JUSTIFYING DISSENT*

LEONARDO BURSZTYN
GEORGY EGOROV
INGAR HAALAND
AAKAASH RAO
CHRISTOPHER ROTH

Dissent plays an important role in any society, but dissenters are often silenced through social sanctions. Beyond their persuasive effects, rationales providing arguments supporting dissenters' causes can increase the public expression of dissent by providing a "social cover" for voicing otherwise stigmatized positions. Motivated by a simple theoretical framework, we experimentally show that liberals are more willing to post a tweet opposing the movement to defund the police, are seen as less prejudiced, and face lower social sanctions when their tweet implies they had first read credible scientific evidence supporting their position. Analogous experiments with conservatives demonstrate that the same mechanisms facilitate anti-immigrant expression. Our findings highlight both the power of rationales and their limitations in enabling dissent and shed light on phenomena such as social movements, political correctness, propaganda, and antiminority behavior. *JEL Codes:* D83, D91, P16, J15.

I. INTRODUCTION

From speaking out against injustice to victimizing protected groups, dissent can be a force for or against social change and therefore plays a consequential role in any society. Fundamental to dissent are rationales—narratives disseminated by political entrepreneurs, social movements, and media outlets—that provide

* We thank the editors, Andrei Shleifer and Lawrence Katz, and five anonymous referees for very insightful comments. We also thank Davide Cantoni, Daniel Gottlieb, Ro'ee Levy, Pietro Ortoleva, Marco Tabellini, David Yang, Noam Yuchtman, and numerous seminar participants for helpful suggestions and Stelios Michalopoulos for a highly constructive discussion. Danil Fedchenko, Maximilian Fell, Takuma Habu, Hrishikesh Iyengar, Melisa Kurtis, and Stan Xie provided outstanding research assistance. We gratefully acknowledge financial support from the Pearson Institute for the Study and Resolution of Global Conflicts and the University of Chicago Social Sciences Research Center. Roth acknowledges funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2126/1-390838866. The research described in this article was approved by the University of Chicago Social and Behavioral Sciences Institutional Review Board.

The Quarterly Journal of Economics (2023), 1–49. https://doi.org/10.1093/qje/qjad007.
Advance Access publication on January 24, 2023.

arguments supporting dissenters' causes. Some rationales spur dissent through persuasion: they change people's views and, as a result, their public behavior. Yet dissent is often limited not because few people hold dissenting opinions, but because these people fear speaking their mind. Indeed, 62% of Americans agree that "The political climate these days prevents me from saying things I believe because others might find them offensive" (Ekins 2020).

Consider Democrats who oppose the movement to defund the police. In many settings, publicly expressing this opposition generates social costs: opposition to police defunding may be seen as a signal of racial intolerance. Suppose that a credible study is publicized suggesting that defunding the police would increase violent crime. This new study might increase a person's willingness to publicly oppose police defunding even if the study does not change her convictions, as long as she is able to attribute her views to the study. The key point is that the availability of this rationale opens up explanations other than racial intolerance for her position, reducing the social costs incurred by voicing it publicly and thus making her more willing to dissent.

In this article we explore the power and potential limitations of rationales in facilitating the expression of dissent. We present a simple theoretical framework demonstrating that rationales introduce "signal-jamming" that has important strategic consequences: by hindering the audience's ability to infer that a dissenter truly holds extreme beliefs, rationales lower the social cost of dissent and thereby increase the share of people willing to express their stigmatized beliefs publicly. Motivated by this framework, we experimentally examine the expression and interpretation of dissent in two contentious and policy-relevant domains: liberals' opposition to defunding the police and conservatives' support for deporting illegal immigrants. We focus on social media, where rationales from both mainstream and fringe sources are abundant and where people often face large social costs of expressing controversial opinions.

We begin by studying opposition to police reform among liberals. In a first experiment, respondents read a *Washington Post* article written by a Princeton criminologist arguing that "One of the most robust, most uncomfortable findings in criminology is that putting more officers on the street leads to less violent crime" (Sharkey 2020). Respondents then choose whether to join a campaign opposing the movement to defund the police and, conditional on doing so, decide whether to post a tweet promoting

the campaign. The experimental manipulation subtly varies the availability of a social cover in the tweet while holding fixed other potential motives to post. In particular, in the *Cover* condition, respondents' tweets indicate that they were shown the article before joining the campaign, while in the *No Cover* condition, respondents' tweets indicate that they were shown the rationale after joining the campaign.[1] The implied timing in the *Cover* condition provides these respondents with a social cover—the (implicit) justification that they joined the campaign because they were persuaded by the article's claims—while the timing implied by the *No Cover* condition eliminates this social cover. Differences in the "willingness to tweet" thus cannot be explained by the persuasiveness of the rationale—all respondents in both groups read the article—or by respondents' expectations that the rationale will persuade their followers—both versions of the tweet contain an identical description of and link to the article.

The availability of a social cover strongly affects posting behavior: in two preregistered waves of the experiment spaced a year apart, respondents are 11 percentage points more likely to post the tweet in the *Cover* condition than in the *No Cover* condition. In two placebo experiments with an identical design, but with tweets expressing support for causes associated with less stigma—as confirmed by an auxiliary survey—we find no difference between posting rates in the *Cover* and *No Cover* conditions. This evidence suggests that effects are indeed driven by (anticipated) changes in the stigma associated with dissenting expression rather than some other independent effect of the treatment. Several additional experiments provide further evidence for this interpretation and insight into the underlying mechanisms.

We conduct a second experiment to examine how the social cover shifts an audience's inferences about the motives underlying dissent and the resulting sanctions levied on dissenters. Respondents are matched with a participant who posted the tweet from the previous experiment—a previous participant assigned to the *No Cover* condition or to the *Cover* condition—and are shown the anti-defunding tweet their matched participant chose to post. They choose whether to deny a bonus to their matched participant, a measure of social sanctions. We also elicit respondents'

1. Both tweets are factually correct, as respondents in both conditions were shown the article both before and after joining the campaign.

inferences about their matched participant's underlying prejudice: respondents guess whether the participant authorized a donation to a pro-Black organization.

The results confirm that the availability of social cover shifts inference and resulting social sanctions. Respondents matched with a participant in the *Cover* condition are 7 percentage points more likely to think that their matched participant authorized the pro-Black donation (relative to a *No Cover* mean of 27%) and are 7 percentage points less likely to deny their matched participant the $1 bonus (relative to a *No Cover* mean of 47%).

We next study the effects of rationales among a different sample, conservatives, and in a different policy context, anti-immigrant policies. Here, supporting the immediate deportation of all illegal immigrants from Mexico is a stigmatized opinion that people may be reluctant to publicly express, but a similar rationale as studied in the previous experiments—concerns about crime—may be effective in shifting inference about motives and thus decreasing social sanctions. In addition to speaking to the robustness of our previous findings and examining the use of rationales by a different population (conservative rather than liberal respondents), these experiments allow us to examine how rationales can generate social cover vis-à-vis different types of audiences. In particular, opposition to police defunding is primarily stigmatized by liberals' in-group (fellow liberals) rather than their out-group (conservatives); in contrast, support for deportation is primarily stigmatized by conservatives' out-group (liberals) rather than their in-group (fellow conservatives).

The experimental manipulation follows the logic in our first experiment: in the *Cover* condition, respondents' tweets indicate that they were exposed to a rationale—a clip of Fox News anchor Tucker Carlson arguing that illegal immigrants commit violent crimes at vastly higher rates than citizens—before joining the campaign, while in the *No Cover* condition, respondents' tweets indicate that they were exposed to the rationale after joining the campaign. Our findings corroborate the importance of rationales in facilitating the expression of dissent: respondents are 17 percentage points more likely to post the tweet in the *Cover* condition than the *No Cover* condition, relative to a *No Cover* mean of 47%. A further experiment shows that this rationale once again has strong effects on inference: respondents matched with a participant who chose to post the *Cover* tweet are 5 percentage points more likely to believe that this participant authorized

the pro-immigrant donation (relative to a *No Cover* mean of 9%) and are 7 percentage points less likely to deny their matched participant the bonus (relative to a *No Cover* mean of 80%).

Taken together, our evidence highlights the importance of rationales in facilitating dissent on both sides of the political spectrum; it sheds light on the mechanisms by which individuals and institutions can influence public behavior by shaping the supply of rationales and perceptions of their social acceptability. Our findings have important implications for how the expression of dissent responds to the availability of new narratives. First, rationales are only effective to the extent to which observers believe that they genuinely change the dissenter's beliefs: an obscure or noncredible rationale may fail to shift inference and may even backfire if it signals the dissenter's underlying type. For example, if only intolerant people tend to read a particular source, citing a novel rationale provided by this source will fail to generate social cover. This implies that the endorsement of rationales by prominent figures such as politicians or celebrities may generate particularly large "social amplifiers": such figures may not only be more credible and directly persuade more people, but also more able to generate common knowledge such that dissenters can claim they were exposed to the rationale without seeking it out directly from stigmatized sources.

Conversely, groups seeking to suppress dissent have strong incentives to silence or marginalize potential sources of rationales (for example, disinviting campus speakers or branding certain news sources as fringe), because these tactics reduce the perceived probability that people will be exposed to rationales "by chance." If successful, these groups can create and sustain a "political correctness" culture—for better or for worse—in which certain rationales are ineffective because citing the stigmatized source undermines social cover. Indeed, at the time of our experiment, only 25% of Democrats privately supported decreasing police funding (Parker and Hurst 2021). By challenging the credibility of rationales or explicitly linking them to stigmatized positions, a vocal group, even a vocal minority, can silence a majority.

I.A. Related Literature

Our work contributes to an emerging literature on narratives as drivers of economic and political behavior (Shiller 2017; Michalopoulos and Xue 2021). Related to our work is

Foerster and van der Weele (2021), which studies the communication of rationales for and against donating to prosocial causes, and Bénabou, Falk, and Tirole (2018), which models the production and circulation of justifications for morally questionable actions. Our contribution to this literature is to characterize and experimentally identify an important channel—the "social cover" effect—through which narratives, specifically rationales, shape the expression and the interpretation of dissent. Our theoretical framework and experimental evidence suggest means by which individuals and institutions can exploit this channel to facilitate or suppress dissent.

Therefore, our work also relates to a literature examining how social norms influence public behavior (Kuran 1997; Bénabou and Tirole 2006; Lacetera and Macis 2010; Ali and Lin 2013; Perez-Truglia and Cruces 2017) and to a theoretical literature on political correctness (Morris 2001; Golman 2021). Like Braghieri (2022), we examine the role of social image concerns in shaping political-correctness equilibria, though we investigate how rationales shape expression and interpretation rather than how differences between private and publicly stated views lead to information loss. As in some of this previous work (Bursztyn et al. 2020; Bursztyn, Egorov, and Fiorin 2020), our article examines how previously stigmatized public behavior can become socially acceptable, but a crucial conceptual difference is that our mechanism conditions social acceptability on the availability of a publicly observable rationale rather than the existence of misperceptions. This has important implications for interpretation and expression of dissenting views. In particular, rationales make public actions less informative about dissenters' underlying type and increase the public expression of dissent by lowering its social cost. This enables moderates who previously would have been unwilling to express dissent for fear of being labeled an extremist to voice their opinions, further hindering inference about dissenters' underlying type. In other words, our mechanism generates a "social amplifier" that magnifies rationales' persuasive effects.[2] We discuss how

2. In contrast to the information aggregation mechanisms examined in Bursztyn et al. (2020) and Bursztyn, Egorov, and Fiorin (2020), rationales may facilitate the expression of views that are privately unpopular. Of course, the two mechanisms may be mutually reinforcing. For example, dissenting views may initially emerge among only a small segment of the population, which may employ rationales to lower the cost of publicly expressing these views to the rest of society. As

political entrepreneurs can strategically supply rationales to make the expression of unpopular views more mainstream.

This latter channel helps explain the mechanisms by which media and propaganda can promote socially undesirable behavior, such as antiminority violence (e.g., Yanagizawa-Drott 2014; Adena et al. 2015; Enikolopov and Petrova 2015). Studies in this vein examining persuasion in field settings often find substantial effects (e.g., Caprettini et al. 2021)—in contrast to the relatively small effects of persuasion typically documented in a vast literature using information provision experiments (Haaland, Roth, and Wohlfart 2021). Among other plausible explanations for this discrepancy is the social-amplifier channel: widespread propaganda creates common knowledge of rationales, generating greater social cover and magnifying the effect of rationales on public behavior. Thus, our work also connects to a literature on populist political movements (e.g., Acemoglu, Egorov, and Sonin 2013; Dreyfuss, Patir, and Shayo 2021; Guriev and Papaioannou 2022) insofar as authoritarian populists are often highly skilled at producing and disseminating rationales normalizing the victimization of minority groups.

Finally, our study relates to a lab experimental literature documenting that individuals seize on even flimsy excuses for selfish behavior.[3] Because behavior is typically private in these settings, these findings can be understood through a behavioral model of self-signaling, as in Bénabou and Tirole (2011) (similarly, Grossman and Van Der Weele 2017 formalize a mechanism by which individuals engage in willful ignorance as an excuse for selfish behavior). Our work holds this "self-excuse" channel constant—all individuals in our experiments privately voice their agreement with the tweet—and we instead examine the role of rationales vis-à-vis others, shedding light on how rationales affect the expression, interpretation, and social punishment of dissent.[4] Our framework highlights levers by which agents

a consequence of this public expression, others may then be privately persuaded. An information aggregation mechanism, such as an election, can then bring these previously fringe views into the mainstream.

