PNAS RESEARCH ARTICLE | SOCIAL SCIENCES

Correcting inaccurate metaperceptions reduces Americans' support for partisan violence

Joseph S. Mernyk[a,1,2], Sophia L. Pink[a,1,2], James N. Druckman[b,c], and Robb Willer[a]

Edited by Samantha Moore-Berg, University of Pennsylvania, Philadelphia, PA; received September 13, 2021; accepted February 26, 2022 by Editorial Board Member Margaret Levi

Scholars, policy makers, and the general public have expressed growing concern about the possibility of large-scale political violence in the United States. Prior research substantiates these worries, as studies reveal that many American partisans support the use of violence against rival partisans. Here, we propose that support for partisan violence is based in part on greatly exaggerated perceptions of rival partisans' support for violence. We also predict that correcting these inaccurate "metaperceptions" can reduce partisans' own support for partisan violence. We test these hypotheses in a series of preregistered, nationally representative, correlational, longitudinal, and experimental studies (total $n = 4{,}741$) collected both before and after the 2020 US presidential election and the 2021 US Capitol attack. In Studies 1 and 2, we found that both Democrats' and Republicans' perceptions of their rival partisans' support for violence and willingness to engage in violence were very inaccurate, with estimates ranging from 245 to 442% higher than actual levels. Further, we found that a brief, informational correction of these misperceptions reduced support for violence by 34% (Study 3) and willingness to engage in violence by 44% (Study 4). In the latter study, a follow-up survey revealed that the correction continued to significantly reduce support for violence approximately 1 mo later. Together, these results suggest that support for partisan violence in the United States stems in part from systematic overestimations of rival partisans' support for violence and that correcting these misperceptions can durably reduce support for partisan violence in the mass public.

political violence | metaperceptions | conflict | political polarization

Significance

Prominent events, such as the 2021 US Capitol attack, have brought politically motivated violence to the forefront of Americans' minds. Yet, the causes of support for partisan violence remain poorly understood. Across four studies, we found evidence that exaggerated perceptions of rival partisans' support for violence are a major cause of partisans' own support for partisan violence. Further, correcting these false beliefs reduces partisans' support for and willingness to engage in violence, especially among those with the largest misperceptions, and this effect endured for 1 mo. These findings suggest that a simple correction of partisans' misperceptions could be a practical and scalable way to durably reduce Americans' support for, and intentions to engage in, partisan violence.

In recent years, scholars and analysts have become increasingly worried about the potential for mass political violence in the United States (1). The general public shares these concerns. One 2016 poll found that more than half of American partisans reported feeling afraid of the other party (2). A 2019 poll found that Americans on average thought the United States was approaching "the edge of civil war" (3). Concerns about the threat of political violence are buttressed by scholarship suggesting troubling levels of support for partisan violence in the American public (4, 5), as well as recent incidents of political violence—such as those in Charlottesville, Portland, and Washington, D.C. Research has identified important determinants of support for political violence (5), but social scientists have only begun to identify ways to productively intervene on this problem.

One promising line of work comes from recent studies showing that American partisans' "metaperceptions" of rival partisans (i.e., their perceptions of rival partisans' views) tend to be highly inaccurate.* For example, research finds that American partisans believe that out-group partisans have higher levels of prejudice and dehumanization toward the in-group (6), are less supportive of democratic norms (8), and are more willing to obstruct the in-group for political gain (7, 9) than they are in reality. These beliefs have the potential to escalate if partisans reciprocate the animosity they perceive among their rival partisans. However, negative outcomes related to exaggerated metaperceptions can be reduced through informational corrections (7, 9).

Here, we build on this and other work, hypothesizing that partisans hold exaggerated metaperceptions of rival partisans' levels of support for and willingness to engage in violence.† These metaperceptions, in turn, exacerbate partisans' own views of political violence. Consistent with this, we expect that partisans' own levels of support for

*Prior work in the field of psychology often defines "metaperceptions" more specifically to refer to how individuals think they are perceived by others (6, 7). Here, we use the term to refer to perceptions of others' perceptions more generally.

†We study both support for and willingness to engage in partisan violence given the distinction between viewing violence as justifiable and being willing to actually engage in violence and to assess if our theoretical reasoning is robust to both attitudes and behavioral intentions regarding partisan violence (10).

Author contributions: J.S.M., S.L.P., J.N.D., and R.W. designed research; J.S.M. and S.L.P. performed research for Studies 1 and 3; J.N.D. performed research for Studies 2 and 4; J.S.M., S.L.P., and J.N.D. analyzed data; and J.S.M., S.L.P., J.N.D., and R.W. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. S.M.-B. is a guest editor invited by the Editorial Board.

See online for related content such as Commentaries.

[1]J.S.M. and S.L.P. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: mernyk@stanford.edu or sopink@stanford.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2116851119/-/DCSupplemental.

Published April 11, 2022.

and willingness to engage in violence can be effectively reduced by correcting these inaccurate metaperceptions.

At least two bodies of research suggest that partisans would overestimate rival partisans' support for partisan violence. First, social identity theory (11, 12) posits that individuals maintain positive views of the groups with which they strongly identify by making favorable relative comparisons with rival out-groups. As a result, high-identifying in-group members engage in out-group derogation—attributing negative characteristics and malicious motivations to the out-group—as a source of group esteem (13). Second, research on political perceptions shows that partisans tend to believe supporters of the rival party hold more extreme partisan views (14, 15) and stereotypical qualities (16) than they do in reality. This pattern is thought to be driven, at least in part, by partisan media environments that highlight negative and extreme views of the out-party (17–22). Inaccurate, overly negative perceptions of rival partisans could lead Americans to believe that out-partisans have more extreme views about supporting and engaging in partisan violence than they do in reality. Thus, we expect partisans will overestimate how much rival partisans support, and intend to engage in, partisan violence (Hypothesis 1).

Further, we expect that these exaggerated metaperceptions lead partisans to increase their own support for violence in response. Prior work finds that people support political violence when they feel a security threat, with violence serving as a form of protection or retribution (10, 23, 24). We predict that the more partisans believe that supporters of the rival party will support or engage in violence, the more likely they are to support or intend to engage in violence against out-partisans themselves (Hypothesis 2). This is particularly concerning as it could generate a cycle of increasing support for partisan violence in which partisans inaccurately perceive the other side to be likely to support or engage in violence, becoming more likely to support or engage in violence themselves. Such an increase in actual support for violence could in turn be perceived by out-partisans, further increasing out-partisans' own support for violence and so on. If true, this feedback loop risks a ratcheting up of partisan tensions that increases the possibility of large-scale outbreaks of actual partisan violence in the United States.

Our theoretical reasoning suggests a straightforward strategy for reducing support for violence: informing people of rival partisans' actual level of support for partisan violence. Such an intervention could prompt partisans to adjust their metaperceptions to be more accurate, in turn decreasing their own support for violence. We expect that partisans who are informed of out-partisans' actual level of support for or willingness to engage in violence will be less likely to support violence or express violent intentions themselves (Hypothesis 3).

