nature human behaviour

Article

https://doi.org/10.1038/s41562-023-01594-w

Why voters who value democracy participate in democratic backsliding

Received: 16 June 2022

Accepted: 3 April 2023

Check for updates

Alia Braley[1] ✉, Gabriel S. Lenz[1], Dhaval Adjodah[2], Hossein Rahnama[2,3] & Alex Pentland[2,3]

Around the world, citizens are voting away the democracies they claim to cherish. Here we present evidence that this behaviour is driven in part by the belief that their opponents will undermine democracy first. In an observational study ($N = 1{,}973$), we find that US partisans are willing to subvert democratic norms to the extent that they believe opposing partisans are willing to do the same. In experimental studies ($N = 2{,}543$, $N = 1{,}848$), we revealed to partisans that their opponents are more committed to democratic norms than they think. As a result, the partisans became more committed to upholding democratic norms themselves and less willing to vote for candidates who break these norms. These findings suggest that aspiring autocrats may instigate democratic backsliding by accusing their opponents of subverting democracy and that we can foster democratic stability by informing partisans about the other side's commitment to democracy.

Around the world, antidemocratic leaders are convincing their supporters to vote away their political rights. While 78% of the world's population reports wanting to live in a representative democracy, democracies continue to erode, with 70% of the 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, citizens who claimed to support democracy the most were no more likely to vote for a democratic candidate[4].

The puzzle deepens when one considers that the modal form of autocratization today is democratic backsliding, in which democracies die a slow death, leaving years for a democracy-loving public to hold their representatives accountable[3,5,6]. Why, then, is democracy slipping away from so many citizens across various regions, cultures and socio-economic conditions[2,7,8]?

Some theories attempt to address this puzzle. For instance, some studies find that citizens' support for democracy in the abstract often fails to translate into support for democracy in the specific[9]. Additionally, citizens may see some power grabs as consistent with an implicit majoritarian definition of democracy[10–14]. They may also prioritize social, political and economic identities or policies over democracy[3,4].

In this paper, we provide an additional explanation for this puzzle that can help make sense of why democracy-loving citizens sometimes eschew widely valued democratic norms, such as those central to free and fair elections. Because democratic survival depends on mutual cooperation[15–17], it resembles the prisoner's dilemma game: if one party suspects the other is defecting, then the best response may be to defect. If citizens believe that opposing partisans will not hold their representatives in check, then they have an incentive to give their politicians leeway to do whatever it takes to save democracy from their opponents. We call this scenario the subversion dilemma: citizens who want to live in a democracy may come to tolerate defection by their representatives to save democracy from their opponents.

This provides would-be authoritarians with a powerful weapon against democracy. They use propaganda to convince their supporters that the other side is undermining democracy. As their supporters come to believe that the other side is defecting, they are more willing to tolerate their leaders' antidemocratic actions, which are seen as merely levelling the playing field now tilted against them.

In the US context, Donald Trump spread misinformation about Democrats subverting democracy from the start. Early in his 2016 campaign, his website stated, "Help Me Stop Crooked Hillary from Rigging

[1]Travers Department of Political Science, University of California, Berkeley, Berkeley, CA, USA. [2]MIT Connection Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. ✉e-mail: alia_braley@berkeley.edu

this Election!"[3]. Throughout the 2016 campaign, he repeated, "This is a rigged election"[18]. 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"[19]. For instance, they accused Democrats of discarding Republican ballots in a ditch in Wisconsin[20]. At the very same time, they cast Trump as the candidate "trying to protect democracy"[21]. This rhetoric probably contributed to the 6 January Capitol attack and to the widespread belief among Republicans that the 2020 election was stolen[22].

Would-be authoritarians' ability to weaponize the subversion dilemma may depend on a larger set of mutually reinforcing polarizations[23–27]. These include increasing partisan identity strength[28–30], polarized views on policy[2,31], dislike of opposing partisans[24], dehumanization of opposing partisans[32,33], stereotypes of opposing partisans[34] and ethnic antagonism[35].

When would-be authoritarians convince their supporters to tolerate backsliding, they potentially scare their opponents into also supporting backsliding. If those opposed to the aspiring autocrat begin to respond in kind, this may start a vicious cycle of antidemocratic action.

In the United States, we see potential signs of the subversion dilemma playing out in Democrats' rhetoric. For instance, in 2019, Representative Ocasio-Cortez tweeted, "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", and in 2020 Senator Elizabeth Warren tweeted, "Health care. Reproductive Freedom. Workers' rights. Dreamers' futures. Our planet. Democracy. Everything is on the line—so everything is on the table"[36,37]. Despite heightened rhetoric, we have not observed much democratic subversion from Democratic leaders. This forbearance is probably upholding democracy in the United States, yet it is unclear how long forbearance can protect democracy among increasing mutual alarm and Republican violations of democratic norms.

To summarize, we think that democracy-loving voters may vote away their political rights in part because would-be authoritarians convince their supporters that the other party is subverting democracy (1), leading them to begin tolerating backsliding by their leaders (2), which then prompts legitimate fears in the other party (3), leading the other party to tolerate drastic action by their own leaders (4). Critically, party leaders opposed to the aspiring autocrat may not take advantage of the increased tolerance among their supporters for antidemocratic action. But if they do, it will probably result in a death spiral for democracy. In this paper, we focus on testing one part of this chain of causal claims that is at the heart of the subversion dilemma. Across three studies, we examine whether partisans support subverting democracy to the extent that they believe the other party is willing to subvert it, the causal claim in points 2 and 4.

Our findings build off a burgeoning literature on how inaccurate perceptions of opposing partisans contribute to backsliding. Studies have found that partisans tend to overestimate opposing partisan policy extremism[38], animosity towards out-partisans[39–42], dehumanization of out-partisans[43–46] and willingness to engage in partisan violence[47]. In the study closest to this one, researchers found that the gap in perceptions between in-partisans' and out-partisans' support for democracy correlates with their willingness to harm the other party even at the expense of the country and its laws[48]. Our study builds on this work by theorizing that partisans will cooperate in a democracy to the extent that they believe opposing partisans will do the same—a tendency that aspiring autocrats may exploit by spreading misinformation that opposing partisans are undermining democracy. Informational interventions that reduce exaggerated beliefs about opposing partisans have successfully reduced partisan animosity[41], dehumanization[44–46] and support for partisan violence[47], though they have not typically increased commitment to democratic norms[42,49,50]. Our studies show that informational interventions can do so when they counter heightened perceptions that the other party supports subverting democracy.

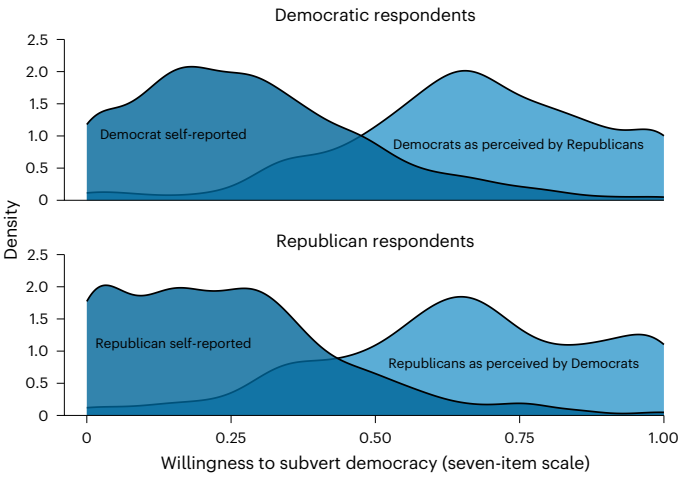

Fig. 1 | Exaggerated misperceptions of opposing partisans' commitment to democracy (study 1). Shown are the distributions of the unweighted average of seven questions that we asked the respondents about their perception of opposing partisans' willingness to subvert democracy and their own willingness to subvert democracy (kernel densities, $N = 1{,}050$ Democratic respondents and $N = 923$ Republican respondents). The questions ranged from reducing polling places near opposing partisans to banning opposing partisan rallies. Members of both parties appear to overestimate opposing partisans' willingness to break democratic norms. On average, Republicans gave Democrats a mean of 0.65 on the 0–1 subversion scale (95% CI, (0.64, 0.67)), but Democrats self-reported a willingness to subvert these norms with a mean of 0.28 on the 0–1 scale (95% CI, (0.27, 0.29)). Similarly, Democrats gave Republicans a mean of 0.67 on the 0–1 subversion scale (95% CI, (0.66, 0.68)), but Republicans self-reported a willingness to subvert these norms with a mean of 0.24 on the 0–1 scale (95% CI, (0.23, 0.29)).

Results

Study 1

Study 1 began by examining whether beliefs that the other side will subvert democracy correlate with partisans' own willingness to subvert democracy with a demographically representative sample of 1,973 US partisans (see Supplementary Tables 1 and 2 for demographics and Supplementary Table 3 for descriptive statistics). Studies suggest a wide variety of understandings of democratic norms among the public[10–14]. We therefore assessed support for democratic norms by asking the respondents about seven actions that would benefit their own party at the expense of a democratic norm. We conceptualized democratic norms as common understandings about when political actors must exercise restraint to uphold democratic institutions, and which have strong support from partisans on both sides in the United States[31,51]. As an example, we asked Democrats, 'Would YOU support reducing the number of voting stations in towns that support REPUBLICANS?' The other six scenarios asked 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.

To assess beliefs about the other party's willingness to subvert, we asked respondents how they think most opposing partisans would respond to 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 asked about 'most' opposing partisans rather than opposing partisan elites because, ultimately, the defence of democracy depends on citizens' willingness to constrain potential office holders who themselves are likely to have varying degrees of democratic commitment[16].

https://doi.org/10.1038/s41562-023-01594-w

Figure 1 presents the distributions of responses to these seven items by party, using a simple additive index rescaled 0–1. Higher values on the scale indicate more willingness to subvert democracy oneself or the belief that the other party is more willing to do so. On average, Republicans believed that Democrats were willing to subvert democratic norms in 5.0 of the 7 scenarios (mean, 0.65 on the 0–1 scale; 95% confidence interval (CI), (0.64, 0.67)), but Democrats self-reported willingness to subvert these norms in only 1.5 of the scenarios (mean, 0.28 on the 0–1 scale; 95% CI, (0.27, 0.29)). Similarly, Democrats believed that Republicans were willing to subvert democratic norms in 5.2 of the scenarios (mean, 0.67 on the 0–1 scale; 95% CI, (0.66, 0.68)), but Republicans self-reported willingness to subvert these norms in only 1.2 of the scenarios (mean, 0.24 on the 0–1 scale; 95% CI, (0.23,0.29)). Supplementary Figs. 1–3 show the distribution of self-reported willingness to subvert for each item by party and by strength of partisanship.