3. See, for example, Dana, Weber, and Kuang (2007); Hamman, Loewenstein, and Weber (2010); Cunningham and de Quidt (2015); Lazear, Malmendier, and Weber (2012); Exley (2016); Golman, Hagmann, and Loewenstein (2017); Saccardo and Serra-Garcia (2020) for work in economics and Shalvi et al. (2015) for a review of the extensive literature in psychology.

4. A seminal contribution in psychology is Langer, Blank, and Chanowitz (1978), which finds that individuals are more likely to comply with a request when

can strategically manipulate the availability or credibility of rationales to influence dissent.

The remainder of the article proceeds as follows. In Section II, we present a simple model of the use and interpretation of rationales facilitating dissenting expression. In Section III, we present experiments studying how the availability of a social cover shapes liberal respondents' willingness to publicly oppose the movement to defund the police, and how this social cover shifts their audience's beliefs about and behavior toward them. In Section IV, we present similar experiments focusing on conservative respondents in the context of anti-immigrant expression. Section V discusses implications of our findings and concludes. We list all main and auxiliary experiments in Online Appendix Table B.1.

II. Theoretical Framework

To organize these ideas and guide the experimental design, we start with a theoretical framework. All formal proofs are provided in Online Appendix A.

II.A. Setup

The society N consists of a continuum of citizens facing a binary policy decision between the status quo (Q) and change (C). There is some objective measure of social welfare from decision C, and we denote this value w. The welfare under the status quo Q is normalized to zero. From the citizens' perspective, this value is distributed normally: $w \sim \mathcal{N}\left(w_0, \sigma_w^2\right)$. This social welfare may incorporate the expected economic payoff to each citizen from enacting decision C, but it may also include externalities to people outside the society or other factors inasmuch as citizens care about them.

Apart from the objective economic consequences captured by w, citizens have idiosyncratic tastes. Specifically, citizen i gets additional utility t_i if policy C, as opposed to Q, is enacted; we refer to t_i as i's type. We assume that t_i is distributed with c.d.f. $H(\cdot)$ and p.d.f. $h(\cdot)$, has finite mean $\mathbb{E}t = \bar{t}$, and satisfies

it is justified by a reason, irrespective of whether the reason is good or bad. The authors interpret this as evidence for the "mindlessness of ostensibly thoughtful action," arguing that people have simply been conditioned to comply with requests accompanied by justifications.

the monotone hazard rate property.[5] To avoid corner cases, we assume that t_i has full support on the real line.

A citizen $i \in N$ is given a chance to publicly state support for change (decision $d_i = 1$) before an audience A. Doing so results in expressive benefit B and social cost S, so $U_i(d_i = 1) = B - S$.[6] We assume that

$$B = \beta\left(\mathbb{E}\left(w \mid *\right) + t_i\right);$$

in other words, the benefit is proportional to the sum of citizen i's posterior belief about w using all available information and i's own type. The social cost S is borne because action $d_i = 1$ may be revealing about i's type t_i, and having a high type is stigmatized by the audience.[7] For simplicity, we assume that stigma is linear in the audience's posterior about citizen i's type:

$$S = \gamma\left(\mathbb{E}_{-i}\left(t_i \mid d_i = 1, *\right) - \bar{t}\right).$$

In other words, a citizen pays a higher social cost if the audience's conditional expectation of their type is higher than the unconditional one; this would be the case, for example, if the relevant audience that pays attention to i's statement and judges citizen i consists of supporters of status quo Q, or if their opinions matter to citizen i disproportionately.

The utility from inaction ($d_i = 0$) is normalized to 0: $U_i(d_i = 0) = 0$.[8]

5. That is, $\frac{h(x)}{1-H(x)}$ is increasing in x, which is satisfied, for example, for the normal and uniform distributions.

6. By "expressive benefit," we mean utility derived from voicing one's true view independently of the social consequences. This might capture aversion to lying and/or staying silent on issues one cares about or other identity considerations.

7. Note that by an audience, we do not necessarily mean the whole society, but the subset of individuals who pay attention to and judge the citizen for supporting the change. For example, a majority of citizens may support the change C, but if the people who listen and make inferences about the sender's type disproportionately support the status quo Q, or if the judgments of these people disproportionately matter to the citizen expressing support for C, the audience should be thought of as mainly consisting of Q-types.

8. We implicitly assume that the audience does not observe that i had a chance to make the action, and thus if he chooses $d_i = 0$ he is pooled with a continuum of citizens who are passive in this model. If the audience observes that inaction is by choice, there may be social consequences in this case as well. Nevertheless, all the results go through as stated.

II.B. Analysis

In the absence of new information, the posterior of citizen i about w equals the prior w_0, and thus the benefit of action $d_i = 1$ is $B = \beta(w_0 + t_i)$. Citizen i makes the decision holding his social cost S fixed, and so chooses $d_i = 1$ if and only if

$$t_i \geqslant \frac{1}{\beta} S - w_0.$$

Thus, any equilibrium takes the threshold form, with the threshold τ_0 satisfying the condition

$$\tau_0 = \frac{\gamma}{\beta} \mathbb{E}\left(t_i \mid t_i > \tau_0\right) + k - w_0.$$

Generally speaking, the threshold need not be unique due to strategic complementarity: if not only extreme but also moderate types choose $d_i = 1$, the social cost is lower, which increases citizens' propensity to choose $d_i = 1$. However, if the distribution of t_i satisfies the monotone hazard rate property, the equilibrium is unique.

PROPOSITION 1. Suppose that $\gamma < \beta$. Then there is a unique equilibrium that takes the form of a threshold: individuals with $t_i > \tau_0$ choose $d_i = 1$ and those with $t_i < \tau_0$ choose $d_i = 0$.

In other words, the equilibrium is unique provided that the citizen's choice is not driven solely by social image concerns and that the expressive benefit from their choice is sufficiently high.

1. *Persuasive Rationales.* Suppose that citizen i, prior to making the decision, receives an informative signal $s = w + \varepsilon$, where $\varepsilon \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$. Citizen i's posterior expectation of w is then equal to

$$w_1 = \mathbb{E}\left(w \mid s\right) = w_0 \frac{\sigma_\varepsilon^2}{\sigma_w^2 + \sigma_\varepsilon^2} + s \frac{\sigma_w^2}{\sigma_w^2 + \sigma_\varepsilon^2},$$

which exceeds w_0 if and only if $s > w_0$. If indeed the signal is positive ($s > w_0$), then for a fixed social cost S, this would prompt more citizens to choose $d_i = 1$ (specifically, all citizens with $t_i \geqslant \frac{1}{\beta} S - w_1$ would do so). This corresponds to a persuasion mechanism. In addition, if the audience is aware that more moderate people

choose $d_i = 1$, the social cost of doing so is lower: intuitively, publicly supporting C is no longer a conclusive sign of extremism. Of course, a decrease in S will prompt even more people to choose $d_i = 1$, which corresponds to a social-amplifier mechanism.

In practice, rationales trigger both persuasion and social-amplifier mechanisms. Our experiments experimentally isolate the latter. To highlight the underlying theory, consider three cases. The equilibrium in each case takes a similar threshold form, but the thresholds themselves, and the social costs of dissenting, vary between cases. In the first case, the rationale is known neither to the sender nor to the audience: we refer to the associated equilibrium cutoff and equilibrium social cost as τ_0 and S_0, respectively. In the second case, the rationale is privately known to the sender, while the audience is unaware that the sender knew the rationale when making decision d_i: we denote the cutoff and social cost as τ_{priv} and S_{priv}, respectively. In the third case, the fact that the sender received the rationale is common knowledge: we denote the cutoff and social cost as τ_{pub} and S_{pub}, respectively. Intuitively, the difference between the first and second cases captures the effect of persuasion, and the difference between the second and third cases captures the role of the social-amplifier mechanism. This is formalized in the following proposition.

PROPOSITION 2. Suppose that the informative signal satisfies $s > w_0$. Then a citizen who received this signal has a higher posterior about w than the prior. The equilibrium thresholds satisfy $\tau_0 > \tau_{priv} > \tau_{pub}$ and $S_0 = S_{priv} > S_{pub}$. Furthermore, an increase in σ_ε^2 weakens all these effects, and as $\gamma \to 0$, the differences between τ_{pub} and τ_{priv} and between S_{pub} and S_{priv} vanish.

In other words, the ex ante probability that citizen i chooses $d_i = 1$ is increasing from the case of no rationale to the private signal case to the public signal case, and the equilibrium social cost is the same in the first two cases, but decreases in the case of the public signal. All these effects are attenuated if the signal is noisier and therefore less informative: citizens update less and are less likely to choose $d_i = 1$, and the associated social cost does not increase as much either. Practically, this means that if the same information is obtained from a less credible source, the changes in behavior and social cost will be smaller, and in the limit, an uninformative signal will have no effect. Finally, in

the absence of social image concerns, the social-amplifier effect disappears: that is, we should observe no difference in behavior between the public and private signal cases in nonstigmatized contexts, though there may still be a persuasion effect.

II.C. Polarizing Rationales

In reality, individuals are often presented with the same evidence, but the evidence has heterogeneous consequences (e.g., some people react favorably to news that a neighborhood is diversifying, while others react unfavorably) or is interpreted differently (e.g., due to differences in background knowledge, cognitive limitations, or behavioral biases). Can rationales still be effective even if they are not persuasive on average—that is, they "dissuade" as many people as they persuade? In Online Appendix A.3, we show that they can. The intuition is that as long as the rationale changes some people's views, the audience faces an inference problem. Assuming for simplicity that citizen i may either get a high signal $s_h > w_0$ or low signal $s_l < w_0$, the audience knows that the citizen i who chose $d_i = 1$ may have done so either because t_i is high, or because i got a high signal s_h. More precisely, the set of citizens who would choose to support change C now contains some types with $t_i < \tau_0$ (moderates who got a high signal s_h) and lacks some types with $t_i > \tau_0$ (extremists who got a low signal $s_l < w_0$). As long as the share of the former is not too small, the posterior of t_i conditional on choosing $d_i = 1$ goes down. As a result, more citizens will choose $d_i = 1$ and will face a lower social cost from doing so. Put differently, for a rationale to be effective, it does not have to be persuasive on average, so long as it hinders inference about the motives underlying the stigmatized action.

III. Opposition to Defunding the Police

The experiments presented in this article examine the expression of dissent on social media. Expression on social media is of direct interest: over 70% of Americans report using social media daily, many politicians and other prominent figures have turned to social media as a primary channel of communication with the public, and social media has been linked to a number of important real-world outcomes: protests (Enikolopov, Makarin, and Petrova 2020), hate crimes (Bursztyn et al. 2019; Müller and Schwarz forthcoming), and social movements (Levy and Mattsson

2021). Second, expressing dissent on social media—like doing so in real-world offline settings, and unlike doing so in more artificial lab settings—may have real social costs vis-à-vis a natural population about whose opinions respondents care—family members, friends, acquaintances, and current and/or future employers. Indeed, a substantial majority of hiring managers report using social media accounts as a screening tool (O'Brien 2018).

Our first two experiments examine the use and interpretation of rationales for opposing the movement to defund the police. The slogan "defund the police" rose to national prominence after the murder of George Floyd in May 2020; advocates seek to decrease funding for police departments, and many favor restricting the responsibilities of law enforcement primarily to violent crime, redirecting resources to specialized response teams such as social workers and conflict resolution specialists to deliver other services (Thompson 2020). Popular opposition to police defunding is relatively high: as of an October 2021 Pew Research survey, only 15% of adults, 25% of Democrats, and 23% of Blacks support reducing spending on policing in their area (Parker and Hurst 2021). Nonetheless, because the movement is closely linked to concerns about racial injustice—most advocates claim that the U.S. law enforcement system is fundamentally racist and requires radical reform (or abolition)—it seems a priori plausible that many liberals would feel uncomfortable publicly voicing opposition to defunding. This is particularly true given that liberal Twitter users are more interested in social justice causes and are more likely to call out perceived injustice than liberals at large (Cohn and Quealy 2019). Indeed, in a preregistered survey (Auxiliary Survey 1), we find that 80% of Democrats anticipate "strong social backlash" or "significant social backlash" if they were to express opposition to police defunding on social media.[9]

III.A. Experiment 1: Rationales and Anti-Defunding Expression

1. *Motivation for Experimental Design.* Experiment 1 studies how the social cover provided by rationales affects respondents' willingness to post a tweet opposing the movement to defund the police. Identifying this effect is challenging from both a design and ethical perspective. From a design perspective, we need to manipulate the availability of a social cover, ruling out

9. The preregistration can be accessed at https://aspredicted.org/7nm5j.pdf. See Online Appendix E.5 for experimental instructions.

other possible reasons for why a rationale might change posting behavior. For example, the rationale may affect posting behavior by changing respondents' private beliefs (persuasion), or respondents might cite the rationale to persuade others (anticipated persuasion). Identifying the cover effect requires us to hold these other channels fixed across experimental conditions. At the same time, we wish to avoid a complicated or heavy-handed intervention to maximize the extent to which our results can speak to the expression of dissent in real-world contexts. From an ethical perspective, while we want to examine the most natural possible outcome—respondents' willingness to tweet—we prefer to avoid leading respondents to actually post political content on Twitter (a particular concern in our similarly structured Experiment 3, which studies willingness to publicly support a campaign to deport all illegal Mexican immigrants). A related and conflicting goal is to avoid explicitly deceiving respondents. We address these design and ethical difficulties with an experiment aiming to (i) hold the persuasion and anticipated persuasion effects constant while varying only the availability of a social cover, (ii) measure respondents' revealed-preference willingness to express dissent on their Twitter account, (iii) avoid having respondents actually posting these tweets, and (iv) avoid explicit deception.