These predictions are nontrivial. First, past work on exaggerated metaperceptions has focused on measures like dehumanization of, and affect toward, rival partisans, sentiments that do not have direct behavioral manifestations. It could be that partisans accurately perceive rival partisans' support for and willingness to engage in violence, given that this behavior is observable and the vast majority of people are never violent. Second, relative to prior domains studied, political violence could be less related to metaperceptions of rival partisans. Prior work finds that support for partisan violence (SPV) is associated with trait aggression and negative views of the political system (5, 25). Thus, it could be that SPV is best viewed as a nonnormative, extrainstitutional political behavior. If so, it may be unaffected by feedback about rival partisans' views.

Further, if supported, our hypotheses could have significant applied value beyond existing work on the causes of partisan violence, which focuses on relatively stable, first-order attributes including aggression, partisan social identity (5), cognitive rigidity (26), epistemic needs for certainty or closure (10), and antiestablishment orientations (25). Dispositional and system-level factors are consequential (27), but also difficult to alter in efforts to ameliorate violence. Elite cues are comparatively easier to implement; however, some prior research suggests elites may have limited impact on violent tendencies (28). Moreover, such cues require cooperation from political actors who may perceive benefits from stoking partisan rancor. Here, we take a different approach to addressing violent tendencies and support for violence in the mass public.

Empirical Overview

We conducted a series of correlational, longitudinal, and experimental studies to test these hypotheses. We first test whether partisans hold inaccurate metaperceptions of out-partisans' support for violence (Study 1) and willingness to engage in violence (Study 2). Then, we test whether correcting these inaccurate metaperceptions reduces support for violence (Study 3) and willingness to engage in violence (Study 4a) and whether the effects persist for several weeks following the correction (Study 4b). The studies were conducted both before and after the 2020 US presidential election and the 2021 US Capitol attack, allowing us to explore the robustness of findings across a shifting political environment.

Results

Study 1. To measure baseline support for partisan violence (SPV) and determine how accurate (or inaccurate) metaperceptions of SPV are among partisans in the United States, we conducted a preregistered, nationally representative, nonprobability survey of American Democrats and Republicans in October of 2020. We measured SPV using a four-item, 100-point scale adapted from prior work (5), measuring support for explicit violence (e.g., "How much do you feel it is justified for [own party] to use violence in advancing their political goals these days?") and support for threats that could make out-partisans fear for their safety (e.g., "When, if ever, is it OK for [own party] to send threatening and intimidating messages to [opposing party] leaders?"). We also measured metaperceptions of out-partisans' SPV—that is, how participants thought the average member of the rival party would respond to the same items—and in-party metaperceptions—that is, how participants thought the average member of their own party would respond to the items.

As shown in Fig. 1, average SPV was approximately 10 on a 100-point scale among both Democrats ($M = 9.3$, $SD = 17.2$) and Republicans ($M = 10.3$, $SD = 21.7$). Importantly, 4.5% of Democrats and 9.8% of Republicans responses to the SPV composite fell above the midpoint of the scale, indicating a small—yet concerning—proportion of participants supporting violence at a relatively high level. We next tested whether partisans' metaperceptions of out-partisans' support for violence were accurate. Democrats' estimates of Republicans' SPV ($M = 35.5$, $SD = 32.2$) significantly exceeded actual levels of support for violence among Republicans ($t[620] = 12.2$, $P < 0.001$), constituting a 245% overestimate. Similarly, Republicans overestimated Democrats' SPV ($M = 37.1$, $SD = 35.1$), significantly exceeding Democrats' actual support for violence ($t[502] = 13.3$, $P < 0.001$) constituting a 299% overestimate. These findings support our preregistered prediction (Hypothesis 1). We do not find significant