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. Nevertheless, it is worth noting that the perception gap we found is larger than the perception gaps documented between partisans on issues of ideological and affective polarization[39,44]. On a 0 to 1 scale, we found that partisans inaccurately perceived opposing partisans by an average of 0.09 on policy views and 0.14 on dehumanization of opposing partisans, while they inaccurately perceived the other side's willingness to subvert democracy by 0.40.

The core claim we examine in this paper is that the belief that the other side will subvert democracy leads partisans to support subverting it themselves. Using the two multi-item scales described above, Fig. 2 reveals 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 with respondents who do not believe that the other party desires to undermine democracy at all (0 on the scale), respondents who believe that the other side is fully willing to undermine democracy (1 on the scale) increase their own willingness to undermine democracy by about 0.25 points (on the 0–1 scale) in a linear regression for Democrats and Republicans (Democrats: $b = 0.29$; s.e. = 0.03; $t(1,029) = 10.2$; $P < 0.001$; 95% CI, (0.231, 0.340); Republicans: $b = 0.24$; s.e. = 0.03; $t(911) = 9.3$; $P < 0.001$; 95% CI, (0.192, 0.295)). Supplementary Figs. 4 and 5 present this relationship with violin plots and by question order.

To address alternative explanations for this relationship, we statistically adjust for factors that may influence both variables with linear regression (we also run experiments in studies 2 and 3). Since polarization along party, ideological and racial lines could influence both variables of interest, we control for partisan identity strength, extremism of policy views, extremism of racial attitudes, dehumanization of opposing partisans and the difference between feeling thermometers for opposing partisans and copartisans. We also control for two sets of beliefs about opposing partisans: perceptions of how extreme their policy views are and how much they dehumanize one's own party members[29,31–33,39,44]. These controls leave the association in Fig. 2 largely unchanged for Democrats and Republicans (Supplementary Table 4). In study 3, we also include a four-item ethnic antagonism battery drawn from Bartels[35], which also leaves the key relationship unchanged.

Study 2

Study 2a. In study 2a, we tested our core claim experimentally: we corrected exaggerated misperceptions of the other side's willingness to break democratic norms and examined the impact on partisans' own willingness to uphold these norms. We used an 'ask–tell' design, which has successfully corrected perceptions in other studies by providing respondents with feedback about how their perceptions align with reality[39,47,52].

In a demographically representative sample of 2,545 US partisans, we administered the ask-tell intervention to 50% of respondents randomly assigned to the treatment condition (Supplementary Tables 1,

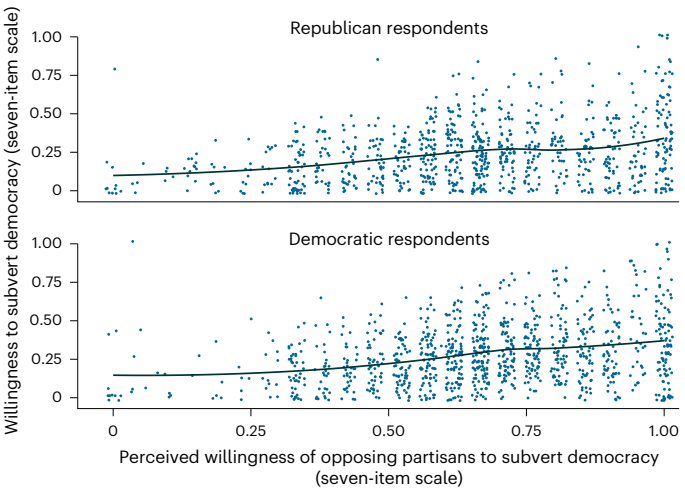

Fig. 2 | The relationship between respondents' willingness to subvert democracy and their perception that opposing partisans support subverting democracy (study 1). Each point shows one respondent ($N = 1,031$ Democratic respondents and $N = 913$ Republican respondents). The points are jittered. The line shows a loess smoother with a span of 0.75. 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 Fig. 1 caption for more details on the seven questions that we asked the respondents. A linear regression produced the following results: for Democrats, $b = 0.29$; s.e. = 0.03; $t(1,029) = 10.2$; $P < 0.001$; 95% CI, (0.231, 0.340); for Republicans, $b = 0.24$; s.e. = 0.03; $t(911) = 9.3$; $P < 0.001$; 95% CI, (0.192, 0.295).

2 and 5). The experiment used the same seven scenarios from study 1 to measure willingness to break democratic norms. In the treatment group, after the participants answered each question about opposing partisans, we told 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 (Supplementary Figs. 1 and 2).

We asked the control group to answer the same questions about opposing partisans, though no feedback was provided. We then asked all respondents about their own willingness to subvert democratic norms using the same seven-item questionnaire from study 1 (Fig. 3 presents this survey flow visually). As in study 1, we took the simple average of the seven items and rescaled it 0–1. This experiment is a preregistered replication of a pilot experiment.

The ask–tell treatment succeeded at lowering perceptions that opposing partisans are willing to break democratic norms. Figure 4 shows the distribution of these perceptions of opposing partisans for the treatment and control conditions by party. Like the findings in study 1, participants in the control condition placed opposing partisans' willingness to subvert democratic norms at 0.64 on the 0–1 scale. By contrast, participants in the treatment condition placed them at 0.40 ($b = -0.236$; s.e. = 0.008; $t(2,543) = -27.8$; $P < 0.001$; 95% CI, (−0.252, −0.219)). Since we administered the ask–tell treatment across seven scenarios, this difference probably underestimates the manipulation effect, as the respondents are not fully treated until after they receive feedback on the final question. Indeed, when we look at responses only for the final (randomized) question, we see a larger difference, with 0.67 in the control condition and 0.37 in the treatment condition ($b = -0.307$; s.e. = 0.013; $t(2,543) = -24.0$; $P < 0.001$; 95% CI, (−0.282, −0.333)). We also find that the manipulation effect increases with each item in the ask–tell treatment (Supplementary Fig. 6).

https://doi.org/10.1038/s41562-023-01594-w

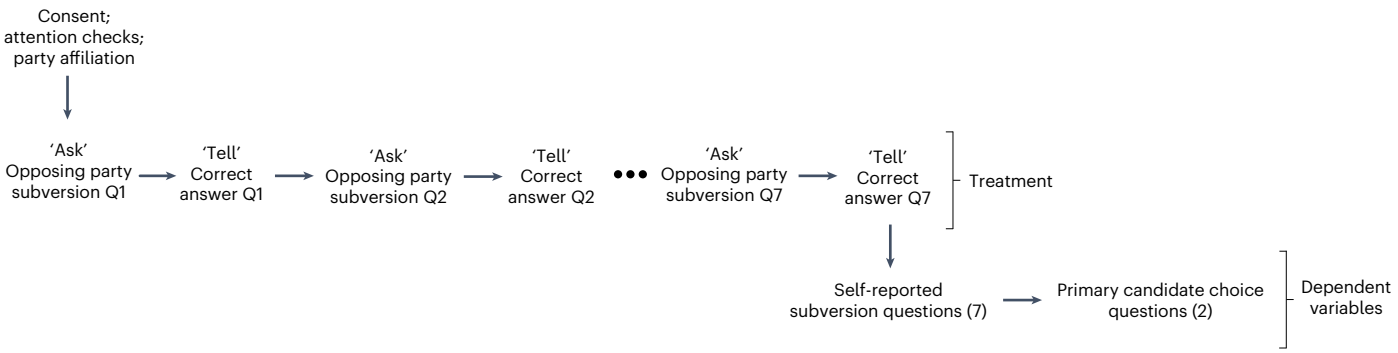

Fig. 3 | Experimental flow in the treatment condition (study 2). After answering each question (Q) about opposing partisans' support for subverting democracy, respondents in the treatment group received information about how opposing partisans actually answered the survey items (based on data from study 1). The control group saw the same flow except for the 'tell' component. We randomized the order of the items within each battery.

If partisans vote away the democracies they cherish because they think the other side is willing to do the same, this successful manipulation should increase support for upholding democratic norms. It does. The treatment group was 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 conditions. The average participant in the treatment group became less willing to subvert democratic norms, shifting from a mean of 0.24 to 0.17 on the 0–1 scale, a 29% relative change and one that is highly unlikely to occur by chance ($b = -0.076$; s.e. = 0.008; $t(2,543) = -9.8$; $P < 0.001$; 95% CI, (−0.091, −0.061)). Respondents in the control condition said they would 'never' support breaking 3.5 of the 7 democratic norms—a number that increased to 4.7 in the treatment group. Both Democrats and Republicans exhibited a response to the treatment, with Democrats falling from 0.26 in the control condition to 0.16 in the treatment condition ($b = -0.095$; s.e. = 0.011; $t(1,419) = -8.9$; $P < 0.001$; 95% CI, (−0.115, −0.074)) and Republicans falling from 0.22 to 0.17 ($b = -0.054$; s.e. = 0.012; $t(1,122) = -4.7$; $P < 0.001$; 95% CI, (−0.076, −0.031)).

We find a statistically significant treatment effect across all seven items in the self-subversion scale, across all but one of our six party-identification levels and across all levels of general political knowledge (Supplementary Figs. 7–9).

While the treatment effect suggests that the observational relationship that we document in study 1 is causal, a key question that follows is: how much of the observational relationship is causal? We shed light on this question by scaling the treatment effect by the degree of successful manipulation—that is, the complier average causal effect (CACE)—using the answer that the participants gave in the final question of the ask–tell battery (order-randomized). The CACE of the treatment through perceptions of opponent subversion is $b = 0.249$; s.e. = 0.024; $t(2,540) = 10.2$; $P < 0.001$; 95% CI, (0.201,0.297)—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 Methods for details and assumptions and Supplementary Tables 6 and 7 for the full results).