2. *Sample and Experimental Design.* We conducted our preregistered Experiment 1 in October 2021 with a sample of 1,122 Democrats and independents.[10] As explained below, this resulted in a final sample for analysis of 523 respondents. We then conducted a preregistered replication of the experiment (Experiment 1R) in October 2022 targeting the same final sample size.[11] For Experiment 1 and Experiment 1R, we recruited respondents from both Luc.id and CloudResearch, two survey providers widely used in the social sciences (Litman, Robinson, and Abberbock 2017; Wood and Porter 2019).

Figure I outlines the structure of Experiment 1. After completing a short attention check, we ask respondents to log in

10. See Online Appendix Table B.1 for all preregistration IDs. The full set of experimental instructions is included in Online Appendix E.1.

11. Due to changes in the sampling interface of our survey provider, we targeted only Democrats in Experiment 1R. The experimental instructions are identical to those in the original experiment with the exception of additional posttreatment questions, as discussed below.

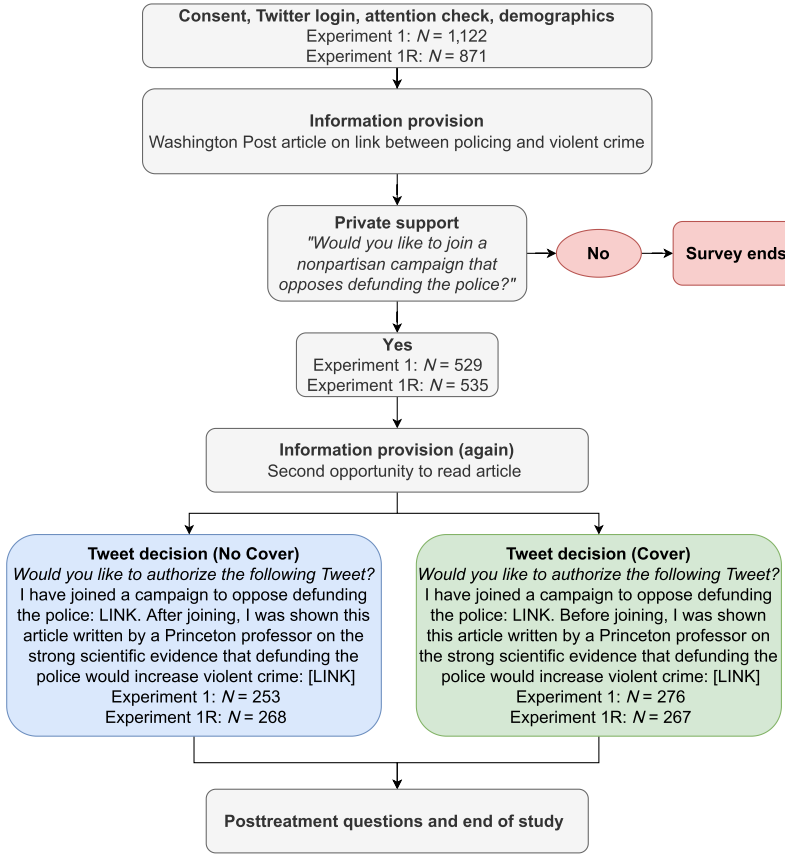

FIGURE I

Experiments 1 and 1R: Experimental Design

to our survey using their Twitter account through "Tweetability," a Twitter application we created using Twitter's Application Programming Interface that allows us to schedule tweets to be posted on the users' accounts at a future date. To an observer, these tweets look as though they were posted by the respondent him- or herself. We automatically capture respondents' Twitter handles after they log in. Respondents are assured that we will never use this application to access any private information from accounts, that all data will be securely stored until its deletion by no later than December 1, 2021 (2022 for Experiment 1R) and that we will never schedule posts on their accounts without their

explicit permission. Respondents then respond to a set of basic demographic and other background questions.

We present respondents with an op-ed written in the *Washington Post* (Sharkey 2020) by Patrick Sharkey, a professor of criminology and public affairs at Princeton University. In the article, Sharkey argues that a vast body of evidence shows that increasing policing decreases violent crime, that defunding the police is thus likely to increase violence, and that other solutions (e.g., granting communities more resources to maintain safety) will likely be more effective. After reading the article, respondents are asked if they would like to join a campaign to oppose the movement to defund the police. The survey terminates for respondents who do not join. Respondents who join are presented with the article again and informed that they can spend as long as they wish reading it.

Once they continue, we inform respondents that the campaign involves circulating a petition on Twitter opposing the movement to defund the police. We show them a screenshot of the tweet and ask if they are willing to schedule the tweet to be posted on their account. We inform respondents that the tweets of all respondents will be posted if and when we have surveyed people in all U.S. counties (a strategy which, as we explain to respondents, is often used in social media campaigns to make certain topics "trend" on users' timelines). In practice, because we target fewer respondents than the number of counties in the United States, we ensure tweets will never be posted. Our outcome can nonetheless be interpreted as revealed preference conditional on respondents believing it sufficiently probable that we will reach respondents in all counties.[12]

Respondents in the *Cover* condition are asked whether they would like to schedule the following tweet:

> I have joined a campaign to oppose defunding the police: [LINK]. Before joining, I was shown this article written by a Princeton professor on the strong scientific evidence that defunding the police would increase violent crime: [LINK]

The tweet is identical for respondents in the *No Cover* condition, with one exception: the second sentence begins "After I joined the

12. It is possible that some respondents believe it unlikely that the tweets will be posted, but for this to bias our estimated treatment effects, we would require not only that this belief is differential across treatment conditions but also that respondents who hold this belief are more or less likely to authorize the *Cover* tweet relative to the *No Cover* tweet.

campaign ..." Both tweets are factually correct (all respondents were in fact shown the article both before and after joining the campaign), but this difference in wording suggests to potential readers of the tweet that respondents in the *Cover* condition had been exposed to the scientific evidence against defunding the police before joining the campaign—and thus had a strong rationale for doing so. In contrast, the *No Cover* tweet suggests that respondents had only been exposed to the evidence after joining, and thus that the evidence could not have led them to join the campaign. This design therefore isolates the cover effect of rationales while fixing the persuasion channel (all respondents are exposed to the same information) and the anticipated persuasion channel (all respondents know their tweet's readers will see a link to the article in the tweet, conditional on the tweet being posted) across conditions.[13] By using a one-word manipulation, we hold other potential confounds, such as the length of the tweet, fixed across conditions. Our final sample is well-balanced on observables across treatment arms (Online Appendix Table B.2).

i. *Discussion of Ethical Considerations.* Although our experiment avoids explicit deception—all statements subjects see are factually true—our design clearly misleads subjects: they believe that their tweets might be posted (if we recruit respondents in every U.S. county), when in fact we purposefully recruit fewer respondents than the number of counties such that there is no chance this condition will ever be met. In experimental economics, deceiving or misleading respondents is often considered problematic due to concerns that it will lead subjects to expect deception in future experiments, potentially changing their behavior. Because subjects do not know, and never learn, that we recruited fewer respondents than the number of U.S. counties, this concern

13. One potential confound, which we cannot fully rule out, is that the respondent updates negatively about the utility they will derive from joining due to anticipated social interactions with other people who joined the campaign. While there are no differences between treatment conditions until and including the screen when respondents choose whether to join the campaign, and thus respondents should have identical beliefs about who joins the campaign, they may particularly care about social interactions with others who post the tweet, not just those who join the campaign. In practice, this is unlikely to significantly bias estimates: as described to respondents in the experimental instructions, the campaign revolves around posting a tweet to one's followers, rather than interacting with other Twitter users who posted the tweet.

TABLE I
WILLINGNESS TO POST ANTI-DEFUNDING TWEET

	Scheduled tweet					
	Main (1)	Replication (2)	Pooled (3)	Main (4)	Replication (5)	Pooled (6)
Cover	0.124***	0.092**	0.108***	0.119***	0.096**	0.108***
	(0.042)	(0.042)	(0.030)	(0.042)	(0.042)	(0.030)
No Cover mean	0.568	0.541	0.554	0.568	0.541	0.554
Controls	No	No	No	Yes	Yes	Yes
Observations	523	535	1,058	523	535	1,058
R^2	0.017	0.009	0.014	0.062	0.056	0.037

Note. The table reports results from Experiment 1 and the replication of Experiment 1 (Experiment 1R). The dependent variable is an indicator taking value 1 if the respondent chose to schedule the post. Columns (1) and (4) limit to the sample from Experiment 1; columns (2) and (5) limit to the sample from Experiment 1R; columns (3) and (6) pool the two samples and include experiment fixed effects. Controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, and a set of education indicators. Robust standard errors are reported. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

does not apply to our experiment. More generally, we concluded that the benefits of protecting participants' privacy and avoiding contributing to a political campaign outweighed the costs of misleading respondents. Moreover, our design ensures that the Twitter followers of the respondents in our survey will not be misled by respondents' tweets as to whether they read the article before or after joining the campaign—given that these tweets are never posted. We discuss the ethical considerations underlying all experimental designs in greater detail in Online Appendix C.

3. *Results.* Table I displays the results separately for the main experiment and the replication. The results are similar in both waves, so we pool the two in the discussion below and in the leftmost comparison in Figure II. Fifty-five percent of respondents authorize the tweet in the *No Cover* condition compared with 66% of respondents in the *Cover* condition ($p < .01$). These effects are stable to the inclusion of controls; the effect size corresponds to 0.25 standard deviations, comparable to or larger than the effects on persuasion generally documented in information provision experiments (Haaland, Roth, and Wohlfart forthcoming) and the effects of image concerns generally documented in experiments varying the observability of decisions

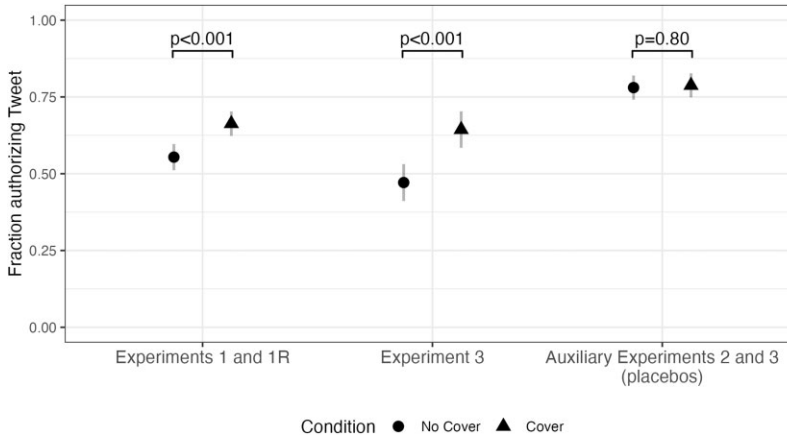

FIGURE II

Fraction Authorizing Tweets across Experiments

The figure presents results from Experiment 1 ($n = 523$) and the replication of Experiment 1 ($n = 535$), from Experiment 3 ($n = 508$, and from Auxiliary Experiments 2 ($n = 315$) and 3 ($n = 524$). We plot the fraction of respondents authorizing the tweet, separately by experiment and treatment condition. Error bars indicate 95% confidence intervals. p-values are obtained from a two-sample t-test of equality of means.

(Bursztyn and Jensen 2015).[14] This relatively large effect underscores the importance of the cover effect in driving the expression of dissent.

i. *Heterogeneity and External Validity.* We can estimate treatment effects only for respondents who were willing to log in via our app and join the campaign. We provide experimental evidence that this selection is not driving our effects in Auxiliary Experiment 5, reported below, but we also shed light on the magnitude of potential selection by investigating treatment effect heterogeneity. In Online Appendix Table B.4, column (1), we show that there is muted treatment effect heterogeneity by age, race and ethnicity, gender, and education; as shown in Online Appendix Table B.5, our estimated treatment effects

14. Indeed, in our preregistered Auxiliary Experiment 1 with the same rationale, we estimate a persuasion effect on private attitudes of 0.12 standard deviations ($p = .059$). See Online Appendix B.1.3 for details, Online Appendix E.6 for experimental instructions, and Online Appendix D for balance and representativeness tables for all auxiliary experiments.

remain stable when we reweight the sample to match the general population on these observables.[15]

III.B. Ruling Out Alternative Explanations

In this section, we consider alternative explanations for the treatment effects presented above.

1. *Direct Evidence on Perceptions of Differential Misleadingness.* To make our instructions as natural as possible, we present a plausible rationale for showing respondents the article again after they join the campaign. In particular, we write, "Since you chose to join the campaign, we wanted to give you more time reading the [article]": a natural offer to someone who had expressed particular interest in the topic. Even so, one potential concern is that respondents are more willing to schedule the *Cover* tweet ("Before joining the campaign ...") than the *No Cover* tweet ("After joining the campaign ...") because they think the latter tweet misleads respondents as to when they joined the campaign relative to reading the article.