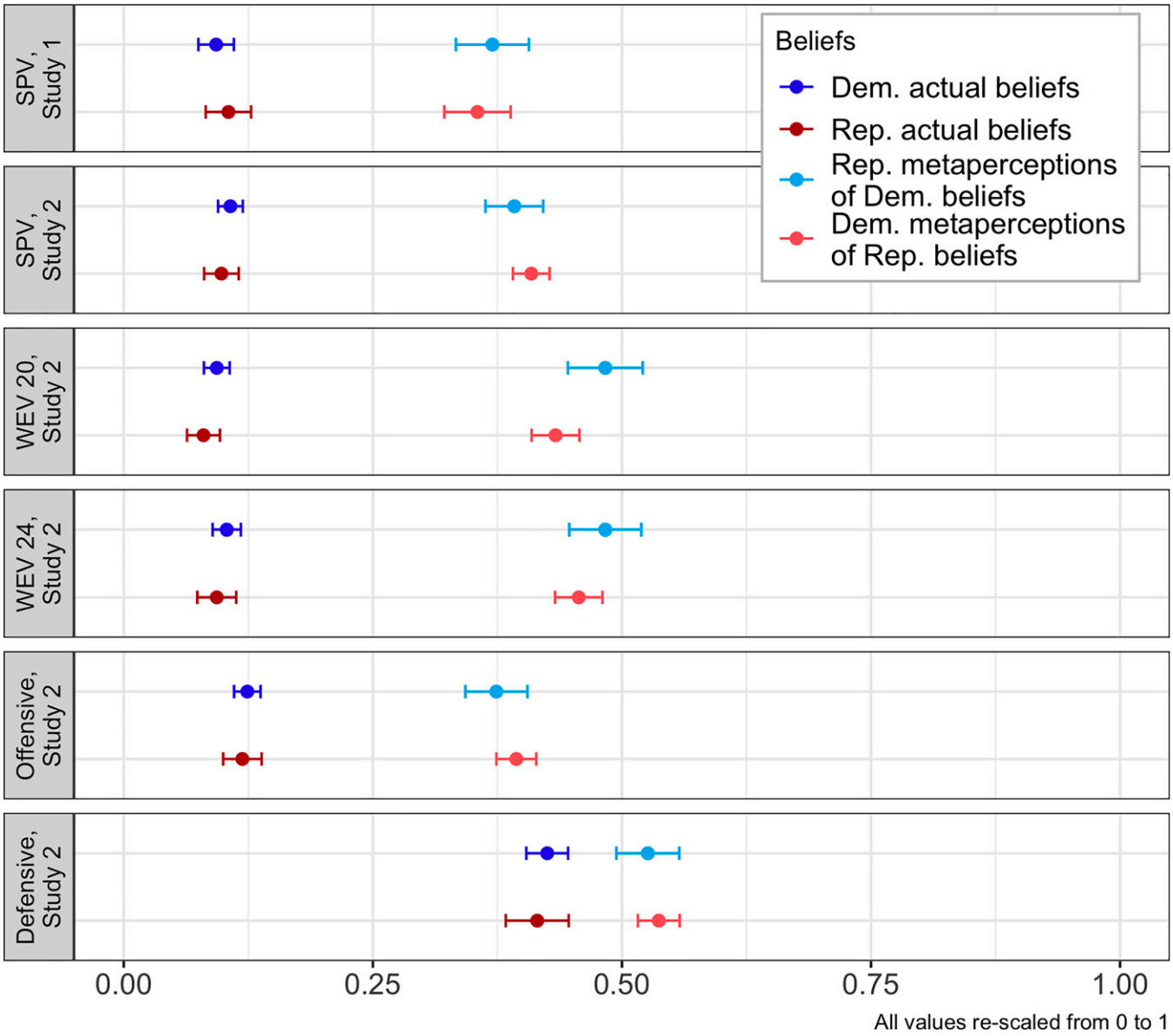

Fig. 1. Actual beliefs vs. out-party metaperceptions of SPV and WEV. Dependent variables were all rescaled to be from zero to one. Defensive indicates support for defensive violence; offensive indicates support for offensive violence. WEV refers to WEV after a contested election in 2020 and in 2024.

differences between Republicans' and Democrats' overestimates of rival partisans' SPV ($t[693] = -0.99$, $P = 0.32$).

We also tested whether participants accurately perceived the metaperceptions of in-partisans. Democrats slightly overestimated other Democrats' SPV ($M = 12.3$, $SD = 18.7$), a 31% overestimate ($t[701] = 2.20$, $P = 0.03$), and Republicans slightly overestimated other Republicans' SPV ($M = 14.0$, $SD = 22.3$), a 36% overestimate ($t[692] = 2.22$, $P = 0.03$). Thus, partisans overestimated SPV among both fellow partisans and rival partisans, although overestimates of SPV among out-partisans were far more pronounced. In-party metaperceptions of SPV were also associated with individuals' own SPV (*SI Appendix*, Table S3).

Consistent with Hypothesis 2, we found that metaperceptions of out-partisans' SPV were associated with individuals' own SPV. In a linear model controlling for participants' gender, age, race, education, income, and party, metaperceptions of out-partisans' SPV were strongly and significantly associated with individuals' own support for violence ($b = 0.097$, $P < 0.001$). This association is consistent with a potential causal effect of metaperceptions of out-partisans' SPV on partisans' actual SPV.

Study 2. We had three primary goals in Study 2. First, we sought to establish whether the inaccurate metaperceptions of out-partisans' SPV that we observed in Study 1 would also exist for perceptions of out-partisans' reported willingness to actually engage in violence (WEV). Second, we sought to assess whether the exaggerated metaperceptions of SPV in Study 1 would replicate after the prominent incidents of partisan violence that followed the 2020 presidential election, in particular the January 6th Capitol attack (our first preregistered prediction). We speculated that actual support for violence might have increased during this period—among Republicans, among Democrats, or among both—potentially reducing or eliminating the gap between actual and perceived SPV we observed in Study 1. Third, we sought to replicate the correlation between individual's SPV and metaperceptions of out-party SPV (our second

preregistered prediction) and test whether this correlation extended to WEV both within the present study and longitudinally. Our data came from two waves of a nationally representative panel survey of Democrats and Republicans. Waves were fielded in October of 2020 (including only the WEV measure) and March of 2021 (including both WEV and SPV measures).

On both waves of the survey, we included an item assessing respondents' WEV in the event of a contested presidential election (scored on a four-point scale from "not at all likely" to "very likely"). In October of 2020, the item referred to the upcoming 2020 presidential election, while in March of 2021, it referred to the 2024 presidential election. Similar to Study 1, we found a small but concerning proportion of participants who responded above the midpoint of the WEV scale. In the October wave, 7.2% of Democrats and 4.4% of Republicans reported being at least somewhat likely to use violence in the event of a contested presidential election. In the March wave, those numbers increased to 9.2% of Democrats and 7.7% of Republicans.

As shown in Fig. 1, in both waves, we found significant gaps between metaperceptions and actual WEV. In the October wave, Democrats overestimated Republicans' WEV ($M = 2.30$, $SD = 1.24$) relative to their actual reported WEV ($M = 1.24$, $SD = 0.59$), an overestimate of 442% ($t[1,557] = 23.0$, $P < 0.001$). Similarly, Republicans overestimated Democrats' WEV ($M = 2.45$, $SD = 1.33$) relative to their actual reported levels of WEV ($M = 1.28$, $SD = 0.67$), an overestimate of 417% ($t[614] = 18.5$, $P < 0.001$). This metaperception gap was similar in March of 2021, when Democrats again overestimated Republicans' WEV ($M = 2.37$, $SD = 1.23$) relative to their actual reported WEV ($M = 1.28$, $SD = 0.69$), now a 389% overestimate ($t[1,618] = 23.0$, $P < 0.001$). Republicans also again overestimated Democrats' WEV ($M = 2.45$, $SD = 1.28$) relative to Democrats' reported levels of WEV ($M = 1.31$, $SD = 0.74$), an overestimate of 368% ($t[702] = 19.1$, $P < 0.001$). These results support Hypothesis 1. Similar to Study 1, we did not find significant differences between Democrats' and Republicans' overestimates of rival partisans' WEV (October wave: $t[906] = -1.5$, $P = 0.13$; March wave: $t[1,003] = -0.78$, $P = 0.43$).

The misperceptions of SPV we documented in Study 1 also replicated following the 2020 presidential election and its aftermath. In March of 2021, Democrats' estimates of Republicans' levels of SPV were higher ($M = 40.9$, $SD = 31.7$) than their actual levels ($M = 9.8$, $SD = 20.5$), a 317% overestimate ($t[1,501] = 24.0$, $P < 0.001$). Likewise, Republicans again overestimated Democrats' SPV ($M = 39.2$, $SD = 34.3$) relative to their actual reported level ($M = 10.7$, $SD = 21.4$), a 266% overestimate ($t[726] = 17.6$, $P < 0.001$). These results support our first preregistered prediction. We again did not find significant differences between Democrats' and Republicans' estimates of rival partisans' SPV ($t[963] = 1.47$, $P = 0.14$).

To test our second preregistered prediction, we examined whether metaperceptions of out-party SPV and WEV were associated with actual SPV and WEV by regressing each violence outcome on demographic controls (gender, age, race, education level, and income) and other variables associated with support for violence (e.g., trait aggression and partisan identity strength). We also included a measure of self-monitoring, which previous work found to be associated with socially desirable response bias (29). We found that out-group metaperceptions were consistently strong predictors of SPV and WEV. Metaperceptions of SPV were significantly associated with SPV ($b = 0.171$, $P < 0.001$). Additionally, metaperceptions of WEV—measured both in October of 2020 and in March of 2021—were significantly associated with reported WEV ($b = 0.085$, $P < 0.001$ and $b = 0.218$, $P < 0.001$, respectively).