If the belief that the other side is willing 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 increased warm feelings towards out-partisans (Republicans: $b = 0.048$; s.e. = 0.021; $t(1,098) = 2.3$; $P = 0.023$; 95% CI, (0.007, 0.089); Democrats: $b = 0.074$; s.e. = 0.018; $t(1,382) = 4.0$; $P < 0.001$; 95% CI, (0.038, 0.11)) (Supplementary Fig. 10). The treatment also decreased the perception that the other party dehumanizes them (Republicans: $b = -0.047$; s.e. = 0.016; $t(1,099) = -2.9$; $P = 0.0039$; 95% CI, (−0.079, −0.015); Democrats: $b = -0.083$; s.e. = 0.014; $t(1,384) = -5.7$; $P < 0.001$; 95% CI, (−0.111, −0.054)) (Supplementary Fig. 11). We do not find that the intervention changed the respondents' own policy views (five-item scale), but it may have reduced Democrats' perception that Republicans are extremists on several policy issues (Supplementary Fig. 12).

Recent experiments have shown that lowering affective polarization fails to increase support for democratic norms, despite these correlating in survey data[50]. The results here suggest that the causal arrow may point in the opposite direction: changes in the perception of opposing partisans' support for democratic norms may influence interpartisan animosity.

We find no sign that respondents guessed the study's purpose on the basis of an open-ended question at the end of the survey, nor do we see evidence consistent with a demand or social desirability pressure account (study 3 and Supplementary Methods study 2). We suspect that the treatment—learning about the level of democratic commitment of opposing partisans—works through multiple mechanisms, including through changing perceptions of descriptive norms[53].

Study 2b. Does correcting perceptions about opposing partisans' willingness to subvert democratic norms translate into behaviours that could limit democratic backsliding? To examine the consequences for voting decisions, we asked the respondents to vote in two hypothetical primary elections at the end of study 2 (study 2b). We then examined the impact of the ask–tell intervention on these voting decisions. Unlike the previous analyses, this analysis is exploratory (not preregistered), though we replicated it in study 3. In each hypothetical primary race, the respondents faced a choice between two candidates from their own party, one who has supported breaking one of the seven democratic norms in our study (chosen at random) and another who has opposed breaking the norm.

To model real-world rhetoric, we told the participants that the candidate who supported breaking a democratic norm did so 'because they believe opposing partisans (Democrats or Republicans) have done the same.' This rhetoric also helped us overcome a social desirability problem, which is that respondents seem unlikely to support a subverting candidate on a survey unless the choice seems justifiable.

In study 2b, we find that the same ask–tell treatment used in study 2a decreases partisan willingness to vote for candidates who support breaking democratic norms by 0.035 on the 0–1 scale, where we code a democracy-protecting-candidate vote 0, a neither vote 0.5 and a democracy-subverting-candidate vote 1. The treatment is statistically significant: $b = -0.035$; s.e. = 0.012; $t(5,050) = -2.97$; $P = 0.003$; 95% CI, (−0.059, −0.012), calculated with robust standard errors clustered at the respondent level, as Fig. 6 shows. This effect represents a shift from 0.259 in the control condition to 0.224 in the treatment condition, a 14% relative change. Republicans and Democrats

https://doi.org/10.1038/s41562-023-01594-w

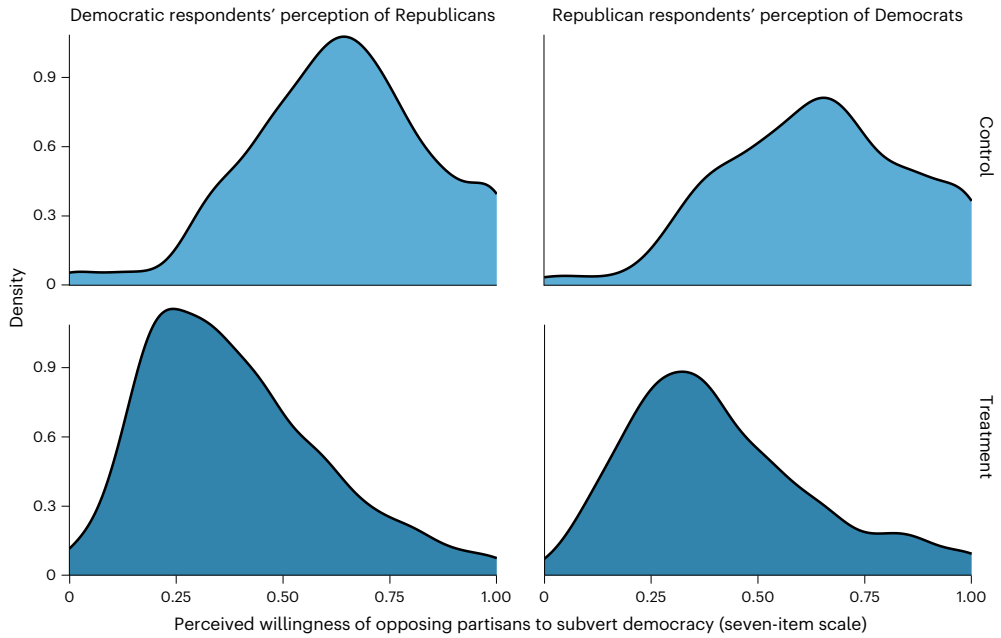

Fig. 4 | Reducing perceptions of opposing partisan willingness to subvert democracy (study 2). The experimental manipulation check for study 2 is shown, using the distributions of the unweighted average of the seven questions that we asked the respondents about their perception of opposing partisans' willingness to subvert democracy (kernel densities, $N = 2{,}545$ respondents). The informational intervention—which targeted beliefs about opposing partisans' commitment to democracy—lead participants to see opposing partisans as less willing to subvert democracy. Since we gave the respondents feedback after they answered each of the seven questions, this figure probably understates the extent of the manipulation effect, because the respondents did not receive the full treatment until they received feedback on their answer to the seventh item. The overall effect for both parties is $b = -0.236$; s.e. = 0.008; $t(2{,}543) = -27.8$; $P < 0.001$; 95% CI, (−0.252, −0.219).

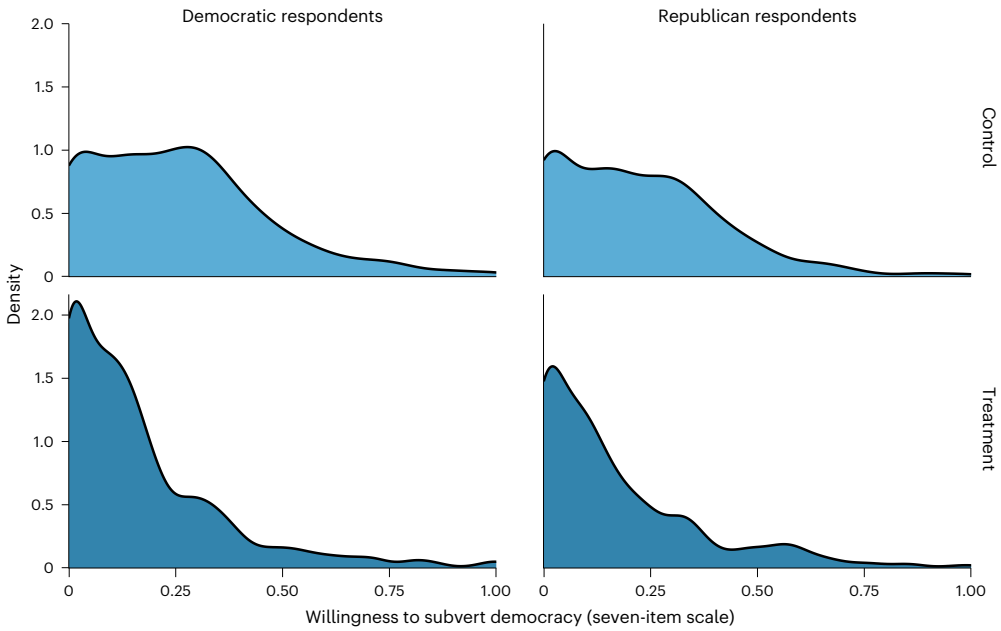

Fig. 5 | Lowering perceptions that other party will subvert democracy generates more support for democratic norms (study 2a). The main experimental effect for study 2a is shown using the distributions of the unweighted average of the seven questions that we asked the respondents about their own willingness to subvert democracy by treatment condition and party identification (kernel densities, $N = 2{,}545$ respondents). They reveal that the informational intervention—which targeted beliefs about opposing partisans' commitment to democracy—decreased the participants' own willingness to subvert democracy. This result suggests that the observational relationship documented in study 1 may be causal. For Democrats, $b = -0.095$; s.e. = 0.011; $t(1{,}419) = -8.9$; $P < 0.001$; 95% CI, (−0.115, −0.074); and for Republicans, $b = -0.054$; s.e. = 0.012; $t(1{,}122) = -4.7$; $P < 0.001$; 95% CI, (−0.076, −0.031).