Our first piece of evidence that this confound is not driving our treatment effects comes from two posttreatment questions we added to Experiment 1R. First, we ask respondents whether they perceived the tweet to be misleading. Second, for those who answer that they did, we ask them to explain why they felt this was the case (in open-ended format), and we hand code the responses.

Only 2% of respondents perceive the tweet to be misleading. As shown in Table II, Panel A, this fraction is in fact 2 percentage points higher in the *No Cover* group than in the *Cover* group,

15. In Experiment 1R, we collected additional information on the characteristics of respondents' Twitter accounts. In Online Appendix Table B.6, we show that treatment effects do not vary significantly by respondents' number of followers. There is some suggestive treatment effect heterogeneity by self-reported perception of the share of followers who would support defunding the police: treatment effects are driven by respondents who perceive this fraction to be between 30% and 70% of their followers. One way to interpret this finding is that respondents whose followers mostly disapprove of defunding the police may not need a cover, while those whose followers mostly approve may not be elastic to social cover given that they still expect substantial social punishment. Finally, in Online Appendix Table B.7, Panel A, we show treatment effects by partisan affiliation. Overall, while there is some evidence of heterogeneity, we are generally underpowered for these comparisons. As shown in column (2), our main treatment effects in Experiment 1 and Experiment 3 are robust to limiting the sample to Democrats and Republicans, respectively.

TABLE II

INTERPRETING EFFECTS OF A RATIONALE ON THE WILLINGNESS TO POST THE ANTI-DEFUNDING TWEET

	Mean		Treatment effect	
	No Cover	*Cover*	Coef. (std. err.)	*p*-value
Panel A: Replication of Exp. 1 (Exp. 1R, $n = 535$)				
Respondent believes tweet is ...				
Misleading	0.04	0.01	−0.02 (0.01)	.11
Misleading about timing	0.00	0.00	0.00 (0.00)	—
Panel B: Rainforest placebo (Aux. Exp. 2, $n = 315$)				
Scheduled post	0.83	0.79	−0.04 (0.04)	.38
Panel C: Daylight saving placebo (Aux. Exp. 3, $n = 524$)				
Scheduled post	0.75	0.79	0.04 (0.04)	.34
Respondent believes tweet is ...				
Misleading	0.06	0.07	0.02 (0.02)	.48
Misleading about timing	0.00	0.00	0.00 (0.00)	—
Panel D: Anticipated persuasion (Aux. Exp. 4, $n = 501$)				
Estimated share persuaded	25.34	27.23	1.90 (2.12)	.37
Panel E: Open-ended motive elicitation (Aux. Exp. 5, $n = 402$)				
Primary motives: respondent mentions ...				
Social cover	0.15	0.25	0.10 (0.04)	.02
Anticipated persuasion	0.07	0.06	−0.01 (0.02)	.67
Information	0.57	0.50	−0.07 (0.05)	.13
Potential confounds: respondent mentions ...				
Unnatural	0.01	0.01	0.01 (0.01)	.32
Misleading	0.00	0.00	0.00 (0.00)	—
Signaling	0.00	0.00	0.00 (0.00)	—
Experimenter demand	0.00	0.00	0.00 (0.00)	—
Panel F: Credibility manipulation (Aux. Exp. 6, $n = 1,017$)				
Hypothetical willingness to post				
Willing to post (high cred.)	0.57	0.67	0.11 (0.04)	.02
Willing to post (low cred.)	0.57	0.62	0.05 (0.04)	.21
Beliefs about social sanctions				
Share denying bonus (high cred.)	53.14	48.05	−5.09 (2.31)	.03
Share denying bonus (low cred.)	53.99	53.00	−0.99 (2.06)	.63

Note. In Panels A and C, "Misleading" and "Misleading about timing" are indicators for whether the respondent found the tweet misleading and whether the respondent found the tweet misleading specifically about when they read the article relative to joining the campaign. In Panel D, the dependent variable is the respondent's guess about the percentage of their followers who would join the campaign if they saw the tweet. In Panel E, the dependent variables are indicators for whether the respondent's motive falls in each of the categories. *p*-values are obtained from a two-sided *t*-test of equality of means.

though the difference is not statistically significant. Of the respondents who indicate that the tweet was misleading, none write anything related to the timing of the information provision, the timing of joining the campaign, or the "before"/"after" wording (the latter being the only difference between treatments). Moreover, restricting the sample to respondents who indicate that the tweet is not misleading leaves treatment effects virtually unchanged.

We turn to a series of experiments designed to provide further evidence against this and other potential confounds and to shed light on the underlying mechanisms. We summarize the results of these experiments in Table II.

2. *Placebo Experiments.* There may be reasons unrelated to the difference in perceived social cover that respondents prefer the *Cover* tweet to the *No Cover* tweet. Respondents may, for example, find the "After" wording strange or unnatural. To rule out that our estimates are a mechanical effect of the "before"/"after" wording, we conduct two "placebo experiments" (Auxiliary Experiments 2 and 8), where by "placebo experiments" we mean that the experiments replicate the manipulation of Experiment 1, but do so in less controversial domains in which, if the underlying mechanism driving our findings in Experiment 1 is indeed social cover, we would expect no treatment differences. One of the placebo experiments is in a relatively political domain, but with an uncontroversial policy where social sanctions are unlikely to exist: support for the conservation of the Amazon rainforest. The other placebo experiment is in a (relatively) apolitical domain where social sanctions are again unlikely to exist: eliminating daylight saving time.[16]

To confirm that social sanctions in either placebo domain are less relevant, we return to the results of Auxiliary Survey 1, in which we ask respondents whether they privately support each of four causes: defunding the police (as in Experiment 1), conserving the Amazon rainforest, eliminating daylight saving time, and immediately deporting all illegal Mexican immigrants (as in Experiment 3). For those who privately support each cause, we ask whether they anticipate that they would face social backlash if they were to express this support on social media. Online Appendix Figure B.1 confirms that respondents who privately support defunding or deportation expect substantial backlash if they were to express their views on social media (59% and 71%,

16. See Online Appendices E.7 and E.8 for experimental instructions.

respectively, expect "significant" or "strong" backlash), while respondents who privately support rainforest conservation or eliminating daylight saving expect far less backlash for expressing these views on social media (20% and 18%, respectively).

Having confirmed that anticipated social backlash is far lower in the rainforest and daylight saving contexts, we turn to the design and manipulation of the placebo experiments, which are identical to Experiment 1 except for the settings and choice of rationales. For the Amazon experiment, the rationale is a Reuters article reporting a new study that finds that over 10,000 species are at risk due to deforestation in the Amazon; for the daylight saving experiment, the rationale is an article written by a Vanderbilt neurologist on the health costs of daylight saving time.[17]

Table II, Panels B and C show no significant difference between posting rates in the *Cover* and *No Cover* conditions for either experiment. Pooling the two placebos in the rightmost comparison of Figure II, we estimate a tight null effect of *Cover* on posting rates. The large and significant difference in effect sizes between the defunding experiments and the placebo experiments suggest that effects are indeed driven by (anticipated) changes in the stigma associated with dissenting expression rather than some other independent effect of the before/after wording.[18]

Ultimately, however, there may be factors specific to the Amazon and daylight saving contexts, or the rationales we use, that lead to the lack of treatment effects of the *Cover* condition. Although highly suggestive, our placebo results cannot definitely prove our preferred interpretation of the results in Experiment

17. The Amazon tweets read: "I've joined a campaign to immediately stop the destruction of the Amazon rainforest! [Before/After] I joined the campaign, I was shown this article about how 10,000 species risk extinction in the Amazon: [LINK]. Join the campaign and sign the petition: [LINK]." The daylight saving tweets read: "I have joined a campaign to eliminate daylight saving time: [LINK]. [Before/After] joining the campaign, I was shown this article by a Vanderbilt professor of neurology on how daylight saving time is connected with serious negative health effects: [LINK]."

18. We cannot definitively rule out the possibility that the lack of treatment effects in either placebo is due to the sum of two countervailing effects: the social-cover mechanism and another mechanism by which people prefer the "after" wording because it signals that they did not have to be informed about the issue to support it. While this confound could plausibly be present in the Amazon context, where people might want to signal that they are a "good" type who does not need to be persuaded to support rainforest preservation, we view it as much less likely in the daylight saving context, in which such signaling motives are implausible.

1. For further evidence for this interpretation, and for evidence on the underlying mechanisms, we turn to a series of auxiliary experiments.

3. *Addressing Anticipated Persuasion.* It remains a possibility that respondents anticipate that the *Cover* tweet will be more persuasive to followers than the *No Cover* tweet, and that this difference drives our estimated treatment effects. Relatedly, it could be the case that respondents believe that their followers are more likely to read the article after seeing the *Cover* tweet than after seeing the *No Cover* tweet.

To mitigate concerns related to such differential anticipated persuasion, we run an auxiliary experiment (Auxiliary Experiment 3). We present Democratic and independent Twitter users with either the *Cover* or *No Cover* tweet and then ask them to estimate the share of their followers who would join the campaign after seeing their tweet, a summary statistic for the combined effects of all channels above.[19] Table II, Panel D shows a small and insignificant 1.9 percentage point difference; we can rule out differences of greater than 4.2 percentage points with 95% confidence. This suggests that differences in posting rates are not driven by differences in the anticipated persuasiveness of the tweets, as respondents' posting decisions would need to be unrealistically elastic to their beliefs about their audience's persuadability in order to generate the 12 percentage point treatment effect documented in Experiment 1. We provide further evidence against this mechanism below.

4. *Direct Evidence on the Social-Cover Mechanism.* We now provide direct evidence that our manipulation varies the perceived availability of social cover, and that this availability is an important consideration on respondents' minds when considering the expression of dissent. We conduct Auxiliary Experiment 4 with a sample of 402 Democrats with Twitter accounts recruited from Prolific. This broader sample allows us to probe the external validity of our findings. In particular, respondents are not required to grant our "Tweetability" app permissions to schedule posts on their Twitter account, which may induce selection into Experiment 1.

19. See Online Appendix B.1.4 for details and Online Appendix E.9 for experimental instructions.

i. *Experimental design*. Respondents begin by reading the article presented in Experiment 1 describing the evidence that defunding the police would increase violent crime. We ask them to imagine that at this stage, they joined a campaign to oppose defunding the police. As in the main experiment, all respondents are then given the chance to read the article again.[20] Then, respondents randomized into the *Cover* condition are asked which of two tweets they would hypothetically prefer to post: the tweet from the *Cover* condition in Experiment 1, or a *Control* tweet omitting any reference to a rationale: "I have joined a campaign to oppose defunding the police: [LINK]." Respondents randomized into the *No Cover* condition are instead asked about their hypothetical preference between posting the tweet from the *No Cover* condition in Experiment 1 or the *Control* tweet above. After respondents choose their preferred tweet, we ask them to "Please explain why you chose this tweet rather than the other tweet." Our object of interest is the difference in respondents' explanations between conditions.

A few comments about the experimental design are in order. First, we separately study preferences for the *Cover* tweet over the *Control* tweet and for the *No Cover* tweet over the *Control* tweet, rather than directly estimating preferences for the *Cover* tweet over the *No Cover* tweet. Our design thus avoids making the before/after distinction between the tweets salient, better capturing behavior both in our main experiment and in real-world settings and reducing the scope for experimenter demand effects. Similarly, our use of open-ended text to elicit motives, rather than structured questions, avoids priming respondents on particular motivations and better captures what naturally comes to mind when making their choice.

We hand code open-ended responses across three categories. "Social cover" responses mention that the respondent's preferred tweet indicates to followers that the article affected the respondent's choice to join the campaign.[21] "Anticipated persuasion" responses mention that the article might persuade others.[22] Finally,

20. See Online Appendix E.10 for experimental instructions.

21. For example, one respondent writes: "I think the evidence provided in the article is an important catalyst in why I would have joined the campaign and without any context that first tweet could be misconstrued, or even cause me to be publicly shamed."

22. For example, one respondent writes: "The tweet is meant to not only inform people of your decision, but to also advertise others to do the same. Having

"Information" responses mention that the article is informative or credible, or that it provides an explanation for why people might want to join the campaign, but do not explicitly relate the information to the respondent's own views or other people's views.[23] Many respondents classified as "information" may have had the "social cover" or "anticipated persuasion" mechanisms in mind, but wrote responses that we could not unambiguously classify into either category. We chose a conservative coding scheme for "social cover" and "anticipated persuasion" to provide a plausible lower bound.

ii. *Results*. We begin by analyzing respondents' preferences over which tweet to post. Eighty-three percent of respondents in the *No Cover* condition prefer the tweet linking to the evidence over the *Control* tweet without the evidence, compared with 87% of respondents in the *Cover* condition.[24] The high fraction choosing the tweet with the rationale (whether the *Cover* or the *No Cover* version) over the *Control* tweet suggests a widespread preference for citing evidence when engaging in dissenting expression, while the high fraction choosing the *No Cover* version constitutes further evidence that respondents do not avoid the "after" wording due to concerns about it being misleading or unnatural.

We turn to the open-ended text. The perceived social costs of dissent in this setting are further evidenced by the substantial number of responses mentioning some form of social sanctions. A relatively large fraction of respondents (20%) explicitly mention the social-cover mechanism, three times the number who mention the anticipated-persuasion mechanism (7%). The majority of responses (53%) fall into the "information" category, though many responses in this category likely meant to convey concerns relating to social cover. Focusing on treatment effects across conditions, reported in Table II, Panel E (upper part), the one-word manipulation indeed induces substantially more respondents to

supporting evidence for your cause will increase the chance of others to side and agree with you. Tweet B does this, tweet A doesn't."