To get further insight into whether metaperceptions may shape actual willingness to engage in violence, we leveraged the longitudinal panel data that measured WEV and WEV metaperceptions in October 2020 and March 2021. We found that between-wave changes in out-party metaperceptions of WEV significantly predicted between-wave changes in WEV ($b = 0.127$, $P < 0.001$) (*SI Appendix*, Table S6).

A final goal of Study 2 was to explore perceptions of different forms of violence that partisans might engage in. In particular, we were interested in partisan' reported likelihood of engaging in offensive partisan violence (engaging in violence when out-partisans did not use violence first) and their likelihood of engaging in defensive partisan violence (engaging in violence if out-partisans did use violence first). Thus, in the March 2021 survey, we measured partisans' intentions to engage in both forms of violence and out-party metaperceptions of these intentions on 0 to 100 scales.

As shown in Fig. 1, both Democrats and Republicans showed much greater willingness to use violence in self-defense than to use violence offensively (Democrats: $M_{diff} = 30.1$, $t[1,884] = 23.5$, $P < 0.001$; Republicans: $M_{diff} = 29.6$, $t[860] = 15.5$, $P < 0.001$). Democrats and Republicans overestimated their rival partisans' willingness to engage in defensive partisan violence but to a lesser extent than other measures of violence in this study (Democrats: $M_{diff} = 12.2$, $t[985] = 6.2$, $P < 0.001$; Republicans: $M_{diff} = 10.1$, $t[1,012] = 5.2$, $P < 0.001$). Democrats and Republicans overestimated their rival partisans' willingness to engage in offensive violence to a much greater extent (Democrats: $M_{diff} = 27.5$, $t[1,481] = 19.3$, $P < 0.001$; Republicans: $M_{diff} = 24.9$, $t[723] = 14.3$, $P < 0.001$). These results indicate that partisans are primarily willing to engage in defensive violence and are much less inclined to offensive violence. Democrats and Republicans did not perceive a large gap in their rival partisans' support for offensive vs. defensive violence, primarily because they greatly overestimated rival partisans' motivations to engage in offensive partisan violence relative to actual levels. Results thus indicate that, when there is a clear security threat (i.e., a need for defensive reaction), the gap between metaperceptions and actual support for violence shrinks.

Study 3. The prior studies illustrate that American partisans greatly overestimate levels of SPV and WEV among out-partisans and that these overestimates are associated with partisans' own levels of SPV and WEV, supporting Hypotheses 1 and 2. However, this is only evidence of a correlational relationship. Thus, we next experimentally test whether providing participants true information about out-partisans' SPV would decrease participants' own SPV (Hypothesis 3). This enables us to assess whether inaccurate metaperceptions of out-party SPV play a causal role in individuals' levels of SPV. We conducted a preregistered experiment in December of 2020 focused specifically on strong partisans ($n = 555$) because pilot testing and past research (5) indicated that these individuals held higher levels of SPV than weak partisans, increasing the likelihood that the effect of our manipulation could be detected. Participants first answered the items measuring metaperceptions of SPV used in Study 1; then, they were randomly assigned to either a correction or a control condition, and finally answered the same SPV items used in Study 1. Before answering the SPV items, participants in the correction condition were presented with the average levels of SPV of respective out-partisans (as measured in Study 1) next to a summary of their own

metaperceptions. Those in the control condition only viewed a summary of their own metaperceptions (treatment and control stimuli located in *SI Appendix*).

To estimate the effect of the correction, we conducted a preregistered multiple regression analysis, regressing SPV on condition and several demographic controls (age, gender, race, education, income, and political party). We found that the correction significantly reduced support for violence in the full sample ($b = -2.8$, $P = 0.01$, Cohen's $D = 0.21$), equivalent to a 34% reduction for the average participant (Fig. 2), supporting our preregistered prediction. This finding is in line with Hypothesis 3, indicating a causal influence of metaperceptions of SPV on actual SPV.

Turning to exploratory analyses, we find that this effect was significantly moderated by the magnitude of metaperception overestimates (the difference between out-party metaperceptions and the true values) (Fig. 3), with participants who believed the out-party was more supportive of violence exhibiting larger treatment effects. Using simple slopes analysis, we find that the effect size of the correction was approximately 2.5 times larger among participants with metaperception overestimates one SD above the mean ($b = -7.5$, $P < 0.001$) compared with participants at the mean level of overestimates. Conversely, participants with smaller overestimates of rival partisans' support for violence, including those with accurate metaperceptions, were not significantly impacted by the treatment. Participants with metaperception overestimates one SD below the mean (which were slight overestimates) were not affected by the treatment ($b = 1.7$, $P = 0.3$). That the correction effect was driven by those with greater misperceptions suggests the effect was not attributable to demand effects since those with more inaccurate metaperceptions would likely be less receptive to corrections. These results show that metaperceptions of violence are causally linked to SPV in the hypothesized direction, can be corrected, and that the impact of the correction depends on the size of the metaperception.

Study 4a. Given the effectiveness of the metaperception correction in Study 3, we next tested whether our findings would replicate following the January 6th US Capitol attack. We also sought to establish the robustness of Study 3 results by focusing on violent behavior intentions using the WEV measure and recruiting a more representative sample of partisans rather than solely targeting strong partisans. We recruited participants from the same panel used in Study 2 to another survey wave in April 2021, implementing a similar correction as the one used in Study 3, although now using the WEV measure instead of SPV and data on actual levels of WEV drawn from the March 2021 wave of the survey for the correction.

Using the same modeling strategy as in Study 3, we found that participants who were exposed to the correction had significantly lower levels of WEV ($b = -0.16$, $P < 0.001$, Cohen's $D = 0.24$) than those not exposed to the correction (Fig. 3), equivalent to a 44% decrease in WEV for the average participant. Consistent with Study 3, we found that the magnitude of overestimates of rival partisans' WEV moderated the effect of the correction on WEV. Among participants with overestimates one SD above the mean, the effect was approximately 80% larger ($b = -0.30$, $P < 0.001$) than the effect on participants at the mean. Among participants with overestimates one SD below the mean (which were slight overestimates), the correction had no significant effect ($b = -0.02$, $P = 0.6$). Additionally, the main effect was not moderated by strength of party identity (*SI Appendix*, Table S16).

There were strong correlations across waves 1 and 2 of our panel survey for WEV ($r = 0.47$) and metaperceptions of WEV ($r = 0.30$). Together, these findings provide further evidence that correcting inaccurate metaperceptions of out-partisans' proclivities for violence is a valid method for reducing not only partisans' own support for violence, but also their willingness to engage in violence. Although the panel we recruited had relatively stable levels of WEV and metaperceptions of WEV over time, our correction was capable of significantly decreasing WEV and interrupting this trend.