exhibited almost identical responses to the treatment, with Democrats falling from 0.27 in the control condition to 0.23 in the treatment condition ($b = -0.035$; s.e. = 0.016; $t(2{,}815) = -2.2$; $P = 0.03$; 95% CI, (−0.067, −0.003)) and Republicans falling from 0.25 to 0.21 ($b = -0.036$; s.e. = 0.018; $t(2{,}233) = -2.1$; $P = 0.04$; 95% CI, (−0.071, −0.002)). The CACE of the treatment through perceptions of opponent subversion

https://doi.org/10.1038/s41562-023-01594-w

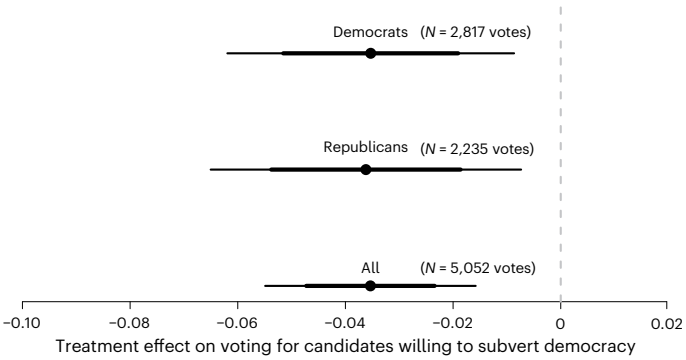

Fig. 6 | Lowering perceptions that the other party will subvert lowers support for candidates who subvert democracy (study 2b). The main experimental effect for study 2b is shown using the treatment-group mean minus the control-group mean coded 0–1, where higher numbers represent votes for more democracy-subverting candidates (N = 5,052 votes). The control-group mean is around 0.24 for both parties. Each point shows the treatment effect estimate from a separate regression. The error bars show 95% (thin) and 68% (thick) CIs and are calculated from standard errors clustered by respondent. The informational intervention—which targeted beliefs about opposing partisans' commitment to democracy—decreased partisan willingness to vote for a primary candidate from their own party who supports subverting democracy. For Democrats, b = −0.035; s.e. = 0.016; t(2,815) = −2.2; P = 0.03; 95% CI, (−0.067, −0.003); and for Republicans, b = −0.036; s.e. = 0.018; t(2,233) = −2.1; P = 0.04; 95% CI, (−0.071, −0.002).

is b = 0.116; s.e. = 0.039; t(5,050) = 3.0; P = 0.003; 95% CI, (0.04, 0.191). This self-reported voting behaviour is correlated with support for democratic norms, and mediation analysis suggests that respondents' own willingness to subvert democracy fully mediates the treatment effect (Supplementary Fig. 13 and Supplementary Tables 8 and 9)[54,55]. Of course, mediation analysis should be taken as suggestive, as it depends on numerous assumptions.

This hypothetical candidate choice study provides suggestive evidence that reducing perceptions that opposing partisans will violate democratic norms may lead voters to prefer candidates who uphold democratic norms. Although not large, the effect may be biased downwards by social desirability and floor effects. It also may not need to be large to matter in the real world, where elections can be closely contested.

Study 3

To further address concerns that the treatment effect arises from respondents guessing the intent of the experiment or from respondents desiring to appear less subversive after exposure to the treatment (treatment-induced social desirability), we conducted a panel version of the study designed to minimize these concerns (study 3). We administered the treatment in an initial survey and collected the dependent variable in a second survey. We designed the surveys to appear as 'seemingly unrelated' as possible. We did so by not highlighting 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 administered this study on Mechanical Turk, where respondents typically take many surveys every day.

In a preregistered study, we recruited 2,523 respondents for the first survey, which took place on 9 and 10 May 2022 (see Supplementary Tables 1 and 2 for the demographics and Supplementary Tables 10 and 11 for descriptive statistics). As in study 2, a random half of the sample received feedback on the accuracy of their perceptions after each question about the extent to which most opposing partisans would support subverting democratic norms (the ask–tell treatment), and the other half did not. We then successfully resurveyed 1,848 (73.2%) between 9 and 14 May 2022. 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 asked the seven self-subversion items (the main dependent variable) and then the seven items about the other party (a manipulation check).

Even with these steps to reduce demand and social desirability, the treatment effect remains similar to that in study 2. The intervention in wave 1 shifted the mean willingness to subvert democratic norms at wave 2 from 0.32 to 0.24 on the 0–1 scale, a 25% relative change and one that is highly unlikely to occur by chance (b = −0.075; s.e. = 0.011; t(1,843) = −6.8; P < 0.001; 95% CI, (−0.097, −0.054)). When we scale the treatment effect by the degree of successful manipulation (CACE) using small subset randomly assigned to first answer the questions about the other party's subversion, the estimate closely matches the study 1 regression estimate and the study 2 CACE estimate (b = 0.227; s.e. = 0.041; t(1,663) = 5.5; P < 0.001; 95% CI, (0.146, 0.304)) (see Methods for details and assumptions and Supplementary Table 12 for the full results).

Figure 7 shows the treatment effect over time, revealing a consistent difference between the control group and the treatment group. The treatment and control lines trend upwards over time, but interestingly, respondents randomly assigned to take the survey later failed to exhibit these upward trends, implying that the time since the previous survey does not cause changes in subversion levels. Instead, this pattern appears to result from non-random 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. It also holds up across all six levels of party identification (Supplementary Figs. 14 and 15). For example, the estimate for 'strong Republicans' is b = −0.097; s.e. = 0.033; t(295) = −2.9; P = 0.0038; 95% CI, (−0.162, −0.032); and for 'strong Democrats' it is b = −0.082; s.e. = 0.02; t(592) = −4.0; P < 0.001; 95% CI, (−0.122, −0.041).

In an exploratory analysis (not preregistered), we replicate the results of the primary choice experiment, which followed study 2 exactly, except that we measured the candidate choice in wave 2 and asked the respondents about three hypothetical primary races. Treated respondents again voted less often for the subverting candidate using the same 0–1 scale (b = −0.058; s.e. = 0.015; t(5,298) = −3.87; P < 0.001; 95% CI, (−0.088, −0.029), calculated with robust standard errors clustered at the respondent level), as Fig. 8 shows. The treatment effect's variation with hours since treatment and correlation with support for democratic norms can be found in Supplementary Figs. 16 and 17. The CACE estimate of the treatment through perceptions of opponent subversion is b = 0.165; s.e. = 0.045; t(4,808) = 3.6; P < 0.001; 95% CI, (0.075, 0.254) on a 0–1 scale with a low control-group mean.

In study 3, we also control for ethnic antagonism (Supplementary Table 13). Although we find that Republican respondents show higher levels of ethnic antagonism overall, interestingly, we find that both Democrat and Republican respondents with higher levels of ethnic antagonism are more likely to support subverting democratic norms. However, controlling for ethnic antagonism leaves the key findings in this paper unchanged.

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

Discussion

Across three studies, we examine whether voters who largely prefer democracy will nonetheless be willing to undermine it to the extent that they believe the other side is doing the same. In an observational study, we find that partisans are willing to break democratic norms to the extent that they believe opposing partisans support breaking them, an association that holds even while controlling for other usual suspects of democratic backsliding, including partisan identity strength, extreme policy preferences, perceptions that the other side

https://doi.org/10.1038/s41562-023-01594-w

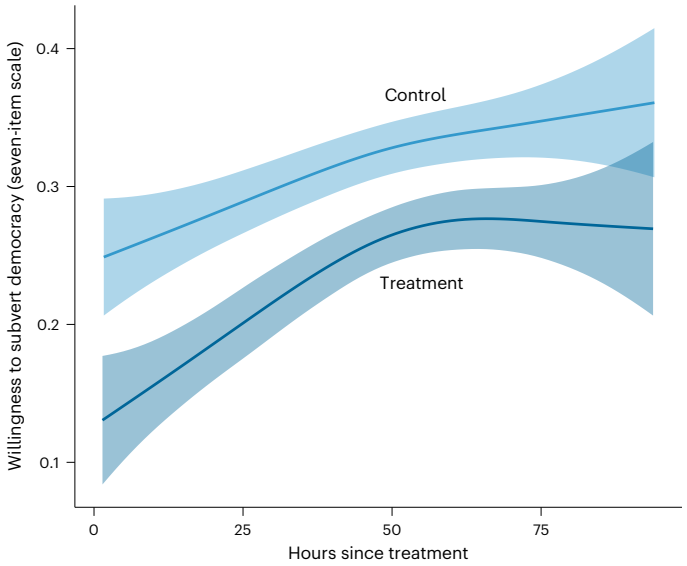

Fig. 7 | Wave 1 ask–tell treatment effect on wave 2 support for subverting democracy by treatment condition and by hours since treatment (study 3). A loess line with a span of 0.6 and with the 95% CI (N = 1,845 respondents) is shown. It reveals that, despite efforts to reduce demand and social desirability with the seemingly-unrelated-studies approach, the treatment effect remains similar to that in 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 do not see the treatment effect diminishing with time. While the treatment and control lines trend upwards over time, this reflects differences between participants based on when they chose to take the second survey. When we analyse the data by when participants were randomly assigned to receive a request to take the second survey, we do not observe these upward trends, implying that the time since the previous survey does not cause changes in subversion levels. The overall treatment effect is b = −0.075; s.e. = 0.011; t(1,843) = −6.8; P < 0.001; 95% CI, (−0.097, −0.054).

holds extreme policy preferences, partisan animosity, dehumanization of the other side, perceptions that the other side dehumanizes oneself, political knowledge and a range of demographic variables. We assess whether this relationship is causal with two experimental studies and find that reducing exaggerated beliefs that the other side is willing to undermine democracy increases support for upholding democratic norms. We also find that reducing these exaggerated perceptions may translate into behavioural outcomes, such as voting for democracy-protecting candidates.

The experimental intervention in this paper recently competed against 25 others (selected from 252) in the Strengthening Democracy Challenge megastudy (N = 32,059)[50]. It ranked first in lowering antidemocratic attitudes and first in lowering an overall composite index of all eight outcomes in the megastudy, including partisan animosity and support for partisan violence. In addition to providing an independent replication, the strength of these results is consistent with the subversion dilemma contributing to backsliding in the United States.

These studies test the core causal claim at the heart of the subversion dilemma—partisans will support subverting democracy to the extent that they believe their opponents support subverting democracy. But these studies have not tested our speculation that aspiring autocrats' rhetoric fosters misperceptions about the other side's commitment to democracy or the claim that actions taken by an aspiring autocrat's party may generate reciprocal democratic defections from their opponents. We leave this to future research, as the ethical complications of experimentally testing these elements of our theory are more formidable. Another important limitation of this paper is that our experiments took place in an online survey setting. Whether this intervention could survive in a real-world setting where an aspiring autocrat may be generating misperceptions between partisans remains an open question for future research.