23. For example, one respondent writes: "I would want others to see this article and know that I have some evidence to back my tweet."

24. The treatment effect is not comparable with the effect estimated in Experiment 1: for example, we might observe zero treatment effect in this experiment and a strong treatment effect in Experiment 1 if most respondents prefer the *Cover* tweet to the *No Cover* tweet, but strongly prefer either tweet to the *Control* tweet (while a minority of respondents exhibit strong preferences for the shorter *Control* tweet). Nonetheless, it is reassuring that the treatment effect is positive (though statistically insignificant, $p = .311$).

mention social cover (a 10 percentage point difference, or a 67% effect relative to the *No Cover* mean).

Consistent with the results of Auxiliary Experiment 4, the manipulation appears to have no effect on the probability that respondents mention that their followers will find the article persuasive. While these two pieces of evidence cannot definitively rule out differences in the anticipated persuasiveness in the tweet, they do suggest that any such differences are unlikely to drive the large treatment effects of the *Cover* condition that we document.

To gauge other potential confounds, we hand code responses along further dimensions. "Unnatural" responses mention that one tweet seems more unnatural or strangely worded than another; "misleading" responses mention that one tweet seems more misleading or deceptive than another; "signaling" responses mention that one tweet suggests that the respondent supports the cause more strongly than the other; "experimenter demand" responses mention that the experimenter wants the respondent to choose one tweet over another, or that the respondents' followers will believe this is the case. As shown in Table II, Panel E (lower part) almost no tweets fall into any of these categories.[25]

Together, the placebo experiments, the anticipated persuasion experiment, and this experiment eliciting participants' reasoning strongly suggest that the treatment effects documented in Experiment 1 are indeed driven by differences in the availability of a social cover.

5. *The Role of Credibility.* In Section II, we showed that the credibility of rationales matters: a rationale that is perceived to come from a questionable source, or whose credibility is otherwise undermined, is likely to be less effective.[26] The wording of the tweet in Experiment 1 emphasizes the credibility of the rationale, explicitly stating that the author is a Princeton professor and that the article is based on strong scientific evidence; our theory implies that reducing the credibility of the rationale will reduce its effect on posting behavior and increase the associated social sanctions.

25. Of the 15% of respondents who choose the *Control* tweet without a rationale, two-thirds cite its shorter length as the reason for doing so. Given that the one-word manipulation in Experiment 1 holds the length of the tweet fixed, preferences for shorter or longer tweets will not affect our results.

26. In particular, it is not necessary that the audience finds the rationale persuasive, but rather that the audience thinks it is plausible that the dissenter him- or herself was persuaded.

We examine the role of credibility with Auxiliary Experiment 5, which we also use to probe another dimension of external validity. In particular, the sample of Experiment 1 consists of respondents who were willing to grant our app permissions to post on their Twitter account, and thus is likely unrepresentative of the population of social media users.[27] To assess the importance of social cover in facilitating dissent among this broader population, we do not ask respondents in Auxiliary Experiment 3 to log in via Twitter; we instead ask whether respondents would have (hypothetically) been willing to post the tweet.

i. *Experimental Design*. The design of Auxiliary Experiment 5 is almost identical to the design of Experiment 1.[28] All respondents who report actively using Facebook and Twitter are eligible to participate. As in Experiment 1, they read the Sharkey article and are given the opportunity to join the campaign to oppose defunding the police; those who do not join are screened out of the survey. Remaining respondents are presented the article a second time. We then explain to them that we are interested in whether they would be willing to make the post in question (either the *Cover* or the *No Cover* post) if it were included as a campaign feature. To probe mechanisms, we also ask an incentivized (postoutcome) question eliciting perceived social sanctions: respondents estimate the share of Democrats who, upon seeing the post, chose to deny the poster a bonus. Finally, and most importantly, we cross randomize a "credibility" manipulation with our previous manipulation of social cover, resulting in four conditions. In particular, to construct "lower-credibility" versions of the tweets, we remove the references to Sharkey's academic credentials and to the scientific evidence underlying the article's claims. The revised lower-credibility tweets read:

27. To speak to the extent of selection by social desirability into "Tweetability" login, we follow Dhar, Jain, and Jayachandran (2022), who use a 13-item version of the Marlowe-Crowne social desirability scale to measure respondents' concern for social approval (Crowne and Marlowe 1960; Reynolds 1982). We implement this scale in Auxiliary Experiment 3. Online Appendix Figure B.2 shows economically and statistically insignificant differences in this score between our experimental sample (which authorized the login) and the general population, suggesting that our sample is not selected on concerns for social approval.

28. See Online Appendix E.11 for experimental instructions.

> I have joined a campaign to oppose defunding the police: [LINK]. [Before/After] joining, I was shown this article arguing that defunding the police would increase violent crime: [LINK]

Our framework predicts that this less credible rationale will generate less social cover and thus will be less effective in facilitating dissent.

ii. *Results*. We present results in Table II, Panel F. Restricting attention to the higher-credibility version of the post (i.e., the version used in Experiment 1), we find an almost identical treatment effect to that documented in Experiment 1, confirming that our results generalize to the broader sample of social media users. Turning to the lower-credibility version, we find a smaller and statistically insignificant treatment effect. While the results are qualitatively consistent with the predictions of Proposition 2, the difference between the high- and low-credibility condition is not statistically significant. Since we are generally not powered to detect significant interaction effects, we view the smaller effect size of the low-credibility condition as suggestive evidence consistent with our theory.

We find a similar pattern when we instead examine respondents' guesses as to the number of Democrats who would deny a person who made the post a bonus (our measure of perceived social sanctions): respondents believe that the social cover is effective in reducing social sanctions when the rationale is highly credible. When the rationale is less credible, the effects on perceived social sanctions are smaller and statistically insignificant (though again we lack the statistical power to detect significant interaction effects).

The perceived treatment effect of the *Cover* condition on social punishment (relative to the *No Cover* condition) implied by our data is 5 percentage points in the high-credibility condition and 2 percentage points in the low-credibility condition. As we show in the next section (Experiment 2 and Auxiliary Experiment 6), the actual treatment effect of *Cover* on social punishment is 7 percentage points in the high-credibility condition and 1 percentage point in the low-credibility condition. In other words, respondents are well calibrated about the treatment effects of *Cover*. They are also fairly well calibrated about the levels of punishment: pooling across all conditions, they expect around half of Democrats to deny the bonus, relative to the actual share of 43%. Thus, our mechanism does not require respondents to over- or underestimate the

share of their audience who would sanction them for expressing dissent, nor does it require this share to be a substantial majority.

iii. *Discussion*. The manipulation arguably generates a fairly modest reduction in credibility (as it still features an article from the *Washington Post*, a well-respected outlet among liberals): far more modest than, for example, citing a right-leaning outlet or making such a claim without any supporting evidence. Nonetheless, even this modest reduction in credibility halves the estimated effect of the rationale on posting. While drawing general conclusions about credibility would require substantially greater evidence than we provide here, our evidence suggests that one way a vocal minority might silence public dissent is by setting the "credibility bar" high, accepting only overwhelmingly conclusive evidence as legitimate.[29] A society that sets this "credibility bar" too high may stifle the expression of legitimate perspectives on issues where strong evidence does not exist. Indeed, if the credibility bar varies between groups—for example, if conservatives are seen as more easily persuaded by fake news than liberals—then groups held to a lower credibility bar can use a wider variety of rationales and thus may be willing to dissent in a wider variety of contexts.

III.C. Experiment 2: Interpretation of Anti-Defunding Rationale

Our theoretical framework implies that rationales lower the social cost of dissent by making the action less informative about type. As documented in Section III.A, respondents are more willing to dissent when they can draw upon credible rationales because they expect such rationales to reduce the informativeness of dissent for prejudice and thus lower the associated social costs. In Experiment 2, we examine whether rationales indeed serve this purpose.

1. *Sample and Experimental Design.* We conducted our preregistered Experiment 2 in November 2021 with a sample

29. Only 25% of Democrats privately support decreasing funding for police in their area, compared with 34% of Democrats who privately support increasing funding (Parker and Hurst 2021). Thus, the results of Experiment 1 and Auxiliary Experiment 5 jointly illustrate how public dissent can be silenced by a vocal minority. Through the lens of our theoretical framework, different audience members may contribute differently to overall social sanctions S: opponents of defunding may not sanction respondents who hold either opinion, while a significant fraction of supporters may heavily sanction opponents.

of Democrats and independents recruited from Prolific.[30] Our final sample of 1,040 Democrats and independents is mostly balanced on observables across treatment arms (Online Appendix Table B.9).

Online Appendix Figure B.3 outlines the structure of Experiment 2. After completing a battery of demographic and other background questions, respondents are informed that they have been matched with a previous survey participant who joined a campaign to oppose the movement to defund the police. They are then randomized into a *Cover* and a *No Cover* condition: respondents in the *Cover* condition are told that their matched participant authorized the tweet corresponding to the *Cover* condition of Experiment 1 ("Before I joined the campaign ...") whereas respondents in the *No Cover* condition are told that their matched participant authorized the *No Cover* tweet ("After I joined the campaign ...").

We begin by asking respondents the following open-ended question: "Why do you think your matched participant chose to join the campaign to oppose defunding the police?" This approach avoids priming respondents to think about particular dimensions and instead directly elicits "what comes to mind" (Gennaioli and Shleifer 2010). As a more direct measure of inference about their matched participant's prejudice, we subsequently tell them that their matched participant had the opportunity to authorize a $5 donation to the National Association for the Advancement of Colored People (NAACP) and ask them to guess whether the participant donated. Finally, we also give respondents the opportunity to authorize a $1 bonus to their matched respondent (at no cost to themselves): declining to do so is our measure of social sanction.

2. *Results.* We estimate statistically and economically significant treatment effects on all three measures of inference. The leftmost comparison in Figure III, Panel A displays the fraction of participants in the *Cover* and *No Cover* condition who believe their matched participant donated to the NAACP (results reported in regression table form in Table III, Panel A, columns (1)–(3)). Twenty-seven percent of respondents in the *No Cover* condition believe their matched participant donated, compared to 35% of respondents in the *Cover* condition ($p = .012$). Similarly, the leftmost comparison in

30. The full set of experimental instructions is included in Online Appendix E.2.

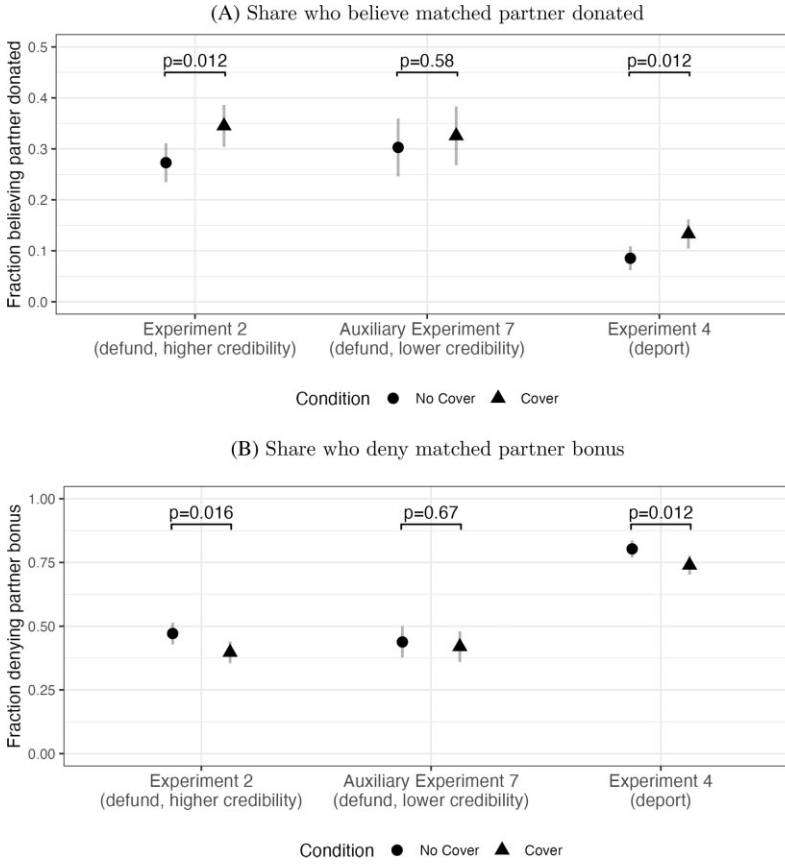

FIGURE III

Interpretation of Tweets

The figure presents results from Experiment 2 ($n = 1{,}040$), Auxiliary Experiment 7 ($n = 506$), and Experiment 4 ($n = 1{,}082$). Panel A plots the fraction of respondents who believe that their matched partner donated to the organization in question: the NAACP for Experiment 2 and Auxiliary Experiment 7, and the U.S. Border Crisis Children's Relief Fund for Experiment 4. Panel B plots the fraction of respondents who deny their partner a $1 bonus. Error bars indicate 95% confidence intervals. p-values are obtained from a two-sample t-test of equality of means.