Study 4b. Given the effectiveness of the correction at reducing WEV in Study 4a, we next test the durability of the effect. We recontacted participants from Study 4a roughly 1 mo after launching the initial study (average of 26 d) and again asked them to report their metaperceptions of rival partisans' WEV and their actual WEV. Eighty percent of the participants from Study 4a and similar proportions of participants in each condition were successfully recruited to the durability test. We find no significant difference across conditions in retention of participants to this follow-up study (*SI Appendix*, Table S18).

In a preregistered multiple regression model, we regressed metaperceptions of WEV on the experimental condition from Study 4a and controls (pretreatment WEV reported in March 2021, age, gender, race, educational attainment, income, political party). We find that the correction has a durable impact on metaperceptions of out-partisans' WEV approximately 1 mo

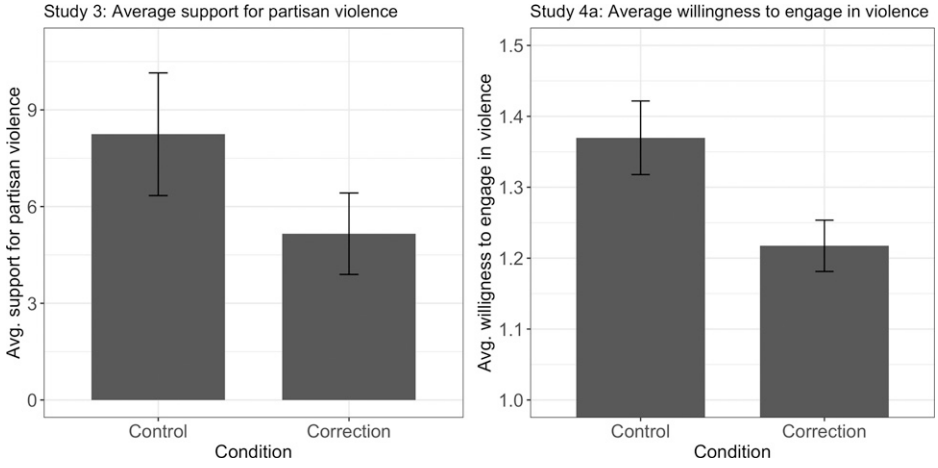

Fig. 2. Support for violence (Study 3) and WEV (Study 4a) by condition. Note that the response options differ between measures. SPV is scaled from 0 to 100; WEV is scaled from one to four (*SI Appendix* has question wording and scale labels).

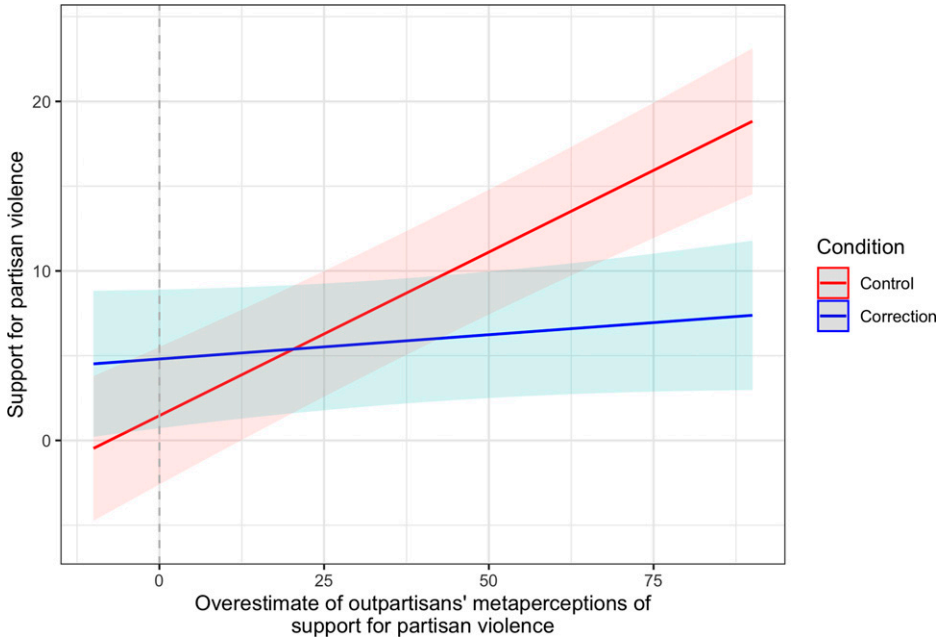

Fig. 3. The magnitude of out-group metaperception overestimates (the difference between out-group metaperceptions and the true values) moderates effect of the correction in Study 3.

after the study. Participants in the correction condition reported significantly reduced metaperceptions of out-partisans' WEV ($b = -0.30$, $P < 0.001$, Cohen's $D = 0.31$). In a parallel preregistered model, we tested whether the metaperception correction had a durable effect on WEV, finding a significant reduction in WEV ($b = -0.07$, $P = 0.03$, Cohen's $D = 0.11$). The results indicate that a brief, informational intervention—correcting participants' responses to a single survey item—continued to significantly improve the accuracy of metaperceptions of out-partisans' WEV several weeks later while also having a significant effect on participants' own reported WEV.

Discussion

Political violence has emerged as a major concern in recent years, with shocking examples of outbreaks of politically motivated violence in Charlottesville, Portland, Washington, D.C., and elsewhere. Here, we provide evidence for a previously unidentified cause of support for, and willingness to engage in, violence among partisans: exaggerated and highly inaccurate metaperceptions of out-partisans' views of violence. In support of our first hypothesis, we find that partisans overestimate rival partisans' SPV by 245 to 317% and WEV by 368 to 441%. These results hold across parties and both before and after prominent, real-world incidents of partisan violence. These inaccurate metaperceptions of out-partisans' support for violence were highly correlated with individuals' own support for violence in observational data, predictive of support for violence in longitudinal data, and causally related to support for violence in experimental data, supporting our second hypothesis. Although other types of inaccurate out-party metaperceptions have been linked to negative out-group attributions (8) and out-group spite (6), this research extends such prior work to the study of partisan violence, a major threat to democratic governance.

Across two experiments, we identified a scalable and durable intervention for reducing SPV in the contemporary United States. Correcting misperceptions of out-party support for violence significantly reduced partisans' own support for and willingness to engage in violence, supporting our third hypothesis. Additionally, participants who received the brief metaperception correction continued to report significantly lower willingness to engage in partisan violence in a follow-up survey nearly a month later. These findings have practical value because correcting inaccurate metaperceptions is likely easier than changing other more stable factors related to partisan violence, such as trait aggression or partisan identity strength, or system-level factors.

Despite the relative ease of correcting inaccurate metaperceptions of support for and willingness to engage in partisan violence, the origins of these exaggerated perceptions remain poorly understood. We speculate that these inaccurate metaperceptions may exist because of an inherent need to view one's own group positively by derogating a rival out-group—as social identity theory suggests—or they may be based on inaccurate stereotypes of out-partisans fostered by partisan media. Most likely, we expect that both of these processes work in tandem, cyclically reinforcing one another. Americans who identify with a political party are motivated to view out-partisans in a negative light, leading them to consume more partisan media, behavior that also further encourages media outlets to produce such content. Upon consuming more negative portrayals of out-partisans, partisans' perceptions of out-partisans become more exaggerated and inaccurate. These perceptions are likely to further increase in-group identification and the motivation to derogate the out-group, potentially creating a feedback loop. While this theoretical account of the emergence of inaccurate metaperceptions is broadly consistent with research on partisanship and media consumption (19), longitudinal field research of media consumption is needed to establish the dynamic, and its effects, with high confidence.