If the dynamic we describe is indeed present in the United States, our findings suggest that Trump's relentless claims about Democrats stealing elections fosters support among Republicans for Trump's overstepping of democratic norms in the name of 'saving democracy', including the 6 January Capitol attack. The ongoing rhetoric from Trump-aligned Republican politicians questioning election integrity in the United States could portend future violations of democracy with wide public support.

The rhetoric from Democrats and third-party observers has understandably focused on the risk posed to democracy by illiberal components of the Republican Party, especially with Trump formally running in the 2024 election. Our work suggests an additional and counterintuitive strategy: focus on convincing everyday Republicans of Democrats' unwavering commitment to democracy. Doing so would probably require a concerted messaging campaign and credible demonstrations of this commitment, such as third-party guarantees or costly signals of good faith. Just as important, supporters of democracy should avoid actions and rhetoric that Trump and others could portray as confirmation of backsliding by Democrats.

In our survey data, we see ample signs that Republicans want to protect democracy and that they are open to information that Democrats want to do the same. Republicans may therefore be open to this messaging and these signals, much more so than many Democrats assume. In study 1, for instance, we asked the respondents whether they would be interested in making a one-to-one pact with a member of the opposing party to never vote for a candidate that subverts democracy. 60% of Democrats said 'probably' or 'definitely yes', while 72% of Republicans did so.

Similarly, Republican leaders concerned about democracy should take a stand against this particularly pernicious type of misinformation being spread by their copartisans and reassure the public about Democrats' commitment to democracy. As an example, gubernatorial candidates Chris Peterson (Democrat) and Spencer Cox (Republican) made a joint ad calling for civility and peaceful transfer for power in their election that went viral across the nation[56].

There may also be institutional factors that could lessen the impact of the subversion dilemma and its contribution to backsliding, such as institutions that help foster a shared commitment to democracy even in the absence of immediate authoritarian threat[57–59]. Meaningful third-party guarantees on democratically normative behaviour in the United States and beyond could come from a re-invigorated international democracy-promoting regime[60,61].

Future research may assess the extent to which the dynamic documented in this paper contributes to backsliding in countries outside the United States. Consistent with the theory, aspiring autocrats during the current "third wave of autocratization"[62] have accused their opposition of subverting democracy in countries throughout the world[3,6,63,64]. For example, Bolsonaro spread doubt about Brazil's electoral integrity leading up to the 2022 election[65]. There is also evidence that the strategies we suggest to counter backsliding may have succeeded in other countries. For instance, democratic parties facing authoritarian challenges in Colombia may have stemmed backsliding by demonstrating a strong commitment to democratic norms during their aspiring autocrat's term in office[6,63]. Rather than calling for impeachment, and so pushing Columbia further into the subversion dilemma, they focused on countering specific democratic violations[6,63,64].

This paper provides a framework for understanding why citizens cede their political rights to aspiring autocrats. It highlights a toxic misperception—the belief that opposing partisans are willing to dismantle democracy—and suggests that countering the spread of this

https://doi.org/10.1038/s41562-023-01594-w

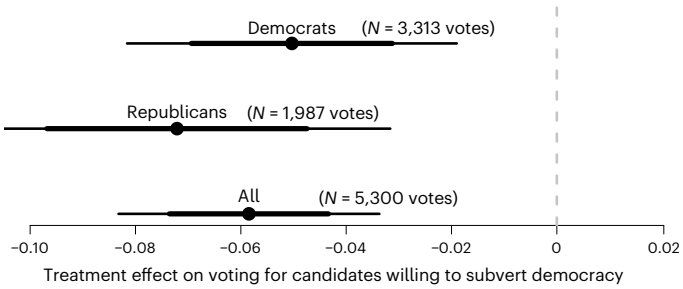

Fig. 8 | Lowering perceptions that the other party will subvert in wave 1 lowers support for candidates who subvert democracy in wave 2 (study 3). This study replicates study 2b, except that we measured the candidate choice in wave 2 and asked the respondents to cast votes in three primary choice elections. The treatment-group mean minus the control-group mean (both coded 0–1) is shown, where higher numbers represent votes for more democracy-subverting candidates (N = 5,300 votes). The control-group mean is around 0.34 for both parties. Each point shows the treatment effect estimate from a separate regression. The error bars show 95% (thin) and 68% (thick) CIs and are calculated from standard errors clustered by respondent. Despite efforts to reduce demand and social desirability with our seemingly-unrelated-studies approach, the treatment effect remains similar to that in study 2. The overall treatment effect is $b = -0.058$; s.e. = 0.015; $t(5,298) = -3.87$; $P < 0.001$; 95% CI, (−0.088, −0.029).

misperception is an important strategy to prevent democracy-loving citizens from falling victim to the slippery slope of the subversion dilemma.

Methods

All studies were approved by the University of California, Berkeley, Committee for the Protection of Human Subjects (protocol no. 2021-01-13949). We complied with all relevant ethical regulations, and all participants provided informed consent. We preregistered study 1 and 2 on 14 January 2021 at https://osf.io/vnd4g and study 3 on 9 May 2022 at https://aspredicted.org/7PJ_6TQ. We administered the surveys with Qualtrics. We recruited study 1 and 2 participants via Lucid Theorem, where participant payment is proprietary. We recruited study 3 participants via Mechanical Turk. The respondents received US$1 for participation. We selected sample sizes many times what would be necessary for 80% power based on pilot tests to examine the consistency of findings across subgroups. Data analysis was not performed blind to the conditions of the experiments. The data were analysed with R version 4.1.2. All studies excluded respondents who failed an attention check[66] or who did not identify with or lean towards the Democratic or Republican party.

Self-reported subversion and opposing partisan subversion questions

In all studies, we measured willingness to subvert democratic norms with seven questions. To measure the respondents' own willingness to subvert democratic norms, each question began, 'Would YOU support…' and was followed by seven items: 'banning FAR-[OTHER WING] rallies in the state capital?', 'ignoring controversial rulings by [OTHER PARTY] JUDGES?', 'freezing the social media accounts of [OTHER PARTY] JOURNALISTS?', 'reducing the number of voting stations in towns that support [OTHER PARTY]s?', 'laws that would make it easier for [OWN PARTY]s (and harder for [OTHER PARTY]s) to get elected?', 'using violence to block major [OTHER PARTY] laws?' and 'significantly reinterpreting the Constitution in order to block [OTHER PARTY] policies?' We randomized the order of the questions for each respondent. The response options were 'Never', 'Probably Not', 'Probably' and 'Definitely'. We measured beliefs about opposing partisan willingness to subvert democratic norms in the same way, except that each question began, 'Would MOST [OTHER PARTY]s support…' and was followed by the same seven items. In these questions, 'OWN WING' or 'OTHER WING' can be 'LEFT' or 'RIGHT' and 'OWN PARTY' or 'OTHER PARTY' can be 'DEMOCRAT' or 'REPUBLICAN,' depending on the respondent. To construct scales of self-reported subversion and opposing partisan subversion, we take the unweighted average of the seven questions and rescale it to a 0–1 range, where 0 is no subversion and 1 is the highest level of subversion (Cronbach alphas in study 1 of 0.89 and 0.82, respectively).

Study 1

We conducted study 1 via Lucid between 15 July and 6 August 2021 with 1,973 participants. From Lucid, we obtained demographic information on the participants including age, gender, household income, ethnicity, education, region, zip code and state (see Supplementary Tables 1 and 2 for the demographics). The demographic distribution closely matches the demographic distribution of US adult residents. After attention checks and a partisan identification question, the respondents were randomly assigned to first answer either the seven self-subversion items or the seven items about how willing the other party is to subvert democracy. At the end of the survey, the respondents were asked about numerous covariates of interest, such as policy preferences and partisan animosity.

We find some sign of order effects, in which those who first answered questions about the other party show lower rates of self-reported subversion (Supplementary Fig. 5).

Study 2

We conducted study 2 via Lucid between 15 and 29 September 2021 with 2,545 participants. From Lucid, we obtained demographic information on the participants including age, gender, household income, ethnicity, education, region, zip code and state (see Supplementary Tables 1 and 2 for the demographics). This was a preregistered replication of a preregistered pilot study (see Supplementary Figs. 18 and 19 for the key pilot findings). After informed consent, attention checks and partisan identification questions, we randomly assigned half of the respondents to the treatment 'tell' condition using Qualtrics survey software. We fail to find significant differential attrition or imbalances on pretreatment covariates (Supplementary Table 14).

The flow of the ask–tell treatment is depicted in Fig. 3. In the treatment condition, the respondents were told, 'Studies have shown that Democrats and Republicans in America DON'T KNOW MUCH about each other', and that they would receive feedback based on real data about 'average, everyday members' of the opposing party. The respondents in the treatment condition received feedback immediately after they answered each question about the extent to which they believe opposing partisans support a given democratic subversion.

When respondents in the treatment condition overestimated the level of support for subverting democratic norms among opposing partisans, they were shown the text: 'Sorry, MOST [OTHER PARTY]s do NOT support this action. Try again next time!' Below this text, there was a cartoon character looking slightly concerned next to the question they had just answered with the accurate answer circled. When respondents in the treatment condition correctly estimated the level of support for subverting democratic norms among opposing partisans for a given question, they were shown the text: 'Great Job! Most [OTHER PARTY]s WOULD NOT support this action.' Below this text, there was a cartoon character happily holding up a trophy next to the question they had just answered with their accurate answer starred. The least common scenario was when respondents underestimated opposing partisan willingness to subvert democratic norms. This occurred when respondents selected that opposing partisans would 'never' undermine a specific norm, when most opposing partisans actually said they would 'probably not' undermine a specific norm. In this case, respondents were shown the text: 'Close—Half Points! MOST [OTHER PARTY]s WOULD NOT support this action.' Below this text, there was

a cartoon character looking encouraging with a thumbs-up next to the question they had just answered with the correct answer starred.

Participants in the control condition answered the seven opposing-partisan-subversion questions without feedback. Afterwards, we asked the treatment and control groups the seven self-subversion items (the main dependent variable) as well as the two hypothetical primary election questions. Finally, we asked the participants about covariates such as policy preferences and partisan animosity.