Figure III, Panel B displays the fraction of participants who deny their matched participant a bonus (results reported in regression table form in Table III, Panel B, columns (1)–(3)). Forty-seven percent of respondents in the *No Cover* condition deny their matched participant a bonus, compared to 40% of respondents

TABLE III

INFERENCE ABOUT AND SOCIAL SANCTIONS TOWARD MATCHED ANTI-DEFUNDING RESPONDENT

	Higher credibility		Lower credibility	
	(1)	(2)	(3)	(4)
Panel A: *Belief partner donated*				
Cover	0.072**	0.072**	0.023	0.023
	(0.029)	(0.029)	(0.041)	(0.042)
No Cover mean	0.273	0.273	0.303	0.303
R^2	0.006	0.024	0.001	0.034
Panel B: *Denied bonus to partner*				
Cover	−0.074**	−0.074**	−0.019	−0.028
	(0.031)	(0.031)	(0.044)	(0.044)
No Cover mean	0.471	0.471	0.438	0.438
R^2	0.006	0.040	0.0004	0.059
Controls	No	Yes	No	Yes
Observations	1,040	1,037	506	506

Note. The table reports results from Experiment 2 (columns (1) and (2)) and Auxiliary Experiment 6 (columns (3) and (4)). The dependent variable in Panel A is an indicator taking value 1 if the respondent reports believing that his or her matched partner donated to the U.S. Border Crisis Children's Relief Fund. The dependent variable in Panel B is an indicator taking value 1 if the respondent denied his or her matched partner a $1 bonus. Controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, and a set of education indicators. Robust standard errors are reported. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

in the *Cover* condition ($p = .016$). As shown in Table III, these estimates are stable to the inclusion of controls.

To analyze the open-ended text, we look for the words or phrases of up to three words that are most characteristic of each condition. More precisely, we follow Gentzkow and Shapiro (2010) to calculate Pearson's χ^2 statistic for each phrase.[31] This statistic is higher when the use of the phrase is more asymmetric across treatment conditions and lower for phrases that are used rarely across both conditions. Online Appendix Table B.11 shows the 10 phrases most characteristic of each condition (i.e., with the most positive and the most negative χ^2 scores); consistent with our framework and the treatment effects on the structured measures

31. This statistic is given by: $\chi^2_p = \frac{\left(n^R_p n^{NR}_{\sim p} - n^{NR}_p n^R_{\sim p}\right)^2}{(n^R_p + n^{NR}_p)(n^R_p + n^R_{\sim p})(n^{NR}_p + n^{NR}_{\sim p})(n^R_{\sim p} + n^{NR}_{\sim p})}$, where n^R_p, n^{NR}_p are the number of times p appears across all responses in the *Cover* condition and *No Cover* condition, respectively, and $n^i_{\sim p}$ is the total number of times a phrase that is *not* p appears in condition i.

of inference, we find that respondents in the *Cover* condition are more likely to use phrases related to the article or the associated evidence—for example, "article," "read," "convincing," or "increase in crime."[32]

3. *Credibility.* To investigate the role of credibility, we run a slightly revised version of Experiment 2 (Auxiliary Experiment 6) with a sample of 506 Democrats and independents: we instead show respondents the "lower-credibility" versions of the tweets, as described in Section III.B.[33] We display results in the center comparisons of Figure III, Panels A and B and Table III, columns (4)–(6). While the point estimate of the effect of the rationale on both structured measures of inference remains positive, it is substantially smaller: 30% of respondents in the *No Cover* condition believe that their matched partner donated, compared to 33% in the *Cover* condition ($p = .58$), and 44% of respondents in the *No Cover* condition deny their matched partner the donation, compared to 42% in the *Cover* condition.[34] Although we are underpowered to conclude that the difference in treatment effects between the high-credibility and low-credibility wordings is statistically significant, the evidence is consistent with this slightly less credible rationale being substantially less effective.

Our revised experiment also speaks to one of the most common complaints surrounding "political correctness" culture: the alleged tendency of people to "take things out of context." The article prominently lists both Sharkey's academic credentials and, in the first few paragraphs, unequivocally states that "One of the most robust, most uncomfortable findings in criminology is that putting more officers on the street leads to less violent crime." Nonetheless, the revised tweet appears substantially less effective in shifting inference and reducing social sanctions (suggesting that most respondents do not read the article before deciding whether to sanction their partner). Requirements for dissenters

32. These open-ended responses also allow us to mitigate concerns about other potential explanations for our findings: for example, that respondents in the *Cover* condition believed that their matched participant felt pressured by the experimenter to join the campaign and this pressure led them to do so. No respondents mention this or other related confounds.

33. See Online Appendix E.12 for experimental instructions.

34. As shown in Online Appendix D, our results are unchanged if we reweight responses to match the demographics of the sample in the higher-credibility variation.

to ensure that no part of their argument can be taken out of context and stripped of accompanying rationales may leave limited scope for expressing nuanced arguments. Conversely, evidence (such as scientific or media articles) may serve as a rationale even if few people actually examine it, as long as it appears compelling at first glance. We discuss implications for the spread of fake and misleading news and for political entrepreneurship in Section V.

IV. Support for Deporting Illegal Immigrants

Our next experiments examine the use and interpretation of rationales among a different population (conservatives) and to justify a different stigmatized position—support for a campaign to immediately deport all illegal Mexican immigrants. We examine our mechanism in this different context for three primary reasons. First, defunding the police is a highly salient but novel policy proposal, and it is thus unclear whether the power of rationales also extends to more "traditional" policy questions, for which there may be more common knowledge about a greater body of evidence and partisan talking points. Second, opposition to defunding the police is likely stigmatized by the in-group (Democrats) but not the out-group (Republicans); in contrast, supporting the immediate deportation of all illegal Mexican immigrants is less stigmatized by the in-group (Republicans), but is highly stigmatized by the out-group (Democrats). This setting thus allows us to examine whether rationales can be used to mitigate social sanctions levied by the out-group as well as by the in-group. Finally, understanding the drivers of anti-immigrant narratives on social media is of direct interest.

As in the previous experiment on the expression of dissent, we study the expression of xenophobia on social media. Given the widespread and growing importance of right-wing media as suppliers of anti-immigrant narratives, we examine a different form of rationale: a 30-second clip from one of the most popular cable news shows in the United States, *Tucker Carlson Tonight*. In the clip, Carlson draws on statistics from the U.S. Sentencing Commission to argue that illegal immigrants commit violent crimes at substantially higher rates than citizens do.[35]

35. The clip is available at https://www.youtube.com/embed/SDdkkTLCUUQ.

IV.A. Experiment 3: Rationales and Pro-Deportation Expression

1. *Sample and Experimental Design.* We conducted our preregistered Experiment 3 in March 2021 with a sample of Republicans and independents.[36] We recruited 1,130 participants through Luc.id. After screening out respondents who did not want to join the campaign (as described below), we are left with a final sample of 508 respondents. Our sample is balanced on observables across treatment arms (Online Appendix Table B.12).

Our experimental design is broadly similar to that of Experiment 1; we provide a diagram in Online Appendix Figure B.4. As in Experiment 1, respondents log into our survey using their Twitter account and respond to a set of demographic and other background questions. Respondents then view the clip from *Tucker Carlson Tonight*, which is embedded in the survey, and are randomized into the *Cover* condition or the *No Cover* condition. Respondents in the *Cover* condition, but not in the *No Cover* condition, are provided with the URL to the video. We then ask all respondents whether they would like to join a campaign to immediately deport all illegal Mexican immigrants. The survey terminates for respondents who do not join the campaign, leaving us with 517 remaining respondents. Those respondents in the *No Cover* group who do join the campaign are provided the URL to the video. In other words, at this point in the survey, the only difference between conditions is whether respondents are provided with the video URL before (*Cover*) or after (*No Cover*) joining the campaign—though all respondents watch the clip before joining the campaign. As we discuss shortly, this difference in timing is key to avoiding explicit deception in our experimental manipulation.

The remainder of the experiment is identical in design to Experiment 1, with respondents given the opportunity to schedule the following tweet in the *Cover* condition:

> I have joined a campaign to immediately deport all illegal Mexican immigrants. Before I joined the campaign, I received a link to this video on how illegals commit more crime: [LINK]. Sign this petition to immediately deport all illegal Mexicans: [LINK]

Respondents in the *No Cover* condition are presented with an identical tweet, but with the "Before I joined the campaign ..." replaced with "After I joined the campaign ..." Although

36. The full set of experimental instructions is included in Online Appendix E.3.

all respondents in fact watched the video before joining the campaign, it is true that respondents in the *Cover* condition received the link to the video before joining, while those in the *No Cover* condition received the link after joining.[37] This difference in wording suggests to potential readers of the tweet that respondents in the *Cover* group had been exposed to the video by Tucker Carlson before joining the campaign—and thus potentially joined because they were convinced by the clip's evidence—whereas respondents in the *No Cover* group had not been exposed before joining the campaign, and thus could not have joined because of the clip. As in Experiment 1, then, this manipulation varies the availability of social cover while fixing the persuasion channel (all respondents are exposed to the same video) and the anticipated persuasion channel (all respondents know their tweet's readers will be exposed to the video, since it is linked in the tweet).[38]

2. *Results.* The central comparison of Figure II displays the results, which we also show in regression table form in Table IV, Panel A. We again find an economically and statistically significant cover effect: 47% of respondents in the *No Cover* condition authorize the tweet, while 64% of respondents in the *Cover* condition authorize the tweet ($p < .01$, a 0.35 standard deviation effect). The fact that the social cover effect is larger than that estimated in Experiment 1 may reflect that Republicans feel greater stigma in joining a pro-deportation campaign than Democrats feel in joining an anti-defunding campaign (which is also consistent with the lower mean authorization rates in this experiment than in Experiment 1); or that Republicans perceive the *Tucker Carlson* video as a more compelling rationale vis-à-vis their Twitter followers than Democrats perceive the *Washington*

37. One potential concern is that providing a link to respondents in the *Cover* condition, but not in the *No Cover* condition, induces differential selection into the campaign. Because we make the source of the clip obvious, we do not view this as a plausible confound. Indeed, we find no statistically significant difference in selection into the campaign between groups (a 2.6 percentage point difference, $p = .474$), and our worst-case estimate under Lee (2009) bounds remains statistically significant at the 1% level.

38. In principle, we could have used a similar design as Experiment 1: showing the video to respondents both before and after they join the campaign. We concluded that such a manipulation would be less natural for a 30-second video than for a longer article, as in Experiment 1.

TABLE IV
EXPRESSION AND INTERPRETATION OF PRO-DEPORTATION TWEET

	Experiment 3	
Panel A: *Scheduled tweet*		
Cover	0.172***	0.179***
	(0.044)	(0.044)
No Cover mean	0.471	0.471
Observations	508	508
R^2	0.030	0.071
	Experiment 4	
Panel B: *Belief partner donated*		
Cover	0.048**	0.049**
	(0.019)	(0.019)
No Cover mean	0.085	0.085
Observations	1,080	1,079
R^2	0.006	0.033
Panel C: *Denied bonus to partner*		
Cover	−0.064**	−0.064**
	(0.026)	(0.026)
No Cover mean	0.803	0.803
Observations	1,080	1,079
R^2	0.006	0.024
Controls	No	Yes

Note. Panel A presents the results of Experiment 3, in which the dependent variable is an indicator taking value 1 if the respondent chose to schedule the post. Panels B and C present the results of Experiment 4. The dependent variable in Panel B is an indicator taking value 1 if the respondent reports believing that his or her matched partner donated to the U.S. Border Crisis Children's Relief Fund. The dependent variable in Panel C is an indicator taking value 1 if the respondent denied his or her matched partner a $1 bonus. Controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, and a set of education indicators. Robust standard errors are reported. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Post article vis-à-vis their followers.[39] Turning to treatment effect heterogeneity, we show heterogeneity by demographic characteristics in Online Appendix Table B.4, column (4); we show in Online Appendix Table B.5 that our estimated treatment effects remain stable when we reweight the sample to match the

39. In our preregistered Auxiliary Experiment 8 designed to measure the persuasiveness of the rationale, we find mixed evidence for persuasive effects on private opinions; see Online Appendix B.2.2 for details and Online Appendix E.13 for experimental instructions. In a previous working paper (Bursztyn et al. 2020), we present a series of related preregistered experiments examining how the availability of an academic rationale affects conservatives' willingness to publicly donate to an anti-immigrant organization. We again find that the rationale increases public anti-immigrant expression.

general population on observables; and we show heterogeneity by partisan affiliation in Online Appendix Table B.7, Panel C.

IV.B. Experiment 4: Interpretation of Pro-Deportation Rationale

Finally, we examine how the availability of the social cover provided by the *Tucker Carlson Tonight* clip shapes an audience's inference about a dissenter's underlying motivations and the resulting social sanctions the dissenter faces.

1. *Sample and experimental design.* We conducted our preregistered Experiment 4 in November 2021 with a sample of 1,082 Democrats and independents recruited from Prolific.[40] We focus on Democrats and independents, as anti-immigrant expression is less likely to be stigmatized among Republicans. Our sample is balanced on observables across treatment arms (Online Appendix Table B.15).

