outcome model (Column 5). The results of the mediation analysis are substantively the same when we exclude pretreatment covariates. All variables are coded on the 0-1 scale. Nonparametric bootstrap confidence intervals with the percentile method. Simulations: 1000.

	Coefficients	95% CI Lower	95% CI Upper	p-values
Average Causal Mediated Effect (ACME)	-0.0463	-0.0545	-0.039	0.0000
Average Direct Effect (ADE)	0.0095	-0.0096	0.031	0.3280
Total effect	-0.0368	-0.0565	-0.016	0.0000
Proportion mediated	1.2568	0.8291	2.932	0.0000

Table S11: Treatment Effect Estimates for Study 3

	DV: Willingness to Subvert Democracy (Wave 2)			DV: Opponents' Will. to Subv. Dem. (Wave 2)		
	WS 1st	WS 2nd	Both	WS 1st	WS 2nd	Both
	1	2	3	4	5	6
Treatment	-0.077*** (0.012)	-0.058 (0.036)	-0.075*** (0.011)	-0.168*** (0.013)	-0.343*** (0.039)	-0.185*** (0.012)
Num.Obs.	1669	176	1845	1637	177	1814
R2	0.026	0.015	0.025	0.094	0.311	0.111
R2 Adj.	0.025	0.010	0.024	0.093	0.307	0.110
F	44.289	2.640	46.828	168.748	78.491	225.900
RMSE	0.24	0.23	0.24	0.26	0.25	0.26
Std.Errors	HC3	HC3	HC3	HC3	HC3	HC3

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

Note: In the Wave 2 survey, most respondents answered the seven items about their own willingness to subvert democracy first, but about 10 percent first answered the seven items about opponents' willingness to subvert democracy. This table shows the treatment effects for both outcomes by this order. WS 1st indicates that respondents saw the willingness to subvert battery first. WS 2nd indicates that respondents saw the willingness to subvert second. Columns 1 and 2 provide the justification for combining both, the estimate of which is shown in Column 3. Columns 4 and 5 show what appears to be a large increase in the treatment effect on the manipulation check variable, opponents' willingness to subvert democracy, when we ask it second. Each column shows a separate least-squares regression with the intercept omitted. Standard errors in parentheses.

Table S12: Observational Regressions for Study 3 on Addressing Order Effects and Ethnic Antagonism in Explaining Support for Subverting Democracy.

	DV: Willingness to Subvert Democracy (Wave 2)					
	Control Group		Treatment & Control Groups			
					Republicans	Democrats
	1	2	3	4	5	6
Opponents' Will. to Subv. Dem. (Wave 1)	0.252*** (0.037)	0.224*** (0.022)	0.218*** (0.022)	0.218*** (0.021)	0.173*** (0.037)	0.266*** (0.026)
Ethnic Antagonisms			0.117*** (0.028)	0.128*** (0.027)	0.148** (0.051)	0.298*** (0.039)
Age				-0.134*** (0.020)	-0.161*** (0.033)	-0.093*** (0.025)
White				-0.012 (0.015)	-0.044 (0.032)	0.010 (0.016)
Hispanic				0.106*** (0.016)	0.186*** (0.026)	0.054** (0.020)
Num.Obs.	835	1841	1841	1841	688	1153
R2	0.054	0.053	0.062	0.113	0.154	0.135
R2 Adj.	0.052	0.053	0.061	0.110	0.148	0.131
F	47.100	103.181	60.866	46.652	24.848	35.763
RMSE	0.23	0.23	0.23	0.23	0.23	0.22

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

Note: Column 1 addresses a potential concern about order effects in Study 1. In Study 3, we can examine the relationship between Willingness to Subvert and Opponents' Willingness to Subvert when these are measured in separate surveys before any other questions about subverting democracy, therefore eliminating concerns about order effects. Column 1 does so in the control group and for respondents who answered the questions about their own willingness to subvert first in Wave 2. The Column 1 estimate reveals that the key coefficient remains very similar in size to that in Study 1 when order concerns are eliminated: about 0.25. In Column 1, we restrict the sample to the control group and to those who were asked the Willingness to Subvert Democracy battery first (most of the sample). In Column 2, we show this relationship in the full sample, which reveals that it is about the same. Since including this group does not change the core relationship much, we conduct the remaining regressions in the full sample to increase precision. Columns 3-6 examine whether an Ethnic Antagonism battery based on Bartels (2020) can account for the key relationship documented in this paper. These columns show that it is a strong predictor of willingness to subvert democracy, but it doesn't explain away the key relationship documented in this paper, which is consistent with the experimental findings in Studies 2 and 3. Each column shows a separate least-squares regression with the intercept omitted. Standard errors in parentheses.

Figure S18: 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 S19: 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 S20: Study 1 perceived willingness of opposing partisans to subvert democracy by strength of partisanship.

Figure S21: 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 Figure 1 note for more details on the seven scenarios we asked respondents about. Both variables coded to the 0-1 range.

Figure S22: 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 by whether we asked the self subversion questions first or second, that is, before or after the questions about the perceived willingness of opposing partisans to subvert. However, in Study 3, we can assess this relationship while avoiding problems with order effects because we ask these questions in separate waves. In the control condition in Study 3, we recover the same average relationship as Study 1. Point size reflects the number of respondents and coloration reflects willingness to subvert.

Figure S23: Fear of opposing partisans and perceived willingness of opposing partisans to subvert democracy in Study 1. The fear question asked: ‘How afraid are you that [Other party: Democrats/Republican] are bending the rules to help them win elections? (By bending the rules, we mean encouraging voter fraud, not accepting election results, limiting voting stations in places that vote for the other party, gerrymandering districts, and making it harder for people to vote.)’ We gave respondents five response options from ‘Not afraid at all’ to ‘Very afraid’. The figure shows the average responses to this question (rescaled 0-1) for each bin.

Figure S24: 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. All variables coded to a 0-1 range. Each point shows the treatment effect estimate from a separate regression. The error bars show 95 percent confidence intervals.

Figure S25: Study 2 treatment effect on the seven-item self subversion scale by strength of partisanship. Partisanship was asked pretreatment. All variables coded to a 0-1 range. Each point shows the treatment effect estimate from a separate regression. The error bars show 95 percent confidence intervals.

Figure S26: 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. All variables coded to a 0-1 range. Each point shows the treatment effect estimate from a separate regression. The error bars show 95 percent confidence intervals.

Figure S27: Study 2 treatment effect on the feeling thermometers and perception of the other party's feeling about you. All variables coded to a 0-1 range. Each point shows the treatment effect estimate from a separate regression. The error bars show 95 percent confidence intervals.

Figure S28: Study 2 treatment effect on respondents' 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. All variables coded to a 0-1 range. Each point shows the treatment effect estimate from a separate regression. The error bars show 95 percent confidence intervals.

Figure S29: Study 2 treatment effect on respondent's dehumanization and meta-dehumanization. Higher values indicate more dehumanization. All variables coded to a 0-1 range. Each point shows the treatment effect estimate from a separate regression. The error bars show 95 percent confidence intervals.

Figure S30: Study 3 treatment effect on the seven-item self subversion scale by number of items correct on a 3-item general political knowledge scale. The figure shows that lower knowledge individuals may exhibit a larger treatment effect. Note that some items political knowledge scale were asked posttreatment. All variables coded to a 0-1 range. Each point shows the treatment effect estimate from a separate regression. The error bars show 95 percent confidence intervals.

Figure S31: Study 3 treatment effect on the seven-item willingness to subvert democracy scale, measured in wave 2, by strength of partisanship. Partisanship was asked pretreatment. All variables coded to a 0-1 range. Each point shows the treatment effect estimate from a separate least-squares regression. The error bars show 95 percent confidence intervals.

Figure S32: Study 3 Wave 1 'Ask-Tell' Treatment Effect on Wave 2 Support for Subverting Candidate by Hours Since Treatment. The loess lines. We asked each respondent to vote in three same-party primaries. We exclude confidence intervals because they won't be clustered by respondent.

Figure S33: 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. (N=668)

Figure S34: 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. ($N=668$)