To further engage the respondents in both the treatment and control conditions, we told them that they would be placed in a league on the basis of the accuracy of their responses. After collecting all data, we showed the respondents whether they were in the 'silver', 'gold', 'diamond' or 'professional' league on the basis of their responses and thanked them for their participation.

Study 3

We conducted study 3 on Mechanical Turk between 9 and 14 May 2022 with 1,973 participants. This was a preregistered panel study of our intervention in study 2 meant to address concerns about demand effects. The first survey included consent, attention checks, demographics, party affiliation, political knowledge and our seven questions about the other party's willingness to subvert democratic norms (see Supplementary Tables 1 and 2 for the demographics). A random half of the participants were assigned to the treatment group using Qualtrics survey software, in which case they received feedback on how they answered each of the seven questions about opposing partisans. The survey closely followed the first half of study 2, except that we excluded the introductory language about respondents not knowing much about each other and we excluded the leagues. We only told respondents in the treatment group that they would get feedback from 'real data'. We find no sign of differential attrition or meaningful imbalances (Supplementary Table 15).

We randomly assigned when the second survey became available to the respondents, 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 (reinterview) began with the seven self-subversion items (the main dependent variable) and also included the seven subversion items about the other party (a manipulation check), three hypothetical primary choice questions and four ethnic-antagonism items[35]. To check for order effects, about 10% of the sample received the battery of questions about opposing partisans first. To reduce demand effects, we did not refer to the second survey as a 'follow-up survey', and we attempted to make each survey visually dissimilar.

Sample representativeness and weighting

Supplementary Table 1 presents the demographic distribution of the samples and compares them to estimates from the Census American Community Survey. Supplementary Table 2 does the same by party by comparing Democrats and Republicans to the American National Election Study (ANES). The Lucid samples appear generally demographically representative. The one notable exception is that the tendency to overrepresent lower-income individuals is pronounced for Republicans, as they are typically higher-income. We think, however, that this overrepresentation of low-income Republicans works against the key effect in this Article. Supplementary Fig. 20 shows the study 2 treatment effect by demographic categories, finding a larger and precisely estimated treatment effect among higher-income Republicans. We therefore take a conservative approach by not weighting the samples, one consistent with the evidence from studies on weighting in survey experiments[67]. Compared with the ANES, the Lucid samples overrepresent strong Democratic partisans, which may also work against a treatment effect. In study 2, for instance, 27.3% are strong Democrats, compared with 23% in the ANES, while 20.9% are strong Republicans, compared with 21% in the ANES. The Mechanical Turk sample underrepresents Republicans. As with the ANES, in all three samples, strong partisans are more common than weak partisans, who are in turn more common than independents who lean towards one party.

Multi-item scales

In constructing the multi-item scales, we take the average of the non-missing values for each respondent. Since we requested responses when respondents attempted 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 Supplementary Tables 3, 5 and 10 for descriptive statistics).

Missing data imputation

So that model estimates do not omit respondents because of missing values on control variables, we impute missing values using demographics (using robust linear regression with M-estimators from the simputation package[68] in R[69]). We never impute values on the key independent or dependent variables (perceptions of the other party's subversion, self-reported subversion or primary vote choice). The number of observations imputed on each control variable is small (typically 20–30), and the model estimates are substantively identical when we do not impute.

Significance tests

We use two-tailed tests for all significance tests. For the significance tests of the main experimental effects in studies 2a and 3, we use HC3 standard errors. The significance tests assume normality, which does not hold with our key independent variable: respondent willingness to subvert. When we use non-parametric bootstrapping with 1,000 resamples for the key experimental tests in studies 2a and 3 on willingness to subvert, the results remain highly statistically significant. For study 2a, no resampled estimate was greater than or equal to zero, which corresponds to a $P < 0.001$ and a 95% CI of (−0.089, −0.063). For study 3, no resampled estimate was greater than or equal to zero, which corresponds to a $P < 0.001$ and a 95% CI of (−0.095, −0.058).

CACE

To scale the experimental estimates by compliance with the treatment so that we can compare them with the observational regressions, we estimate the CACE. A key assumption for this analysis is that the treatment can influence respondents only through its effect on the perceived willingness of out-partisans to subvert, not through other observed or unobserved variables (the exclusion restriction). In the two-stage least-squares (2SLS) estimates for study 2, we instrument the final (seventh) item that the respondents answered about opponents' willingness to subvert with the treatment indicator. We presented these items to the respondents in random order, and we use their answer to the final one because respondents in the treatment group had mostly received the treatment by that time (they had 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 doing so underestimates compliance even more. The 2SLS estimates reveal a CACE estimate that is close to the regression estimate from study 1 (both around 0.25), and one that is precisely estimated. Supplementary Tables 6 and 7 present the reduced-form regression and the 2SLS estimates for study 2.

Since we measured many covariates post-treatment, we can include those in separate 2SLS estimates to attempt to block those causal paths. Although including post-treatment variables can introduce other biases, the stability of the key estimate is reassuring. When we include measures of dehumanization, meta-dehumanization, the feeling thermometer difference between parties, policy extremism and beliefs about the other side's policy extremism (not shown), we find only the slightest decrease in the 2SLS estimate, and the estimate

remains precisely estimated. Of course, these analyses do not rule out exclusion restriction violations on unobservables and should be seen as only suggestive.

In study 3, we can adopt an arguably better approach. For a small subset of respondents in the reinterviews, we asked the battery about perceived willingness of the other party to subvert democracy before any other questions. We estimate the first stage in this sample and the second stage among the other respondents who were first asked the self-subversion items (in the second-wave interview). This approach is called two-sample 2SLS. To calculate the standard errors, we use Stata's weaktsiv package[70], and we control for age and folded-seven-point party identification, both measured pretreatment, neither of which substantially affect the CACE estimates. When we adopt a single-sample approach, we estimate a much larger CACE because asking the self-subversion battery first notably suppresses the treatment effect on the opponents' subversion battery, lowering the estimate of compliance and so increasing the CACE. The sensitivity of the estimates to question wording order is yet another reason to be sceptical of CACE estimators.

Demand and social desirability

There are additional reasons to believe that experimentally-induced socially desirable responding is not driving the results in study 2. First, although the treatment influenced perceptions of out-partisans, the respondents still saw out-partisans as more willing to subvert democracy than they reported being themselves. This is true even when we look only at the last (randomized) item we asked them about (when they would have received six of the seven ask–tell treatments). On this last item, mean perception of out-partisans' willingness to subvert democratic norms is still about twice as high for the other party as for themselves (0.39 versus 0.17 on the 0–1 scale). Respondents could, in theory, admit to more willingness to subvert democratic norms while keeping 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 yields a similar estimate.

Interpretation of experimental findings

One interpretation of the experimental findings is that respondents learned about a 'descriptive norm' from the treatment and adjusted their answers to the self-subversion questions to be more socially acceptable. Although this interpretation may be plausible, several findings bolster our interpretation. In study 1, respondents who most perceived that the other party wanted to subvert democracy also expressed the most fear of the other party (Supplementary Fig. 21), which is exactly what our theory predicts. We also found that the study 2 intervention increased partisans' positive feelings about each other, decreased the perception that the other party dislikes them, reduced Democrats' perception that Republicans are extremists on a number of policy issues, decreased blatant dehumanization of the other party and decreased the perception of being blatantly dehumanized by the other party (Supplementary Figs. 10–12). Likewise, in Stanford's Strengthening Democracy Challenge megastudy, this same intervention reduced partisan animosity as well as support for interpartisan violence[50]. If the treatment were only providing information about how to answer the seven subversion questions in a publicly acceptable way, it should not affect these other outcomes. Of course, none of this rules out the possibility that some respondents are merely learning what to say on a survey.

Preregistration deviations

We report several slight deviations from our preregistration. For the CACE estimates, the pre-analysis plan stated that we would subtract perceptions of opposing partisans' willingness to subvert on the final subversion item (7/7) from the answer they gave on the first question (1/7). We subsequently preferred to estimate CACE with only the final question. When we conduct the preregistered analysis, however, the results are the same (see Supplementary Table 16 and compare to Supplementary Table 7).

The only other deviations involve minor changes in question wording, described in the following paragraphs. Our preregistration listed eight subversion items, but we ended up dropping one item before running studies 1 and 2. This question asked 'Would YOU support a [OWN PARTY] governor ruling by executive order if [OTHER PARTY]s don't cooperate?' The responses to this question had such high levels of support that we determined that this question was not perceived as normatively prohibited in the US democratic context. As a result, the preregistration notes that the index for subversion and opposing partisan subversion would include eight questions coded 0–3 for a total index range of 0–24. Instead, we have seven questions coded 0–3, which are added together and normalized to 0–1 for ease of interpretation.

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?' With these changes, all questions are asking about the participants' support for certain actions rather than their support mediated through a hypothetical representative.

In studies 1 and 2, we received demographic variables except for party identification from Lucid, not from the respondents. The preregistration does not list the covariates used to examine alternate hypotheses in our observational survey and to test moderators in our experiments. These include policy polarization, perceptions of opposing partisan policy polarization, blatant dehumanization, perceptions of opposing partisan dehumanization, level of political knowledge and level of warmth towards Democrats and Republicans in studies 1 and 2.

Due to uncertainties about our budget, we conservatively estimated that we would survey 500 people in study 1 and 1,000 in study 2, but we surveyed 2,000 (1,973) in study 1 and about 2,500 (2,545) in study 2.

In the preregistration for study 3, we forgot to note that, as in all our studies, we ended the survey if respondents failed the initial attention checks (two attention checks immediately following consent), a procedure adopted because of the high level of inattention among respondents on Lucid and Mechanical Turk.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Replication data for the main text and Supplementary Information are available at https://doi.org/10.17605/OSF.IO/4BDET (ref. 71). Source data are provided with this paper.

Code availability

Replication code for the main text and Supplementary Information is available at https://doi.org/10.17605/OSF.IO/4BDET.

References

1. Wike, R., Simmons, K., Stokes, B. & Fetterolf, J. *Globally, Broad Support for Representative and Direct Democracy* (Pew Research Center, 2017).