Experiment 4 follows the structure of Experiment 2; Online Appendix Figure B.3 outlines the structure of the experiments (with red text corresponding to Experiment 4). Respondents are informed that they have been matched with a previous survey participant, who joined a campaign to deport all illegal Mexican immigrants. As in Experiment 2, they are then randomized into a *Cover* and a *No Cover* condition: respondents in the *Cover* condition are told that their matched participant authorized the tweet corresponding to the *Cover* condition of Experiment 3 ("Before I joined the campaign ...") whereas respondents in the *No Cover* condition are told that their matched participant authorized the *No Cover* tweet ("After I joined the campaign ..."). Subsequently, they guess whether their matched participant authorized a $5 donation to the U.S. Border Crisis Children's Relief Fund (an organization that seeks to provide care and basic hygiene items to children along the U.S.–Mexico border) when given the opportunity to do so, and they choose whether to deny a $1 bonus to their matched participant.[41]

2. *Results.* The rightmost comparisons of Figure III display the fraction of participants in the *Cover* and *No Cover* condition who believe their matched participant donated to the

40. The full set of experimental instructions is included in Online Appendix E.4.

41. We randomized the order of these two different outcomes and detect no significant order effects.

pro-immigrant organization and the corresponding fractions of participants who deny their matched respondent a bonus. Of respondents in the *No Cover* condition, 8.5% believe their matched participant donated, compared to 13.4% of respondents in the *Cover* condition ($p = .01$); 80% of respondents in the *No Cover* condition deny their matched participant a bonus, compared to 74% of respondents in the *Cover* condition ($p = .011$). As shown in Table IV, Panels B and C, these estimates are stable to the inclusion of demographic controls.

We plot results from our analysis of open-ended text in Online Appendix Table B.17, using the same procedure described in Section III.C.2. As in Experiment 2, respondents in the *Cover* condition are substantially more likely to use words referencing the rationale—"watched a video," "fear mongering," "convinced"—whereas respondents in the *No Cover* condition mention phrases such as "Republican" and "racial." This evidence underscores that rationales shift inference about underlying motives.

V. Discussion and Conclusion

This article examines how rationales facilitate dissent by lowering the social cost of expressing controversial opinions. In our model, rationales change some people's private views, but they also change an audience's inference about dissenters' motivations and thus can be used to enable dissent. We explore these mechanisms among both liberal and conservative respondents, focusing on a natural setting and outcome: willingness to express dissent on social media. First, we show that liberal respondents are more likely to authorize a tweet opposing the movement to defund the police when they can credibly ascribe their views to strong scientific evidence. Consistent with our framework, a credible rationale shifts an audience's inference about the respondents and reduces resulting social sanctions. Similarly, conservative respondents are more likely to authorize a tweet calling for the deportation of all illegal immigrants from Mexico—and are seen as less intolerant after doing so—when they can ascribe their views to a Fox News clip.[42]

42. While our experiments explore settings in which there is pressure to express more liberal views—and thus, the rationale supports a more centrist view in Experiment 1 and a more right-wing view in Experiment 3—our conceptual

We now discuss some implications of our framework and empirical results, which may provide fruitful avenues for future research.

V.A. *Political Correctness and the Limitations of Rationales*

In a "political correctness" culture, certain rationales cannot be voiced because they are seen as legitimizing dangerous or undesirable causes, and so anyone who uses such a rationale is seen as supporting the cause itself. For example, people who argue for the presence of reverse discrimination against men in labor markets may be seen as sexist: that is, even scientific rationales such as correspondence studies—which may be effective rationales in other settings—may fail to provide a social cover. In some cases, this may be socially desirable: for instance, equating the use of a rationale with sexism may prevent sexist people from citing rationales they do not believe or cherry-picking rationales to support their claims. In other cases, political-correctness culture may stifle socially important forms of dissenting expression by stigmatizing rationales that would typically be seen as highly credible.[43]

Individuals or institutions seeking to eliminate certain forms of public behavior may use multiple levers to silence dissenters. One lever, explored in Section III.B, is to undermine the credibility of rationales directly. Another lever is to manipulate the real or perceived correlation between knowledge of a rationale and the underlying type, tying the rationale directly to the stigmatized motive.[44] Indeed, in the limit in which only people with stigmatized motives are aware of a certain rationale—for

framework generalizes to any context in which certain types are stigmatized and public expression is informative about type.

43. The announcement of new ethics requirements in the prominent journal *Nature Human Behaviour* highlights this tension (see "Science Must Respect the Dignity and Rights of All Humans." *Nature Human Behaviour*, (2022, 1029–1031) "In some cases ...potential harms to the populations studied may outweigh the benefit of publication. Academic content that undermines the dignity or rights of specific groups ...or promotes privileged, exclusionary perspectives raises ethics concerns that may require revisions or supersede the value of publication ...[but] ensuring that no research is discouraged simply because it may be socially or academically controversial, is as important as preventing harm."

44. For example, during the second Red Scare, Joseph McCarthy and his allies explicitly tied several rationales for dissenting with government policy to communist sympathies. Famously, physicist J. Robert Oppenheimer was stripped of his security clearances when political opponents attributed his opposition to the development of the hydrogen bomb to alleged Soviet loyalties (Cassidy 2019).

example, because only they consume the extreme news sources through which the rationale is broadcast—the rationale is completely ineffective, as to use it is to reveal one's motives with certainty. Tactics to manipulate this real or perceived correlation include disallowing controversial opinions a public platform (e.g., disinviting campus speakers or banning social media accounts) or branding particular media sources or speakers as fringe.[45] Further exploring the conditions under which rationales are most effective, and heterogeneity in the types and sources of rationales that are effective across different groups, is an important direction for future research.[46] For example, evidence that noncredible rationales can backfire, leading to greater social sanctions, would have important implications for understanding social dynamics and the supply side of political narratives. Similarly, evidence on how people endogenously acquire rationales and the supply-side implications of such strategic behavior might shed light on both the causes and consequences of increasing polarization in the media.

V.B. Political Entrepreneurship and Populism

Populist politicians often scapegoat minorities (Bursztyn et al. 2022; Guriev and Papaioannou 2022). While the persuasive effects of political propaganda are doubtless important (Adena et al. 2015), propaganda may also generate social cover, enabling supporters to speak their mind more openly and spread the message through their social circle (Satyanath, Voigtländer, and Voth 2017; Caesmann et al. 2021). The strength of this social amplifier depends not only on the number of individuals who hold stigmatized views but also on the number of individuals who previously could not express these views. This may be one reason the Nazis were able to leverage social networks and associations more effectively than other groups, such as communists: if anti-Semitism was stigmatized, but relatively common and

45. This can also help explain how censorship techniques such as China's "Great Firewall" can be highly effective in repressing discourse unfriendly to the regime, even if citizens can bypass them relatively easily (Chen and Yang 2019).

46. Policy makers can also use rationales to affect behavior in nonpolitical settings. For instance, in settings where educational investments are stigmatized (Austen-Smith and Fryer 2005), providing monetary incentives for exerting educational effort (Levitt et al. 2016) might enable students to attribute educational investments not to academic interest but to the incentive. For similar reasons, cold-calling might be preferable to allowing students to volunteer answers.

persistent (Voigtländer and Voth 2012; Cantoni, Hagemeister, and Westcott 2019), then anti-Semitic Nazi rhetoric generated a large social amplifier. In contrast, blaming economic elites was less stigmatized and thus generated smaller amplifiers.

A more recent rhetorical strategy is dog-whistling: "sending a message to certain potential supporters in such a way as to make it inaudible to others whom it might alienate or deniable for still others who would find any explicit appeal along those lines offensive" (Goodin and Saward 2005). Historians and political scientists have argued that the Republican Party's "Southern strategy" to win white support in the South was characterized by extensive racial dog whistling (Haney-López 2014). In a 1981 interview, Republican strategist and Republican National Committee chairman Lee Atwater described the approach as follows:

> You start out in 1954 by saying, "N—, n—, n—." By 1968 you can't say "n—": that hurts you. Backfires. So you say stuff like forced busing, states' rights and all that stuff. You're getting so abstract now [that] you're talking about cutting taxes, and all these things you're talking about are totally economic things and a byproduct of them is [that] blacks get hurt worse than whites. (Perlstein 2012)

Such dog whistles generate two types of social cover: one for the politician vis-à-vis the greater public, and one for the politician's supporters vis-à-vis others who disapprove of the statement, allowing them to publicly support the politician and his or her policies without incurring social stigma.

V.C. Fake and Misleading News on Social Media

Our findings speak to the influence of fake and misleading news on social media. Some recent studies suggest that the persuasive effect of fake and misleading news is limited (Allcott and Gentzkow 2017; Nyhan 2018), whereas others suggest the opposite: that such stories can affect behavior (Barrera et al. 2020) and that individuals may have trouble distinguishing between fake and real news (Angelucci and Prat 2021) or between facts and opinions (Bursztyn et al. forthcoming). Our results highlight the potential importance of mechanisms beyond persuasion. Specifically, fake and misleading news can generate a social amplifier: rationales that plausibly persuade a small group can change public behavior among a much larger group. This is particularly concerning given that the breadth of rationales, especially those for fringe views, is far greater on social media than on traditional outlets.

Among other platforms, Facebook and Twitter have conducted small-scale experiments evaluating strategies to curtail the spread of misinformation, including warning users before they post an article flagged as fake news and flagging fake or misleading news when it appears on others' timelines. The effects of such interventions are typically modest (Jahanbakhsh et al. 2021). Yet because these changes have not been rolled out at scale, users retain social cover when sharing fake news: they can credibly claim that they did not know the news was fake. Scaling these initiatives to the entire user base, thus generating common knowledge that all users are warned before posting fake news, would eliminate this cover. For this reason, current (partial equilibrium) estimates of the effects of debunking on users' propensity to share fake news may substantially understate the general-equilibrium effects that would be realized if platforms were to fully scale up the feature. At the same time, the evidence from Barrera et al. (2020) emphasizes the importance of platforms' credibility when debunking rationales: when platforms lack credibility, fake and misleading news retains its power to generate social cover for the expression of stigmatized views.

V.D. Dynamics

Our experiments investigated a snapshot of the United States between 2021 and 2022. But what are the mechanisms by which social norms surrounding the expression of particular views vary over time and by which particular rationales become more or less effective? Both the credibility of any given rationale and the stigma associated with certain positions (including the rationales and positions we study here) are likely to change over time, either due to strategic manipulation by certain individuals and institutions or due to broader social dynamics. For example, particular topics may become normalized, or particular sources may become delegitimized or associated with stigmatized causes. Rationales that were effective at one time may no longer be effective at a later date—or they may no longer be needed, either because the view has become normalized or because those holding the view care less about the sanctions imposed by the out-groups that would disagree. Understanding these dynamics is an exciting direction for future work.

UNIVERSITY OF CHICAGO AND NATIONAL BUREAU OF ECONOMIC RESEARCH, UNITED STATES
KELLOGG SCHOOL OF MANAGEMENT AND NATIONAL BUREAU OF ECONOMIC RESEARCH, UNITED STATES
NHH NORWEGIAN SCHOOL OF ECONOMICS, NORWAY
HARVARD UNIVERSITY, UNITED STATES
UNIVERSITY OF COLOGNE AND MAX PLANCK INSTITUTE FOR RESEARCH ON COLLECTIVE GOODS, GERMANY, AND CENTRE FOR ECONOMIC POLICY AND RESEARCH, UNITED KINGDOM

SUPPLEMENTARY MATERIAL

An Online Appendix for this article can be found at *The Quarterly Journal of Economics* online.

DATA AVAILABILITY

The data underlying this article are available in the Harvard Dataverse, https://doi.org/10.7910/DVN/YVUIFF (Bursztyn et al. 2023).

REFERENCES

Acemoglu, Daron, Georgy Egorov, and Konstantin Sonin, "A Political Theory of Populism," *Quarterly Journal of Economics*, 128 (2013), 771–805. https://doi.org/10.1093/qje/qjs077

Adena, Maja, Ruben Enikolopov, Maria Petrova, Veronica Santarosa, and Ekaterina Zhuravskaya, "Radio and the Rise of The Nazis in Prewar Germany," *Quarterly Journal of Economics*, 130 (2015), 1885–1939. https://doi.org/10.1093/qje/qjv030

Ali, Nageeb S., and Charles Lin, "Why People Vote: Ethical Motives and Social Incentives," *American Economic Journal: Microeconomics*, 5 (2013), 73–98. https://doi.org/10.1257/mic.5.2.73

Allcott, Hunt, and Matthew Gentzkow, "Social Media and Fake News in the 2016 Election," *Journal of Economic Perspectives*, 31 (2017), 211–236. https://doi.org/10.1257/jep.31.2.211

Angelucci, Charles, and Andrea Prat, "Is Journalistic Truth Dead? Measuring How Informed Voters Are about Political News," Social Science Research Network Working Paper 3593002, 2021.