This research contributes to theory on false out-group metaperceptions and metaperception corrections in two ways. First, we find that partisans can have exaggerated metaperceptions for measures with direct behavioral manifestations, like violence. Prior work in this space has focused primarily on attitudes of out-partisans (such as partisan dehumanization and animosity), which are not directly observable. Previous research in the perceptual correction domain has not explored measures with direct behavioral manifestations, and one may expect people to have

more accurate perceptions of out-group attitudes that have directly observable manifestations. Second, we find that a metaperception correction can influence nonnormative, extrainstitutional outcomes like willingness to engage in political violence, not only more normative measures like partisan animosity.

Our focus on violence also has implications for the study of affective polarization. We found no significant relationship between affective polarization and SPV in Studies 1 and 2, which is in line with prior work that found a weak negative relationship between the two (30). Additionally, correcting misperceptions about out-party support for violence had no effect on measures of affective polarization in Study 3 (*SI Appendix*, Table S10). These variables may, thus, be less related than previous work has assumed (31). This is intriguing in light of other recent work that questions the relationship between affective polarization and various downstream outcomes (32, 33). While further research is needed to better understand why these factors are independent, one possibility is that affective polarization measures tap feelings of warmth and connection to out-partisans in the context of the extant party system, while support for and willingness to engage in partisan violence tap support for actions that would restructure or overturn this system. Thus, the latter sentiments toward political violence may be more related to antiestablishment orientations—which recent research suggests are on the rise in recent years (25)—than to routine partisan animosity.

While our research contributes to understandings of partisan violence, metaperceptions, and affective polarization, it also has several limitations. First, we rely solely on self-reported responses rather than actual behavior. It is possible that participants either overreport their willingness to engage in and support for violence to show that they dislike the other party (expressive responding) or underreport it due to social desirability bias. To address the latter, in Study 2 we included a standard measure of self-monitoring (29), a trait that is associated with socially desirable responding, finding the same patterns of results when controlling for it in multivariate analyses of results from Studies 2, 4a, and 4b (*SI Appendix*). Nonetheless, behavioral evidence of these dynamics would further increase confidence in these findings.

Second, additional work could also investigate perceptions of support for violence among copartisans and the extent to which it affects individuals' own support for violence. We found a high correlation between in-party metaperceptions and support for violence, possibly due to a projection dynamic (34). The high accuracy of in-party metaperceptions, however, leaves little room for a correction intervention.

We also do not test to what extent projection of participants' own attitudes about violence drives metaperceptions of violence. In Studies 1 and 2, we could only test for correlations. We ran additional models testing whether individuals' support for violence and WEV predict overestimates of rival partisans' support for and willingness to engage in violence (*SI Appendix*, Tables S4 and S8). We do find that SPV and WEV are significant predictors of overestimates of rival partisans' SPV and WEV. Although in Study 3, we find a strong, causal effect in the opposite direction, future research is needed to determine if there is any causal impact of SPV on metaperceptions of SPV.

Overall, our findings extend the larger body of research on misperceptions of polarization in the United States (18, 21). Prior demonstrations of partisans misperceiving the views of their rivals are concerning because these misperceptions can become self-fulfilling, with individuals reciprocating the negative views they inaccurately perceive in their rivals (7). This is perhaps nowhere as concerning as in the case of partisan violence, where there is a risk of escalating political violence based in part on misperceptions that threaten to make themselves true. The present research illustrates the potential to interrupt this dynamic, reducing misperceptions through informational corrections. However, if we are right that misperceptions of rival partisans' violent intents can become real over time, it suggests that informational corrections may be very time sensitive and most useful before the perceived risk of partisan violence coalesces into real violence.

Materials and Methods

Ethics Statement. Studies 1 and 3 were approved by the Stanford University Institutional Review Board, and Studies 2 and 4 were approved by the Northwestern University Institutional Review Board. All subjects provided informed consent. Materials, anonymized data, preregistrations, and analysis code are available at https://osf.io/rsyeh/.

Study 1. A power analysis using G*Power (35) indicated that a sample size of 580 would be required to detect a small to medium effect ($d = 0.3$) with 95% power. We recruited a sample of 1,071 Democrats and Republicans from ForthRight, a panel managed by Bovitz Inc., in October 2020 to complete the study. Although they are not probability samples, we follow common vernacular by referring to samples drawn from an internet panel to match census benchmarks for several demographic variables as "nationally representative." In total, 369 participants were excluded from analyses per our preregistered exclusion criteria (failing a simple attention check, self-identifying as an Independent, changing indicated gender, or not completing our key dependent variables), resulting in a final sample size of 702 participants, which was representative in terms of age, gender, educational attainment, race, and income. *SI Appendix*, Table S1 shows full demographic data and comparisons with population benchmarks.

Support for partisan violence (SPV) (Cronbach's $\alpha = 0.94$) was measured using four slider-style questions with scales from 0 to 100. Questions were adapted from Kalmoe and Mason (5) and included two items about direct violence ("How much do you feel it is justified for [in-partisans] to use violence in advancing their political goals these days?" and "How much do you feel it is justified for [in-partisans] to use violence if the [out-party] wins the next presidential election?") and two items about harassment ("When, if ever, is it OK for [in-partisans] to send threatening and intimidating messages to [out-party] leaders?" and "When, if ever, is it OK for an ordinary [in-partisan] in the public to harass an ordinary [out-partisan] on the internet, in a way that makes the target feel frightened?"). We measured metaperceptions of rival partisans' SPV (Cronbach's $\alpha = 0.96$) using the same items as above, but participants were asked to guess how they thought an "average [out-partisan]" would respond to each question. We measured metaperceptions of in-partisans' SPV using the same items, but participants were asked how an "average [in-partisan]" would respond.

Study 2. In October 2020, a nationally representative panel of ForthRight respondents answered one item about WEV and one item about metaperceptions of out-party violence in a survey. In March 2021, we returned to this panel and asked them questions about WEV and SPV. We only included participants who participated in both waves, identified as a Republican or a Democrat, and did not change self-identified political partisanship between the two waves. The final sample was 1,679 participants. The panel had completed multiple waves of surveys prior to those included in this paper on topics orthogonal to this research. It was nationally representative during the first wave but became slightly less representative over time due to attrition. A comparison between representative demographic benchmarks and actual demographic breakdowns in this study is included in *SI Appendix* (*SI Appendix*, Table S1), and we find that there are not large differences in key results when recalculated using representative sample weights (*SI Appendix*, Table S5).