2. Lindberg, S. I. & Boese, V. A. *Autocratization Changing Nature?* (Democracy Report, 2022).
3. Levitsky, S. & Ziblatt, D. *How Democracies Die* (Broadway Books, 2018).
4. Svolik, M. When polarization trumps civic virtue: partisan conflict and the subversion of democracy by incumbents. *Q. J. Polit. Sci.* **15**, 3–31 (2020).
5. Bermeo, N. On democratic backsliding. *J. Democr.* **27**, 5–19 (2016).
6. Gamboa, L. Opposition at the margins: strategies against the erosion of democracy in Colombia and Venezuela. *Comp. Polit.* **49**, 457–477 (2017).
7. Waldner, D. & Lust, E. Unwelcome change: coming to terms with democratic backsliding. *Annu. Rev. Polit. Sci.* **21**, 93–113 (2018).
8. Fish, M. S. *Postcommunism and the Theory of Democracy* (eds Roeder, P. G. et al.) 54–95 (Princeton Univ. Press, 2002).
9. McClosky, H. Consensus and ideology in American politics. *Am. Polit. Sci. Rev.* **58**, 361–382 (1964).
10. Ahmed, A. Is the American public really turning away from democracy? Backsliding and the conceptual challenges of understanding public attitudes. *Perspect. Polit.* https://doi.org/10.1017/S1537592722001062 (2022).
11. Goodman, S. W. 'Good citizens' in democratic hard times. *Ann. Am. Acad. Polit. Soc. Sci.* **699**, 68–78 (2022).
12. Grossman, G., Kronick, D., Levendusky, M. & Meredith, M. The majoritarian threat to liberal democracy. *J. Exp. Polit. Sci.* **9**, 36–45 (2022).
13. Krishnarajan, S. Rationalizing democracy: the perceptual bias and (un)democratic behavior. *Am. Polit. Sci. Rev.* **117**, 474–496 (2023).
14. Jacob, M., Wunsch, N. & Derksen, L. The demand side of democratic backsliding: how divergent understandings of democracy shape political choice. *OSF* https://doi.org/10.31219/osf.io/c64gf (2022).
15. Putnam, R. *Making Democracy Work* (Princeton Univ. Press, 1993).
16. Weingast, B. R. The political foundations of democracy and the rule of law. *Am. Polit. Sci. Rev.* **91**, 245–263 (1997).
17. Przeworski, A. Democracy as an equilibrium. *Public Choice* **123**, 253–273 (2005).
18. Reuters. US presidential election is rigged, says Donald Trump. *Guardian* (18 October 2016); https://www.theguardian.com/us-news/video/2016/oct/18/us-presidential-election-rigged-donald-trump-wisconsin-video
19. Ingraham, L. The Ingraham Angle. *Fox News* (9 September 2020).
20. Carlson, T. Tucker Carlson Tonight. *Fox News* (24 September 2020).
21. Ingraham, L. The Ingraham Angle. *Fox News* (1 October 2020).
22. Rose, J. & Baker, L. 6 in 10 Americans say U.S. Democracy is in crisis as the 'big lie' takes root. *NPR* (3 January 2022).
23. McCarty, N. *Polarization: What Everyone Needs to Know* (Oxford Univ. Press, 2019).
24. Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N. & Westwood, S. J. The origins and consequences of affective polarization in the United States. *Annu. Rev. Polit. Sci.* **22**, 129–146 (2019).
25. Pierson, P. & Schickler, E. Madison's constitution under stress: a developmental analysis of political polarization. *Annu. Rev. Polit. Sci.* **23**, 37–58 (2020).
26. Fishkin, J., Siu, A., Diamond, L. & Bradburn, N. Is deliberation an antidote to extreme partisan polarization? Reflections on 'America in one room'. *Am. Polit. Sci. Rev.* **115**, 1464–1481 (2021).
27. Lee, A. H.-Y., Lelkes, Y., Hawkins, C. B. & Theodoridis, A. G. Negative partisanship is not more prevalent than positive partisanship. *Nat. Hum. Behav.* **6**, 951–963 (2022).
28. Finkel, E. J. et al. Political sectarianism in America. *Science* **370**, 533–536 (2020).
29. Graham, M. H. Does partisan identity reduce support for electoral fairness? *OSF* https://doi.org/10.31235/osf.io/vwe36 (2020).
30. Arbatli, E. & Rosenberg, D. United we stand, divided we rule: how political polarization erodes democracy. *Democratization* **28**, 285–307 (2021).
31. Graham, M. H. & Svolik, M. W. Democracy in America? Partisanship, polarization, and the robustness of support for democracy in the United States. *Am. Polit. Sci. Rev.* **114**, 392–409 (2020).
32. Cassese, E. C. Partisan dehumanization in American politics. *Polit. Behav.* **43**, 29–50 (2021).
33. Martherus, J. L., Martinez, A. G., Piff, P. K. & Theodoridis, A. G. Party animals? Extreme partisan polarization and dehumanization. *Polit. Behav.* **43**, 517–540 (2021).
34. Ahler, D. J. & Sood, G. The parties in our heads: misperceptions about party composition and their consequences. *J. Politics* **80**, 964–981 (2018).
35. Bartels, L. Ethnic antagonism erodes Republicans' commitment to democracy. *Proc. Natl Acad. Sci. USA* **117**, 22572–22579 (2020).
36. Ocasio-Cortez, A. 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. *Twitter* (27 August 2019); https://twitter.com/AOC/status/1166502815717568512
37. Warren, E. Health care. Reproductive Freedom. Workers' rights. Dreamers' futures. Our planet. Democracy. Everything is on the line—so everything is on the table. *Twitter* (26 September 2020); https://twitter.com/ewarren/status/1309876174231949312
38. Levendusky, M. S. & Malhotra, N. (Mis)perceptions of partisan polarization in the American public. *Public Opin. Q.* **80**, 378–391 (2016).
39. Lees, J. & Cikara, M. Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nat. Hum. Behav.* **4**, 279–286 (2020).
40. Lees, J. & Cikara, M. Understanding and combating misperceived polarization. *Phil. Trans. R. Soc. B* **376**, 20200143 (2021).
41. Ruggeri, K. et al. The general fault in our fault lines. *Nat. Hum. Behav.* **5**, 1369–1380 (2021).
42. Voelkel, J. G. et al. Interventions reducing affective polarization do not necessarily improve anti-democratic attitudes. *Nat. Hum. Behav.* **7**, 55–64 (2022).
43. Kteily, N., Hodson, G. & Bruneau, E. They see us as less than human: metadehumanization predicts intergroup conflict via reciprocal dehumanization. *J. Pers. Soc. Psychol.* **110**, 343–370 (2016).
44. Moore-Berg, S. L., Ankori-Karlinsky, L.-O., Hameiri, B. & Bruneau, E. Exaggerated meta-perceptions predict intergroup hostility between American political partisans. *Proc. Natl Acad. Sci. USA* **117**, 14864–14872 (2020).
45. Landry, A. P., Ihm, E., Kwit, S. & Schooler, J. W. Metadehumanization erodes democratic norms during the 2020 presidential election. *Anal. Soc. Issues Public Policy* **21**, 51–63 (2021).
46. Landry, A. P., Schooler, J. W., Willer, R. & Seli, P. Reducing explicit blatant dehumanization by correcting exaggerated meta-perceptions. *Soc. Psychol. Pers. Sci.* **14**, 407–418 (2021).
47. Mernyk, J. S., Pink, S. L., Druckman, J. N. & Willer, R. Correcting inaccurate metaperceptions reduces Americans' support for partisan violence. *Proc. Natl Acad. Sci. USA* **119**, e2116851119 (2022).
48. Pasek, M. H., Ankori-Karlinsky, L.-O., Levy-Vene, A. & Moore-Berg, S. L. Misperceptions about out-partisans' democratic values may erode democracy. *Sci. Rep.* **12**, 16284 (2022).
49. Broockman, D., Kalla, J. & Westwood, S. Does affective polarization undermine democratic norms or accountability? Maybe not. *Amer. Jour. Pol. Sci.* https://doi.org/10.1111/ajps.1271 (2022).

50. Voelkel, J., Stagnaro, M., Chu, J., Pink, S. & Mernyk, J. Megastudy identifying successful interventions to strengthen Americans' democratic attitudes. *OFS* https://doi.org/10.31219/osf.io/y79u5 (2023).
51. Carey, J. M., Helmke, G., Nyhan, B., Sanders, M. & Stokes, S. Searching for bright lines in the Trump presidency. *Perspect. Polit.* **17**, 699–718 (2019).
52. Ahler, D. J. Self-fulfilling misperceptions of public polarization. *J. Politics* **76**, 607–620 (2014).
53. Bicchieri, C. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms* (Oxford Univ. Press, 2016).
54. Imai, K., Keele, L. & Tingley, D. A general approach to causal mediation analysis. *Psychol. Methods* **15**, 309–334 (2010).
55. Tingley, D., Yamamoto, T., Hirose, K., Keele, L. & Imai, K. Mediation: R package for causal mediation analysis. R package version 4.5.0 (2014).
56. Ads by political rivals in Utah go viral. *Good Morning America* https://www.youtube.com/watch?v=A8yaf5OJRu8 (2020).
57. Mazepus, H. & Toshkov, D. Standing up for democracy? Explaining citizens' support for democratic checks and balances. *Comp. Polit. Stud.* **55**, 1271–1297 (2022).
58. Şaşmaz, A., Yagci, A. H. & Ziblatt, D. How voters respond to presidential assaults on checks and balances: evidence from a survey experiment in Turkey. *Comp. Polit. Stud.* **55**, 1947–1980 (2022).
59. Simonovits, G., McCoy, J. & Littvay, L. Democratic hypocrisy and out-group threat: explaining citizen support for democratic erosion. *J. Politics* **84**, 1806–1811 (2022).
60. Matanock, A. M. How international actors help enforce domestic deals. *Annu. Rev. Polit. Sci.* **23**, 357–383 (2020).
61. Hyde, S. D. Democracy's backsliding in the international environment. *Science* **369**, 1192–1196 (2020).
62. Lührmann, A. & Lindberg, S. I. A third wave of autocratization is here: what is new about it? *Democratization* **26**, 1095–1113 (2019).
63. Cleary, M. R. & Öztürk, A. When does backsliding lead to breakdown? Uncertainty and opposition strategies in democracies at risk. *Perspect. Polit.* **20**, 205–221 (2022).
64. Gamboa, L. *Resisting Backsliding: Opposition Strategies Against the Erosion of Democracy* (Cambridge Univ. Press, 2022).
65. Al Jazeera. Brazil election: 'It is over,' Bolsonaro tells supreme court. *Al Jazeera and News Agencies* (2 November 2022).
66. Ternovski, J. & Orr, L. A note on increases in inattentive online survey-takers since 2020. *J. Quant. Descr. Digit. Media* **2** (2022).
67. Miratrix, L. W., Sekhon, J. S., Theodoridis, A. G. & Campos, L. F. Worth weighting? How to think about and use weights in survey experiments. *Polit. Anal.* **26**, 275–291 (2018).
68. van der Loo, M. simputation: simple imputation. R package version 0.2.8 (2022).
69. R Core Team. *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, 2021).
70. Choi, J., & Shen, S. Two-sample instrumental-variables regression with potentially weak instruments. *Stata J.* **19**, 581–597 (2019).
71. Braley, A. & Lenz, G. S. Replication code and data for 'Why voters who value democracy participate in democratic backsliding'. *OSF* https://doi.org/10.17605/OSF.IO/4BDET (2023).