Austen-Smith, David, and Roland G. Fryer, Jr., "An Economic Analysis of 'Acting White'," *Quarterly Journal of Economics*, 120 (2005), 551–583. https://doi.org/10.1093/qje/120.2.551

Barrera, Oscar, Sergei Guriev, Emeric Henry, and Ekaterina Zhuravskaya, "Facts, Alternative Facts, and Fact Checking in Times of Post-Truth Politics," *Journal of Public Economics*, 182 (2020), 104123. https://doi.org/10.1016/j.jpubeco.2019.104123

Bénabou, Roland, Armin Falk, and Jean Tirole, "Narratives, Imperatives, and Moral Persuasion," NBER Working Paper 24798, 2018. https://doi.org/10.3386/w24798

Bénabou, Roland, and Jean Tirole, "Incentives and Prosocial Behavior," *American Economic Review*, 96 (2006), 1652–1678. https://doi.org/10.1257/aer.96.5.1652

———, "Identity, Morals, and Taboos: Beliefs as Assets," *Quarterly Journal of Economics*, 126 (2011), 805–855. https://doi.org/10.1093/qje/qjr002
Braghieri, Luca, "Political Correctness, Social Image, and Information Transmission," Stanford University Working Paper, 2022.
Bursztyn, Leonardo, Georgy Egorov, Ruben Enikolopov, and Maria Petrova, "Social Media and Xenophobia: Evidence from Russia," NBER Working Paper no. 26567, 2019. https://doi.org/10.3386/w26567
Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin, "From Extreme to Mainstream: The Erosion of Social Norms," *American Economic Review*, 110 (2020), 3522–3548. https://doi.org/10.1257/aer.20171175
Bursztyn, Leonardo, Georgy Egorov, Ingar Haaland, Aakaash Rao, and Christopher Roth, "Scapegoating During Crises," *AEA Papers and Proceedings*, 112 (2022), 151–155. https://doi.org/10.1257/pandp.20221069
———, "Replication Data for: 'Justifying Dissent,'" (2023), Harvard Dataverse, https://doi.org/10.7910/DVN/YVUIFF.
Bursztyn, Leonardo, Alessandra L. González, and David Yanagizawa-Drott, "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia," *American Economic Review*, 110 (2020), 2997–3029. https://doi.org/10.1257/aer.20180975
Bursztyn, Leonardo, Ingar K. Haaland, Aakaash Rao, and Christopher P. Roth, "Disguising Prejudice: Popular Rationales as Excuses for Intolerant Expression," NBER Working Paper no. 27288, 2020. https://doi.org/10.3386/w27288
Bursztyn, Leonardo, and Robert Jensen, "How Does Peer Pressure Affect Educational Investments?," *Quarterly Journal of Economics*, 130 (2015), 1329–1367. https://doi.org/10.1093/qje/qjv021
Bursztyn, Leonardo, Aakaash Rao, Christopher Roth, and David Yanagizawa-Drott, "Opinions as Facts," *Review of Economic Studies*, forthcoming. https://doi.org/10.1093/restud/rdac065
Caesmann, Marcel, Bruno Caprettini, Hans-Joachim Voth, and David Yanagizawa-Drott et al. "Going Viral: Propaganda, Persuasion and Polarization in 1932 Hamburg," Centre for Economic Policy Research Technical Report, 2021. https://doi.org/10.5167/uzh-205774
Cantoni, Davide, Felix Hagemeister, and Mark Westcott, "Persistence and Activation of Right-Wing Political Ideology," Rationality and Competition Discussion Paper Series 143, CRC TRR 190, 2019.
Caprettini, Bruno, Marcel Caesmann, Hans-Joachim Voth, and David Yanagizawa-Drott, "Going Viral: Propaganda, Persuasion and Polarization in 1932 Hamburg," Center for Economic and Policy Research Working Paper 16356, 2021.
Cassidy David, C., *J. Robert Oppenheimer and the American Century* (Lexington, MA: Plunkett Lake Press, 2019).
Chen, Yuyu, and David Y. Yang, "The Impact of Media Censorship: 1984 or Brave New World?," *American Economic Review*, 109 (2019), 2294–2332. https://doi.org/10.1257/aer.20171765
Cohn, Nate, and Kevin Quealy, "The Democratic Electorate on Twitter Is Not the Actual Democratic Electorate," New York Times, April 9, 2019.
Crowne, Douglas P., and David Marlowe, "A New Scale of Social Desirability Independent of Psychopathology," *Journal of Consulting and Clinical Psychology*, 24 (1960), 349–354. https://doi.org/10.1037/h0047358
Cunningham, Tom, and Jonathan de Quidt, "Implicit Preferences Inferred from Choice," Social Science Research Network Working Paper 2709914, 2015. https://doi.org/10.2139/ssrn.2709914
Dana, Jason, Roberto A. Weber, and Jason Xi Kuang, "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness," *Economic Theory*, 33 (2007), 67–80. https://doi.org/10.1007/s00199-006-0153-z
Dhar, Diva, Tarun Jain, and Seema Jayachandran, "Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India," *American Economic Review*, 112 (2022), 899–927. https://doi.org/10.1257/aer.20201112

Dreyfuss, Bnaya, Assaf Patir, and Moses Shayo, "On the Workings of Tribal Politics," Social Science Research Network Working Paper 3797290, 2021. https://doi.org/10.2139/ssrn.3797290

Ekins, Emily E., "Poll: 62% of Americans Say They Have Political Views They're Afraid to Share," Social Science Research Network Working Paper 3659953, 2020.

Enikolopov, Ruben, Alexey Makarin, and Maria Petrova, "Social Media and Protest Participation: Evidence from Russia," *Econometrica*, 88 (2020), 1479–1514. https://doi.org/10.3982/ECTA14281

Enikolopov, Ruben, and Maria Petrova, "Chapter 17 - Media Capture: Empirical Evidence," in Simon P. Anderson, Joel Waldfogel, and David Strömberg, editors, *Handbook of Media Economics*, volume 1. (Saint Louis: Elsevier Science & Technology, 2015), 687–700.

Exley, Christine L., "Excusing Selfishness in Charitable Giving: The Role of Risk," *Review of Economic Studies*, 83 (2016), 587–628. https://doi.org/10.1093/restud/rdv051

Foerster, Manuel, and Joël J. van der Weele, "Casting Doubt: Image Concerns and the Communication of Social Impact," *Economic Journal*, 131 (2021), 2887–2919. https://doi.org/10.1093/ej/ueab014

Gennaioli, Nicola, and Andrei Shleifer, "What Comes to Mind," *Quarterly Journal of Economics*, 125 (2010), 1399–1433. https://doi.org/10.1162/qjec.2010.125.4.1399

Gentzkow, Matthew, and Jesse M. Shapiro, "What Drives Media Slant? Evidence from U.S. Daily Newspapers," *Econometrica*, 78 (2010), 35–71. https://doi.org/10.3982/ECTA7195

Golman, Russell, "Acceptable Discourse: Social Norms of Beliefs and Opinions," Carnegie Mellon University Working Paper, 2021. http://dx.doi.org/10.2139/ssrn.4160955

Golman, Russell, David Hagmann, and George Loewenstein, "Information Avoidance," *Journal of Economic Literature*, 55 (2017), 96–135. https://doi.org/10.1257/jel.20151245

Goodin, Robert E., and Michael Saward, "Dog Whistles and Democratic Mandates," *Political Quarterly*, 76 (2005), 471–476. https://doi.org/10.1111/j.1467-923X.2005.00708.x

Grossman, Zachary, and Joël J. Van Der Weele, "Self-Image and Willful Ignorance in Social Decisions," *Journal of the European Economic Association*, 15 (2017), 173–217. https://doi.org/10.1093/jeea/jvw001

Guriev, Sergei, and Elias Papaioannou, "The Political Economy of Populism," *Journal of Economic Literature*, 60 (2022), 753–832. https://doi.org/10.1257/jel.20201595

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart, "Designing Information Provision Experiments," *Journal of Economic Literature* (forthcoming). http://dx.doi.org/10.2139/ssrn.4160955

Hamman, John R., George Loewenstein, and Roberto A. Weber, "Self-Interest through Delegation: An Additional Rationale for the Principal-Agent Relationship," *American Economic Review*, 100 (2010), 1826–1846. https://doi.org/10.1257/aer.100.4.1826

Haney-López, Ian, *Dog Whistle Politics: How Coded Racial Appeals Have Reinvented Racism and Wrecked the Middle Class* (New York: Oxford University Press, 2014).

Jahanbakhsh, Farnaz, Amy X. Zhang, Adam J. Berinsky, Gordon Pennycook, David G. Rand, and David R. Karger, "Exploring Lightweight Interventions at Posting Time to Reduce the Sharing of Misinformation on Social Media," *Proceedings of the ACM on Human-Computer Interaction*, 5 (2021), 1–42. https://doi.org/10.1145/3449092

Kuran, Timur, *Private Truths, Public Lies: The Social Consequences of Preference Falsification* (Cambridge, MA: Harvard University Press, 1997).

Lacetera, Nicola, and Mario Macis, "Social Image Concerns and Prosocial Behavior: Field Evidence from a Nonlinear Incentive Scheme," *Journal of Economic Behavior and Organization*, 76 (2010), 225–237. https://doi.org/10.1016/j.jebo.2010.08.007

Langer, Ellen J., Arthur Blank, and Benzion Chanowitz, "The Mindlessness of Ostensibly Thoughtful Action: The Role of 'Placebic' Information in Interpersonal Interaction," *Journal of Personality and Social Psychology*, 36 (1978), 635–642. https://doi.org/10.1037/0022-3514.36.6.635
Lazear, Edward P., Ulrike Malmendier, and Roberto A. Weber, "Sorting in Experiments with Application to Social Preferences," *American Economic Journal: Applied Economics*, 4 (2012), 136–163. https://doi.org/10.1257/app.4.1.136
Lee, David S., "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76 (2009), 1071–1102. https://doi.org/10.1111/j.1467-937X.2009.00536.x
Levitt, Steven D., John A. List, Susanne Neckermann, and Sally Sadoff, "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance," *American Economic Journal: Economic Policy*, 8 (2016), 183–219. https://doi.org/10.1257/pol.20130358
Levy, Ro'ee, and Martin Mattsson, "The Effects of Social Movements: Evidence from #MeToo," Social Science Research Network Working Paper 3496903, 2021. http://dx.doi.org/10.2139/ssrn.3496903
Litman, Leib, Jonathan Robinson, and Tzvi Abberbock, "TurkPrime.com: A Versatile Crowdsourcing Data Acquisition Platform for the Behavioral Sciences," *Behavior Research Methods*, 49 (2017), 433–442. https://doi.org/10.3758/s13428-016-0727-z
Michalopoulos, Stelios, and Melanie Meng Xue, "Folklore," *Quarterly Journal of Economics*, 136 (2021), 1993–2046. https://doi.org/10.1093/qje/qjab003
Morris, Stephen, "Political Correctness," *Journal of Political Economy*, 109 (2001), 231–265. https://doi.org/10.1086/319554
Müller, Karsten, and Carlo Schwarz, "From Hashtag to Hate Crime: Twitter and and Anti-Minority Sentiment," *American Economic Journal: Applied Economics*, forthcoming.
Nature Human Behaviour, "Science Must Respect the Dignity and Rights of All Humans," *Nature Human Behaviour*, 6, (2022), 1029–1031. https://doi.org/10.1038/s41562-022-01443-2
Nyhan, Brendan, "Fake News and Bots May Be Worrisome, but Their Political Power Is Overblown," *New York Times*, February 3, 2018.
O'Brien, Sarah, "Employers Check your Social Media Before Hiring. Many then Find Reasons not to Offer you a Job," *CNBC*, August 10, 2018.
Parker, Kim, and Kiley Hurst, "Growing Share of Americans Say they Want More Spending on Police in their Area," Pew Research Center Report, 2021.
Perez-Truglia, Ricardo, and Guillermo Cruces, "Partisan Interactions: Evidence from a Field Experiment in the United States," *Journal of Political Economy*, 125 (2017), 1208–1243. https://doi.org/10.1086/692711
Perlstein, Rick "Exclusive: Lee Atwater's Infamous 1981 Interview on the Southern Strategy." *The Nation*, 13 (2012). https://www.thenation.com/article/archive/exclusive-lee-atwaters-infamous-1981-interview-southern-strategy/
Reynolds, William M., "Development of Reliable and Valid Short Forms of the Marlowe-Crowne Social Desirability Scale," *Journal of Clinical Psychology*, 38 (1982), 119–125. https://doi.org/10.1002/1097-4679(198201)38:1%3C119::AID-JCLP2270380118%3E3.0.CO;2-I
Saccardo, Silvia, and Marta Serra-Garcia, "Cognitive Flexibility or Moral Commitment? Evidence of Anticipated Belief Distortion," Social Science Research Network Working Paper 3676711, 2020.
Satyanath, Shanker, Nico Voigtländer, and Hans-Joachim Voth, "Bowling for Fascism: Social Capital and the Rise of the Nazi Party," *Journal of Political Economy*, 125 (2017), 478–526. https://doi.org/10.1086/690949
Shalvi, Shaul, Francesca Gino, Rachel Barkan, and Shahar Ayal, "Self-Serving Justifications: Doing Wrong and Feeling Moral," *Current Directions in Psychological Science*, 24 (2015), 125–130. https://doi.org/10.1177/0963721414553264

Sharkey, Patrick, "Why Do We Need the Police?," Washington Post, June 12, 2020.
Shiller, Robert J., "Narrative Economics," *American Economic Review*, 107 (2017), 967–1004. https://doi.org/10.1257/aer.107.4.967
Thompson, Derek, "Unbundle the Police," *The Atlantic*, June 11, 2020.
Voigtländer, Nico, and Hans-Joachim Voth, "Persecution Perpetuated: The Medieval Origins of Anti-Semitic Violence in Nazi Germany," *Quarterly Journal of Economics*, 127 (2012), 1339–1392. https://doi.org/10.1093/qje/qjs019
Wood, Thomas, and Ethan Porter, "The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence," *Political Behavior*, 41 (2019), 135–163. https://doi.org/10.1007/s11109-018-9443-y
Yanagizawa-Drott, David, "Propaganda and Conflict: Evidence from the Rwandan Genocide," *Quarterly Journal of Economics*, 129 (2014), 1947–1994. https://doi.org/10.1093/qje/qju020