In the first wave, participants were asked about their WEV if the rival party candidate won the 2020 presidential election and how willing they thought the rival party would be to engage in violence if the participant's own party candidate won the 2020 presidential election (metaperception of WEV). Since the second wave of the survey was conducted after the 2020 presidential election, we

asked the same questions in the context of the 2024 presidential election. In addition, SPV (Cronbach's $\alpha = 0.96$) and metaperceptions of out-party SPV (Cronbach's $\alpha = 0.96$) measures from Study 1 as well as measures of support and metaperceptions of out-party support for offensive and defensive violence were added to the second wave. WEV was measured using the items "If [an out-partisan] is declared the winner of a contested election, how likely would you be to engage in violence?" (for the October wave) and "Think about the next presidential election in 2024. If [an out-partisan] is declared the winner of a contested election, how likely would you be to engage in violence?" (for the March wave; both on a four-point scale from "not at all likely" to "very likely"). Out-party metaperceptions of violence in a contested election were measured using the items "If [an in-partisan] is declared the winner of a contested election, how likely do you think [out-party] voters would be to engage in violence?" (for the October wave) and "Think about the next presidential election in 2024. If [an in-party] candidate is declared the winner of a contested election, how likely do you think [out-party] voters would be to use violence? (for the March wave; both on a four-point scale from "not at all likely" to "very likely"). To measure support for defensive partisan violence, we asked participants, "If [out-partisans] engage in violence to achieve their political goals, how justified do you think it is for [in-partisans] to use violence in response?" (0- to 100-point scale). Out-group metaperceptions of support for defensive partisan violence were measured in the same way as the previous item, but asked participants how the average rival party member would respond to the same question. We measured support for offensive partisan violence with the item "How justified do you think it is for [in-partisans] to use violence to achieve political goals, even if [out-partisans] do not engage in violence first?" (0- to 100-point scale). Out-party metaperceptions of support for offensive partisan violence were measured in the same way as the previous item, but asked participants how the average rival party member would respond to the same item. In linear models regressing metaperceptions on each measure of violence support, we controlled for gender, age, race, education, and income as well as trait aggression, party as a social identity, difference in feelings toward the in-party and out-party, political knowledge, and self-monitoring.

Study 3. A power analysis based on pilot data indicated that a sample size of 578 would be required to detect a small to medium effect ($d = 0.3$) of our correction with 95% power. In total, 732 participants who had previously self-identified as strong partisans were recruited from a panel of Amazon Mechanical Turk users to a study conducted in December 2020; 175 participants were excluded from analyses per our pre-registered exclusion criteria (failing an attention check or not identifying as strong partisans), resulting in a final sample size of 557 participants (49.5% Democrat, 50.5% Republican) (*SI Appendix* includes full demographics). We also analyzed data from the 95 participants who did not identify as strong partisans and see slightly stronger results when these participants are included in the analyses (*SI Appendix*, Table S9).

Participants first answered the metaperceptions of out-party SPV measure from Study 1. Next, participants were randomly assigned to the correction or control condition. Participants in the correction condition were shown a table with each of the four meta-SPV questions, their guesses, and the actual average responses to each of these questions from members of the out-party (which were collected in Study 1). In the control condition, participants were also shown a summary of their responses that was identical to that of the correction group but without information about actual responses from out-party members. (Simulated treatment and control measures included in *SI Appendix*.) Participants then completed the SPV measure from Study 1. Post hoc sensitivity analyses show that we are over 70% powered to detect the effect. With the addition of participants who did not identify as strong partisans, we are 85% powered to detect the effect.

Study 4a. Participants from the ForthRight panel used in Study 2 were recontacted to participate in this study; 1,803 participants (68% Democrat, 32% Republican) (*SI Appendix* includes full demographics) completed the survey. Only participants who identified as a Republican or a Democrat and did not switch parties between waves were included in analyses. The study was fielded in April 2021. The procedure was identical to Study 3, except that the SPV and metaperceptions of SPV items were replaced with the WEV and metaperception of WEV items from the March wave in Study 2. (*SI Appendix* has treatment and control messages.) Although recontacting this panel made it less representative (due to attrition), it contained sufficient variance to explore many potential moderators (for which we found little evidence). In this sense, we are confident that our results generalize to a nationally representative sample (36–38). Post hoc sensitivity analyses show that we are over 90% powered to detect the effect.

Study 4b. The study was fielded in May 2021. Participants who completed Study 4a were recontacted; 1,447 participants completed the follow-up survey. Only participants who identified as a Republican or a Democrat were included in analyses. The average time between participating in Studies 4a and 4b was 26.06 d. When participants entered the study, they answered three questions in the following order: metaperceptions of WEV, WEV, and one item from the SPV scale. Roughly equal proportions of participants who had been assigned to the correction and control groups in Study 4a completed Study 4b. Although we did not preregister a model excluding those who switched parties between Studies 2, 4a, and 4b, we believe it is a better measure of the durability effect because participants who switched parties were exposed to different questions in each wave (i.e., the out-party would be switched between waves). The exclusion of these participants does not impact our findings, and full results of the preregistered model and a model applying these exclusion criteria are in *SI Appendix*. Post hoc sensitivity analyses show that we are over 60% powered to detect the effect.

Data Availability. Anonymized data, analysis code, materials, and preregistration data have been deposited in the Open Science Framework (https://osf.io/rsyeh/). All other study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. We thank the Stanford Center on Philanthropy and Civil Society and the Institute for Policy Research at Northwestern University for funding this research.

Author affiliations: [a]Department of Sociology, Stanford University, Stanford, CA 94305; [b]Department of Political Science, Northwestern University, Evanston, IL 60208; and [c]Institute for Policy Research, Northwestern University, Evanston, IL 60208

1. L. Diamond, L. Drutman, T. Lindberg, N. P. Kalmoe, L. Mason, Americans increasingly believe violence is justified if the other side wins. *Politico Magazine*, 1 October 2020. https://www.politico.com/news/magazine/2020/10/01/political-violence-424157. Accessed 28 January 2022.
2. Pew Research Center, How do the political parties make you feel? (2016). https://www.pewresearch.org/politics/2016/06/22/6-how-do-the-political-parties-make-you-feel/. Accessed 28 January 2022.
3. Georgetown University Institute of Politics and Public Service, New poll: Voters find political divisions so bad, believe U.S. is two-thirds of the way to "edge of civil war" (2019). https://politics.georgetown.edu/2019/10/23/new-poll-voters-find-political-divisions-so-bad-believe-u-s-is-two-thirds-of-the-way-to-edge-of-a-civil-war/. Accessed 28 January 2022.
4. K. Arceneaux, R. Truex, Donald Trump and the lie. PsyArXiv [Preprint] (2021). https://psyarxiv.com/e89ym (Accessed 28 January 2022).
5. N. Kalmoe, L. Mason, "Lethal mass partisanship: Prevalence, correlates, & electoral contingencies" in *National Capital Area Political Science Association American Politics Meeting* (2019). https://www.dannyhayes.org/uploads/6/9/8/5/69858539/kalmoe_mason_ncapsa_2019_lethal_partisanship_final_lmedit.pdf. Accessed 28 January 2022.
6. S. L. Moore-Berg, L. O. Ankori-Karlinsky, B. Hameiri, E. Bruneau, Exaggerated meta-perceptions predict intergroup hostility between American political partisans. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 14864–14872 (2020).
7. J. Lees, M. Cikara, Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nat. Hum. Behav.* **4**, 279–286 (2020).
8. M. Pasek, L.-O. A. Karlinsky, A. Levy-Vene, S. Moore-Berg, Biased and inaccurate meta-perceptions about out-partisans' support for democratic principles may erode democratic norms. PsyArXiv [Preprint] (2021). https://psyarxiv.com/qjy6t (Accessed 28 January 2022).
9. K. Ruggeri *et al.*, The general fault in our fault lines. *Nat. Hum. Behav.* **5**, 1369–1380 (2021).
10. D. Webber, A. Kruglanski, E. Molinario, K. Jasko, Ideologies that justify political violence. *Curr. Opin. Behav. Sci.* **34**, 107–111 (2020).
11. H. Tajfel, J. Turner, "An integrative theory of intergroup conflict" in *The Social Psychology of Intergroup Relations*, W. G. Austin, S. Worchel, Eds. (Brooks/Cole, Monterey, CA, 1979), pp. 33–47.
12. N. Branscombe, N. Ellemers, R. Spears, B. Doosje, "The context and content of social identity threat" in *Social Identity: Context, Commitment, Content*, N. Ellemers, R. Spears, B. Doosje, Eds. (Blackwell, Malden, MA, 1999), pp. 35–38.
13. M. Rubin, M. Hewstone, Social identity theory's self-esteem hypothesis: A review and some suggestions for clarification. *Pers. Soc. Psychol. Rev.* **2**, 40–62 (1998).
14. J. R. Chambers, D. Melnyk, Why do I hate thee? Conflict misperceptions and intergroup mistrust. *Pers. Soc. Psychol. Bull.* **32**, 1295–1311 (2006).
15. J. Westfall, L. Van Boven, J. R. Chambers, C. M. Judd, Perceiving political polarization in the United States: Party identity strength and attitude extremity exacerbate the perceived partisan divide. *Perspect. Psychol. Sci.* **10**, 145–158 (2015).
16. D. Ahler, G. Sood, The parties in our heads: Misperceptions about party composition and their consequences. *J. Pol.* **80**, 964–981 (2018).

17. S. Rathje, J. J. Van Bavel, S. van der Linden, Out-group animosity drives engagement on social media *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2024292118 (2021).
18. M. Levendusky, N. Malhotra, Does media coverage of partisan polarization affect political attitudes? *Pol. Commun.* **33**, 283–301 (2016).
19. A. Wilson, V. Parker, M. Feinberg, Polarization in the contemporary political and media landscape. *Curr. Opin. Behav. Sci.* **34**, 223–228 (2020).
20. E. Peterson, A. Kagalwala, When unfamiliarity breeds contempt: How partisan selective exposure sustains oppositional media hostility. *Am. Pol. Sci. Rev.* **115**, 585–598 (2021).
21. A. Enders, M. Armaly, The differential effects of actual and perceived polarization. *Pol. Behav.* **41**, 815–839 (2019).
22. J. N. Druckman, S. Klar, Y. Krupnikov, M. Levendusky, J. B. Ryan, (Mis)estimating affective polarization. *J. Pol.*, https://doi.org/10.1086/715603 (2022).
23. N. Kteily, G. Hodson, E. Bruneau, They see us as less than human: Metadehumanization predicts intergroup conflict via reciprocal dehumanization. *J. Pers. Soc. Psychol.* **110**, 343–370 (2016).
24. N. Kteily, E. Bruneau, Backlash: The politics and real-world consequences of minority group dehumanization. *Pers. Soc. Psychol. Bull.* **43**, 87–104 (2017).
25. J. Uscinski *et al.*, American politics in two dimensions: Partisan and ideological identities versus anti-establishment orientations. *Am. J. Pol. Sci.* **65**, 877–895 (2021).
26. L. Zmigrod, A. Goldenberg, Cognition and emotion in extreme political action: Individual differences and dynamic interactions. *Curr. Dir. Psychol. Sci.* **30**, 218–227 (2021).
27. H. Bartusevicius, A. Bor, F. Jørgensen, M. Petersen, The psychological burden of the COVID-19 pandemic is associated with antisystemic attitudes and political violence. *Psychol. Sci.* **32**, 1391–1403 (2021).
28. K. Clayton *et al.*, Elite rhetoric can undermine democratic norms. *Proc. Natl. Acad. Sci. U.S.A.* **118**. (2021).
29. E. Connors, Y. Krupnikov, J. Ryan, How transparency affects survey responses. *Public Opin. Q.* **83**, 185–209 (2019).
30. N. P. Kalmoe, L. Mason, Most Americans reject partisan violence, but there is still cause for concern. Voter Study Group (7 May 2020). https://www.voterstudygroup.org/blog/has-american-partisanship-gone-too-far. Accessed 15 December 2021.
31. E. J. Finkel *et al.*, Political sectarianism in America. *Science* **370**, 533–536 (2020).
32. J. G. Voelkel *et al.*, Interventions reducing affective polarization do not improve anti-democratic attitudes. OSF [Preprint] (2021). https://osf.io/7evmp/ (Accessed 28 January 2022).
33. D. Broockman, J. Kalla, S. Westwood, Does affective polarization undermine democratic norms or accountability? Maybe not. OSF [Preprint] (2020). https://osf.io/9btsq/ (Accessed 28 January 2022).
34. J. M. Robbins, J. I. Krueger, Social projection to ingroups and outgroups: A review and meta-analysis. *Pers. Soc. Psychol. Rev.* **9**, 32–47 (2005).
35. F. Faul, E. Erdfelder, A. Buchner, A. G. Lang, Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behav. Res. Methods* **41**, 1149–1160 (2009).
36. J. N. Druckman, C. D. Kam, "Students as experimental participants: A defense of the 'narrow data base'" in *Cambridge Handbook of Experimental Political Science*, J. N. Druckman, D. P. Greene, J. H. Kuklinski, A. Lupia, Eds. (Cambridge University Press, New York, NY, 2011), pp. 41–57.
37. K. J. Mullinix, T. J. Leeper, J. N. Druckman, J. Freese, The generalizability of survey experiments. *J. Exp. Political Sci.* **2**, 109–138 (2015).
38. A. Coppock, T. J. Leeper, K. J. Mullinix, Generalizability of heterogeneous treatment effect estimates across samples. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 12441–12446 (2018).