Acknowledgements

For helpful feedback, we thank C. Amat, L. Barden-Hair, J. Barker, A. Berinsky, C. Bicalho, J. Chu, D. Bischof, D. Broockman, J. Druckman, S. Fish, J. Fishkin, M. Graham, A. Guess, K. Hansen, C. Hosam, S. Hyde, H. Jefferson, M. Kagan, J. Krosnick, M. Landau-Wells, N. Malhotra, A. Matanock, J. Mernyk, C. Mo, E. Moro, J. Pan, S. Pink, D. Rand, C. Redekopp, E. Schickler, R. Slothus, N. Stagnaro, J. Voelkel, R. Willer, A. Wojtanik, A. Yan and S. S. You. We also thank J. Levy for research assistance. We also thank Aarhus University, Aletheia, MIT Connection Science, the MIT Media Lab Human Dynamics Group, the Stanford Communications Department, the Stanford Polarization and Social Change Lab, the Strengthening Democracy Challenge, the UC Berkeley Political Science Department, the Broockman-Lenz Lab, the International Conference on Computational Social Science (IC2S2 2020), the Bridging Divides & Strengthening Democracy Conference (2022), the Society for Personality and Social Psychology Conference (SPSP 2023), the Association for Psychological Science (APS 2023), and the American Political Science Association (APSA 2023). Funding for this study was provided by the Massachusetts Institute of Technology (A.P.) and the University of California, Berkeley (G.L.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

A.B. developed the concept and designed studies 1 and 2a. A.B. and G.S.L. collaborated on the design of studies 2b and 3. A.B. and G.S.L. fielded the studies, performed the final analysis, constructed the figures and wrote the paper. G.S.L., D.A., H.R. and A.P. supervised studies 1 and 2a, and G.S.L. supervised studies 2b and 3.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-023-01594-w.

Correspondence and requests for materials should be addressed to Alia Braley.

Peer review information *Nature Human Behaviour* thanks Michael Pasek, Jennifer McCoy and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Alia Braley

Last updated by author(s): Apr 26, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The survey data for this paper was generated using Qualtrics software, Versions [July 2021, August 2021, September 2021, and May 2022]. The Lucid Platform was used to sample subjects in July 2021, August, 2021, and September 2021, and the Mechanical Turk platform was used to sample subjects in May 2022.
Data analysis	R version 4.1.2 was used to analyze the data including the simputation package version 0.2.7 and the mediation package version 4.5.0. Stata version 15.1 was also used in one analysis including the weaktsiv package 9/20/2019. The code and data to replicate all results in this paper and in the Supplementary Information is available at www.doi.org/10.17605/OSF.IO/4BDET

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data generated during and analyzed in this paper is available at the following OSF project page: https://osf.io/4bdet/

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	We collected data on self-reported gender in Study 1 and Study 2 via Lucid's data on respondent gender. In Study 1 (N=1,973), 57% reported female, and in Study 2 (N=2,545), 52% reported female. Our analysis finds no significant relationship between gender and our outcome of interest - willingness to subvert democratic norms. As a result, we do not ask participants their gender in Study 3, run on Mechanical Turk.
Population characteristics	See "Behavioural & Social Sciences Study Design" section.
Recruitment	See "Behavioural & Social Sciences Study Design" section.
Ethics oversight	The University of California, Berkeley Committee for the Protection of Human Subjects (CPHS) and Office for the Protection of Human Subjects (OPHS) granted exemption for these studies under Principle Investigator, Gabriel Lenz, "as it satisfies the Federal and/or UC Berkeley requirements under categor(ies) 2,3. This is effective from January 21, 2021 through January 20, 2031. CPHS Protocol Number: 2021-01-13949. CPHS Protocol Title: Playing Dirty, The Dynamics of Democratic Backsliding in the United States. The approval is issued under the University of California, Berkeley Federalwide Assurance #00006252. Amendments to the original IRB approval were granted on 6/1/2021, 7/9/2021, 5/6/2021, and 5/25/2022.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The paper includes three online survey-based quantitative studies. Study 1: A sample of 1,973 US partisans were asked the extent to which they support upholding seven different democratic norms at the expense of partisan interests. They were also asked how they believe most opposing partisans would answer the same set of questions. Questions were randomized within and between blocks and covariates were measured including partisan identity strength, policy views, partisan animosity, and partisan dehumanization. The main analysis compared partisan expectations about the extent to which most opposing partisans would support democratic norms to real answers provided by partisans in the study. Study 2: A sample of 2,545 US partisans were randomized into treatment and control groups (50/50 split). The control group was asked how they think most opposing partisans would answer the seven questions about upholding democratic norms. The treatment group was asked the same set of questions, but after each question they were told corrective information about how most opposing partisans actually answered the question. All participants were then asked about their own willingness to uphold democratic norms in the seven scenarios as the primary DV. The secondary DV presented participants with questions about how they would vote in two different hypothetical primary elections, where one candidate supported breaking a given democratic norm from the list of seven, and one candidate opposed breaking it. Covariates were measured as in Study 1. The main analysis compared self-reported willingness to break democratic norms or vote for anti-democratic candidates between treatment and control groups. Study 3: A sample of 2,523 US partisans were randomized into treatment and control groups and provided with the questions (or, in treatment, the questions and answers) about opposing partisan commitment to democratic norms as in Study 2. In order to test durability and reduce socially desirable responding, we then successfully re-surveyed 1,848 of the sample at random between 1.4 to 47 hours after the first survey. The second survey asked the same DV questions as in Study 2, with the exception that we asked about three hypothetical primaries rather than two, and we added a covariate question about ethnic antagonism. At the end of the second survey, participants were also asked how they think most opposing partisans would answer the seven questions about democratic

	norms as a manipulation check. The main analysis compared treatment and control groups on our main dependent variables over time.
Research sample	In all three studies, we sought demographically representative samples of US partisans. Below, in parenthesis are data from the ACS 2019, which we use as benchmarks. Study 1: The research sample of 1,973 was recruited via Lucid. 54% were female (compared to 51%), 11% were 18-24 (compared to 12%), 13% were 25-34 (compared to 18%), 26% were 35-49 (compared to 25%), 32% were 50-64 (compared to 25%), and 18% were 65+ (compared to 20%). Study 2: The research sample of 2,545 was recruited via Lucid. 52% were female (compared to 51%), 13% were 18-24 (compared to 12%), 17% were 25-34 (compared to 18%), 24% were 35-49 (compared to 25%), 28% were 50-64 (compared to 25%), and 17% were 65+ (compared to 20%). Study 3: The research sample of 2,523 was recruited via mTurk. Information on gender was not collected. 6% were 18-24 (compared to 12%), 32% were 25-34 (compared to 18%), 28% were 35-49 (compared to 25%), 28% were 50-64 (compared to 25%), and 6% were 65+ (compared to 20%).
Sampling strategy	Lucid (Lucid Theorem) is an opt in online survey providers that uses demographic quotas to provide a sample that is demographically representative of the US based. Mechanical Turk is a online marketplace where participants can take surveys. On Mechanical Turk, there is no sampling procedure or quotas. Studies have generally found that experiments in representative samples of the US replicate in Lucid samples and Mechanical Turk samples. We selected sample sizes many times what would be necessary for 80% power based on pilot tests because they would allow us to show the consistency of findings across subgroups. Consistent with our preregistration, we excluded respondents from taking part in the studies if they failed initial attention checks, if they lacked a partisan affiliation or, if independent, did not lean towards one of the parties. We also excluded respondents from taking part if they were flagged by Qualtrics as likely bots or repeat respondents.
Data collection	The survey data for this paper was generated using Qualtrics software, Versions [July 2021, August 2021, September 2021, and May 2022]. Participants in these studies were recruited and participated in these studies online, and so during the time the participants participated in the experiment, they had no interaction with the researcher and both participants and the researcher were blind to individual assignment to treatment and control. At the point of data analysis, the researcher was not blind to the treatment and control status of participants.
Timing	Study 1: July 15-August 6, 2021. Study 2: September 15-29, 2021. Study 3: May 9-14, 2022.
Data exclusions	We did not exclude any respondents from the analysis.
Non-participation	Study 1: 80/2053 respondents cleared the sampling strategy rules (see above) but then dropped out. Study 2: 102/2645 respondents did the same. Study 3: In wave 1, all respondents were counted as participating. In wave 2, 2/1866 failed to answer any of the self-subversion items. As we discuss in the paper, we reinterviewed 73% of wave 1 respondents in wave 2, a high reinterview rate, so we have 27% nonparticipating in wave 2. As we report in the paper, we do not find differential attrition by condition in wave 2.
Randomization	Study 1 was an observation study in which every participant was in the same condition. Studies 2 and 3 used Qualtrics survey software to randomize respondents into treatment and control conditions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging